Performance Evaluation of Crowdsourced HCI User Studies

February - May 2015

AALBORG UNIVERSITY
STUDENT REPORT

Authors:
MTA151035
Allan Christensen
Simon André Pedersen

AALBORG UNIVERSITET

Department of Architechture,
Design and Media Technology
Medialogy, 10th Semester Master Project

Title:
Performance Evaluation of Crowdsourced HCI User Studies

Theme:
Master's Thesis

Project period:
3rd of February 2015 -
27th of May 2015

Project group:
MTA151035

Members:
Allan Christensen

Simon André Pedersen

Supervisor:
Hendrik Knoche

Total no. of pages: 57

Abstract:

In this study, we examined the viability for HCI user studies to use crowdsourcing as a participant group. Potentially it could yield higher attendance for the studies and studies would not rely on subjects classified as WEIRD. We conducted a preliminary study to determine if touch or tilt controlled a game better within a lab environment. We found that touch outperformed tilt. Following this study, we examined if an informed crowd (informed about being in an experiment) and uninformed crowd could perform equivalent to participants in a controlled lab environment. The study showed that in our first level, a touch controlled game, the lab environment outperformed both of the crowds, while the informed crowd performed better than the uninformed. The second level featured a device human resolution experiment through Fitts' law, to determine the smallest selectable target with little effort. The test revealed that the lab consistently produced fewer errors and we saw a significant increase in errors between a Fitts' ID of 3.70 and 4.64. For the informed crowd we saw a spike in errors for a Fitts' ID between 2.81 and 3.70. The uninformed crowd had generally too many errors to determine a significant increase in errors. The smallest selectable target for all three groups combined, was between 2 mm and 4 mm for touch devices.

Preface

This report is written in 2015 as a 10th semester master thesis project by group MTA151035 at Medialogy, Aalborg University. The purpose of the project is to examine if there are performance differences between conducting HCI user studies on touch devices, in a controlled lab environment compared to on crowdsourced participants in their own environment.

This report functions as a portfolio with full descriptions included in all sections. It is a long version of the paper, in order to give more information to the reader, where it might be needed.

A DVD added to the report contains an AV-production showing the purpose of the project along with a demonstration of the application, a pdf file of the report and paper, and a file containing the application.

Table of Contents

Chapter 1

Introduction

In the scientific society Human Computer Interaction (HCI) user studies are conducted all the time. These experiments are conducted on both small and large user groups, where participants are recruited from various target groups in order to analyse their behavior and reactions when interacting with computers or software in general.

However, many of these experiments also come with certain disadvantages. When users are brought into a testing setup or environment, they know that their results matter greatly to the people they are testing for, which quickly enables users to respond differently due to the testing environment they are placed in. Furthermore, the setups can positively influence the data as the setup does often not suffer from for example noise in the environment, and it can therefore be discussed how good the ecological validity of the data actually is and if the setups are comparable to the actual environments that products eventually will be used in.

Therefore it is relevant to examine if there is a difference between users participating in a user study in an experimental environment and users interacting with the same application in their natural environment.

1.1 Concept

One of the issues often seen when testing in an experimental setup is that it is the same participants used repeatedly for different scenarios. These test subjects are often students (mainly undergraduate) testing an experiment for a friend, professor, or being paid by a fellow student to do so. Actually 96% of subjects in behavioral science research come from Western industrialized countries, where 67% of the American samples, and 80% of the samples from other countries are undergraduates in psychology courses [1, 17]. The question is then how representative these typical subjects actually are.

Henrich et al. [17] conducted research in this field of study and created a category of what they called WEIRD subjects, which is an acronym for Western, Educated, Industrialized, Rich, Democratic. One of the issues they address is the fact that it seems like most behavioral science experiments draw their samples from the WEIRD subjects, which consists of an extremely narrow slice of the human diversity, who might also be a peculiar subpopulation. They found that social decision-making experiments showed that the

WEIRD subjects often occupy the extreme end of the behavioral science distribution and therefore provide very different results than many other subject groups. Another issue is that most behavioral scientists do not necessarily acquire these subjects only due to laziness, but also routinely assume that their findings from the WEIRD subjects actually generalize the entire species. Their findings also show that Westerners in general have more independent views of self than non-Westerners, which means that they are more likely to:

- Demonstrate a positively biased view of themselves
- Have a heightened valuation of personal choice
- Have an increased motivation to "stand out" rather than to "fit in"

The heightened sense of choice can also be seen in Section 2.2 on gamification, where we see that people prefer games where is they feel like they have a sense of choice and autonomy. Furthermore, there have only been a few studies that examined if undergraduates or college-educated Americans differ from people who are not students, or were never college-educated. This is, however, mainly in the area of psychological measures [17].

In their thesis, Henrich et al. conclude that WEIRD people are outliers in many key domains of behavioral science, which may actually make them one of the worst subpopulations to study, when trying to generalize human behavior. They further conclude that much research is missing and that they cannot accurately evaluate how unusual the WEIRD population actually is, meaning that they want to encourage researchers to *span the globe* and prove that the population is actually representative of the entire human population [17]. This problem, and the solution, is also suggested by Baumard and Sperber [3], who further suggests that the results may actually have more to do with methodological problems and processes that some of these experiments are intended to illuminate, rather than actual cultural differences. Furthermore, the analysis should reflect if the differences in the data is actually due to the differences in the interpretation of the experimental situation, rather than actual differences.

Gächter [11] also points out that the issues might not necessarily occur because of the WEIRD population, but rather in the ways that the population is being used. Currently most, if not all, research in many experiments and papers use the WEIRD population for testing through the entire process and the type of test does not matter. Gächter suggests that researchers should consider their research question before selecting a population, as the WEIRD population could in some cases be the best subject pool. He further states that the WEIRD subject pool could, and should, be used as a benchmark or starting point for investigating generalizability to other social groups. This means that the data could be used for early analysis, and then expand the experiments later if it seems relevant [11]. In hindsight of the research stated above, we have chosen to investigate the differences between the WEIRD population (in this case mainly students), against a wider general population. In order to do this we have chosen to examine the WEIRD population in an experimental setup against the general population who are testing in their natural environment.

Chapter 2

Research

2.1 Crowdsourcing Experiments

A huge successor of crowdsourcing is Wikipedia. Wikipedia is having a crowd of people working instead of paying article workers to updates all the articles [23]. The word crowdsourcing is a combination of the words "Crowd" and "outsourcing". This usually mean instead of in-house development or designing this process is given to a large pool of people to solve. A force of using crowdsourcing is e.g. designing a logo, as many freelancers can send in their proposal for a design, and the company get to choose the best, and only pay a relatively low price for the logo. When investigating research and crowdsourcing one of the most common solutions is Amazon Mechanical Tuck, which is a platform for people to upload their projects or other micro tasks they want solve [23, 36]. The company then pays per task solved by the people participating. This crowd know this they are testing products for companies, and research show that the quality of results and dedication from the crowd is following the payment from the companies. The benefit of crowdsourcing using Amazon Mechanical tuck compared to testing the same in lab environments is the amount of participants participating in the test. Payment for the work completed is higher per participant in lab environments, and time used testing is higher as well [22].

2.1.1 Current crowdsourced mobile research games

Relatively little research exist on research projects using crowdsourcing as participant pool on mobile devices. Currently in the app stores, few project exists that depends on crowdsourcing [19, 23]. One of these projects, made by Henze et al. [19], has developed a mobile game that got over 108.000 installations in 72 days. They gathered touch events and other unidentifiable information from the participants without letting the participants know of the data collection or purpose of collecting data. Their participants played on average $\sim$21 levels. They investigated selection of targets and the result showed error rate increased rapidly on targets below 15mm and error rate increased to over 40% when going below 8 mm, these results were gathered from over 8.6 million touch events across 4 devices. Google recommend a target size range between 7-10 mm, which contradicts the results from Henze et al. However, Google do not mention how they achieved their results [13]. Therefore, unknown variables influence the data and one answer is that people crowdsourcing try to be as fast as possible compared to participants in lab environments

try to show the best results to the researcher [19, 18].

2.1.2 Validity

When conducting experiments in lab environments, a low number of participants appears frequently throughout studies and further the tasks tested has low generalizable results [23]. External Validity and the counterpart internal validity try to explain the difference between a lab environment and real world testing. When testing in a lab, the internal validity is high, because the researchers control the lab conditions such as location, time, lighting and noise. However, the external validity is low because the experimental factors is highly controlled and therefore the results observed might only be valid within the lab environment. When considering an app in the app store, there is potentially, a lot higher amount of participants participating in the experiment, and further the participants will be within their natural setting. An high external validity is expected, since results observed can be considered generalizable, however, this compromise the internal validity, with all the confounding variable, this mean the results cannot be assured coming from participants performance because it might be due to unknown sources.

2.1.3 Data Logging

When conducting HCI experiments usually 14 - 30 participants deliver valid data entries [16, 23]. Releasing the experiment to the app store the number of participants and valid data potentially increase, however, Henze et al. [19], found that the number of installations highly depends on type of app and more importantly how many data entries depends on how participant are told that data is being logged. They released five different apps with each a new approach to let the user know of data logging. They discovered that telling the user without letting the user opt-out is only marginally worse than not telling the user at all. When letting the user opt-out of the experiment data entries fall from 27% and up to 80% [19, 26].

2.1.3.1 Ethics of Data Logging without Consent

When conducting internet based research Nosek et al. [30] discovered differences between lab research and the crowdsourced research. They found with the absence of the research potentially raise issues with research ethics. The debriefing of the participant may be missing if the participant leave the study early due to e.g. boredom. Protection of minor in a study is essential, it will be very visible in a lab environment and easy to solve however, in crowdsourced studies it can be hard to control who participate. Asking for age before the study will remove the minor from the data but they are still able to complete the study. Nosek et al, mention the absence of a research can be positive in terms of ethics regarding participation withdrawal because participants might feel forces to complete the study even though they feel uncomfortable participating. When a study is completed face-to-face with a researcher people will not opt-out and with internet-based research, the participant is more likely to voluntarily opting-out [19, 23, 30].

2.2 Gamification

In order to examine how we can disguise the real intention of our application from the users we build upon the term of gamification.

Gamification is a term that originates from the digital media industry. In recent years the term has become a more popular subject in academic research and in the service industry [15]. Furthermore, Gartner [12] has stated that more than 50 percent of companies and organizations will gamify their innovation processes by 2015. There are different definitions of gamification, but one of the more acknowledged is the use of game design elements in non-game contexts proposed by Deterding et al. [8]. The point that the definition of gamification is trying to establish is that by applying gamification (or gamifying) a non-game product or service we are able to motivate user engagement and make the product more enjoyable [10, 38]. This also indicates a higher replay-value of the product, as users feel that it is worth using the application again in order to for example beat an existing score or compete and engage with others, if there is a community using the product.

In order to further help define gamification, Deterding et al. defined five levels of abstraction. These can be seen in Table 2.1. One of the main points that Deterding et al. criticize in their definition of gamification is the current state of gamification, where it is currently used as putting points, badges and leaderboards on everything to make it seem like a game [8, 14]. One of the issues that they also address is for example that children will draw more pictures in lesser quality if they are paid to be drawing pictures, however, as soon as the children are stopped being paid, their motivation for drawing pictures is reduced. This effect (also known as *overjustification*) can also occur in games with point systems, where if the points are removed, it creates a negative attitude towards the game [14].

Because of this behavior they introduce three principles that should be applied:

- **Relatedness:** The universal need to interact and be connect with others.
- **Competence:** The universal need to be effective and master a problem in a given environment.
- **Autonomy:** The universal need to control one's own life.

Relatedness refers to the fact that the users have a need for personal goals and be connected to a meaningful community. In order to reach a higher relatedness within the game for the users, the developers can add a *meaningful story*. This can also be added by making up a story around different activities to make them seem more meaningful and relatable.

Competence refers to the fact that we as humans have a need to master different aspects of our lives, be it games, our job, our hobby etc. A lot of video games are based around puzzles and being able to complete puzzles in a short enough time. Without knowing it the players are challenging themselves while having fun at the same time. Additionally, in a lot of games people are mastering the game while doing for example math at the same time. This is why it is important as a game designer to create interesting challenges in the game, as it can make the user learn while having fun and not necessarily knowing that

they are learning [25]. The tasks or puzzles should also increase in difficulty, making the game harder and harder as you play it, but still be possible to complete it without making the user frustrated because the challenges are too hard. This should also match the *Flow* theory, which describes a users state of immersion in an activity [6]. The flow theory is illustrated in Figure 2.1.

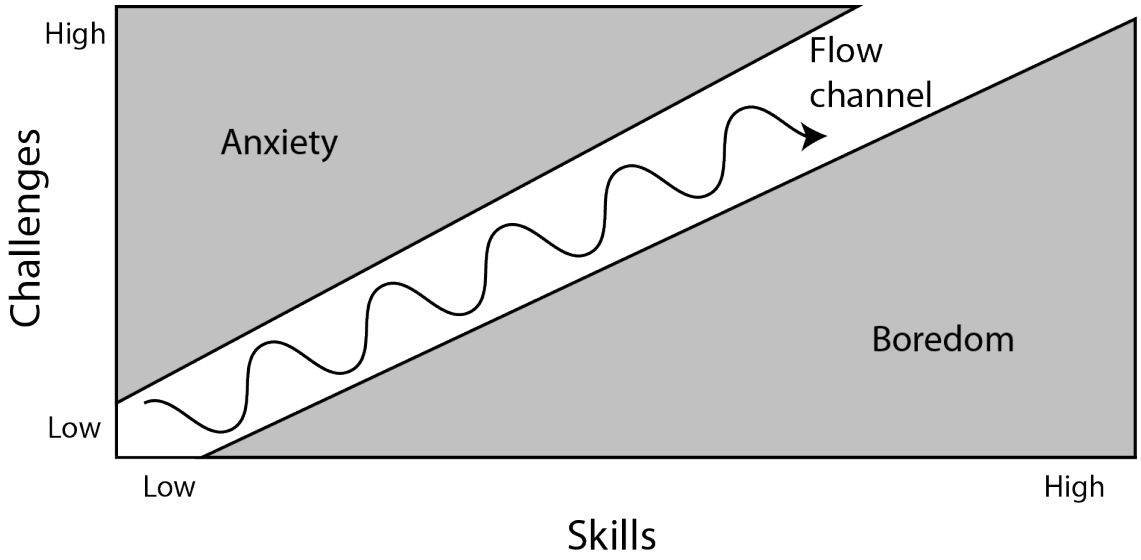


Figure 2.1. *Illustration of the Flow theory [32].*

The Flow theory describes when a person is in a state of flow, also called in the zone of immersion. When people are in "flow", they are deeply immersed in the game and their enjoyment is high. Csíkszentmihályi developed the theory [7]. He, along with his researchers, determined three conditions for flow, which had to be in place to achieve state of flow.

- One must be involved in an activity with a clear set of goals and progress. This adds direction and structure to the task.
- The task must have clear and immediate feedback. This helps the person negotiate any changing demands and allows them to adjust their performance to maintain the flow state.
- One must have a good balance between the perceived challenges of the task and their own perceived skills. One must have confidence in one's ability to complete the task. [9]

Therefore, to maintain flow there has to be a growth principle [7, 9]. When one is in flow, the person is fully immersed while trying to master the task. To maintain flow, the game has to provide greater challenges to the person, which fit the competence level of the person. If the game is overly challenging the person, he will eventually lose the flow state, because the task becomes too difficult. This zone is called *anxiety*. On the other hand, if the game is too easy, the user will be bored and lose motivation to keep playing. This zone is called *boredom*.

Lastly, autonomy is when users feel that they are making their own decisions and not being forced to do certain tasks or being controlled while playing the game. If the users feel that they are being controlled, they will lose their autonomy and the game can become a demotivating experience. This is in general found if the game presents an "if - then" reward system [8, 14].

Level	Description	Example
Game interface design patterns	Common, successful interaction design components and design solutions for a known problem in a context, including prototypical implementations	Badge, leaderboard, level
Game design patterns and mechanics	Commonly reoccuring parts of the design of a game that concern gameplay	Time constraint, limited resources, turns
Game design principles and heuristics	Evaluative guidelines to approach a design problem or analyze a given design solution	Enduring play, clear goals, variety of game styles
Game models	Conceptual models of the components of games or game experience	challenge, fantasy, curiosity; game design atoms;
Game design methods	Game design-specific practices and processes	Playtesting, playcentric design, value conscious game design

Table 2.1. *Levels of game design elements* [8, 14].

In order to help describe engaging, fun, and pleasurable experiences for users we examine the Playful Experience Framework (PLEX), which categorizes playful experiences using 22 PLEX cards [24, 27, 29] It was first investigated by Costello and Edmunds [5]. They gathered the views of game designers, philosophers and researchers to create what they called *pleasure framework* and from it they created 13 categories of pleasure. This initial starting point was later used in the creation of the PLEX framework, which focuses on experiences, pleasures, emotions, elements of play, and the reasons as to why people play [27], and extended the initial 13 categories to 22 categories as seen in Table 2.2.

Experience	Description	Experience	Description
Captivation	Forgetting one's surroundings	Fellowship	Friendship, communality or intimacy
Challenge	Testing abilities in a demanding task	Humour	Fun, joy, amusement, jokes, gags
Competition	Contest with oneself or an opponent	Nurture	Taking care of oneself or others
Completion	Finishing a major task, closure	Relaxation	Relief from bodily or mental work
Control	Dominating, commanding and regulating	Sensation	Excitement by stimulating senses
Cruelty	Causing mental or physical pain	Simulation	An imitation of everyday life
Discovery	Finding something new or unknown	Submission	Being part of a larger structure
Eroticism	A sexually arousing experience	Subversion	Breaking social rules and norms
Exploration	Investigating an object or situation	Suffering	Experience of loss, frustration and anger
Expression	Manifesting oneself creatively	Sympathy	Sharing emotional feelings
Fantasy	An imagined experience	Thrill	Excitement derived from risk and danger

Table 2.2. *The 22 PLEX categories, as well as descriptions of each category* [28].

The cards/categories can be used on mundane everyday tasks and make these tasks more

worthwhile by approaching them through some form of play or game, but can also be used when gamifying existing products, by examining each element with the PLEX cards and making it a playful experience for the users.

One of the most important things that we need to address is if the game we are developing is actually considered gamification or if it creating a new game using game design elements. The current research has shown that while gamification of a product can be done very easily through for example points and leaderboards, most researchers do not recommend this approach. Instead we need to consider how the development of a complete game could be done and then add some of the design and game elements that are normally used when creating a game. Furthermore, we need to address the concerns of replay-value and if the game is playful and engaging, as well as fulfills the three principles (relatedness, competence and autonomy) to make sure that the user experience of the game is as high as possible. Lastly, the flow theory should be considered. This can for example be done by making the game increasingly difficult to make sure that users are engaged in the tasks and find the "puzzle-solving" part of the game to be enjoyable.

Chapter 3

Design

In this chapter we describe the design of our application and games. The goal with the design is that each individual level in the application becomes a new user study. Therefore, whenever a level is added, a new design section for that level needs to be created describing:

- The experiment
- Why it is relevant (background research)
- How it is gamified
- The data logging needed
- How it fits the story and the rest of the levels

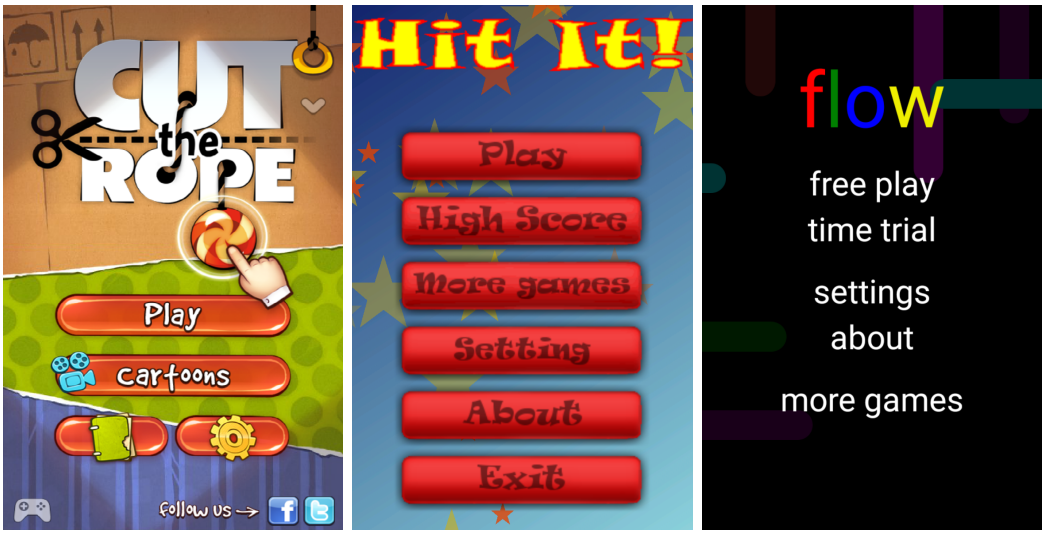
For the prototype of our application we want to develop two levels that each incorporate a user-centered study, gamifies it, and gathers data that can be used for later analysis. Furthermore, each level should consider the three principles of gamification (relatedness, competence, and autonomy) [8, 14] as these can increase user immersion, as well as the replay-value of the application and each level in it.

3.1 Requirements

Before we examine the design of each level in the application (app) we have created some requirements that it needs to fulfill or at least something that needs to be considered when implementing the different elements.

3.1.1 Menu

First of all the app needs a menu to enable different selections for the user. The menu is also there to make sure that the user can press the start button when he/she is actually ready to start the game. The design of the menu system is based on research by Jones and Marsden [21], as well as existing applications such as Cut the Rope, Hit It, and Flow Free. Examples of the menus from these games can be seen in Figure 3.1. As seen in the figures below, each menu uses a linear menu system with no icons, meaning only text descriptions of each button. Furthermore, some of them have an options/settings button which functions as a hierarchical menu system, where pressing the button will lead to a submenu, where the user is able to select between different options.



(a) Cut the Rope (b) HitIt (c) Flow Free

Figure 3.1. *Illustrations of three different menus from existing games on Google Play store.*

This means that for our menu system we will create a linear menu with at least a start/play, score, and credits/about button. As promotion of other similar games is not the goal, we will leave this button out and focus on creating a small and easily understandable menu. For future development of the app, an option could be enabled where the user can select which level he/she wants to start on after all previous levels had been completed at least once.

3.1.2 Consent

One question that needs to be answered is if the application should get the user's consent, that the data from using the app will be used as experimental research. In the app Hit It by Henze et al. [19], they present their users with a consent popup button, as seen in Figure 3.2, where the user has to press an 'okay' button in order to view the menu and eventually play the game.

They state that around 4% of users chose not to use the app due to this consent popup. This means that if we add the consent button we can also expect that at least 4% of the users installing our application will not give their consent for the data to be used in experimental analysis. However, there are a lot of questions about ethics if a button like this is not added to the application. Furthermore, we need to ask ourselves if adding a consent button ruins the entire purpose of the research, as we initially did not want the crowd to know about their participation in a study, as it might influence the data. The goal of the study is also to examine if there is a difference between participants that know that they are taking part of a study and participants who do not know it.

In conclusion, there is no real answer to this question. We have chosen to divide the crowd group into two separate groups. This means that half the crowd participants will get a consent window before the game starts, thereby informing them that they are participating in a study, whereas the other half will not. This is done to see if the consent

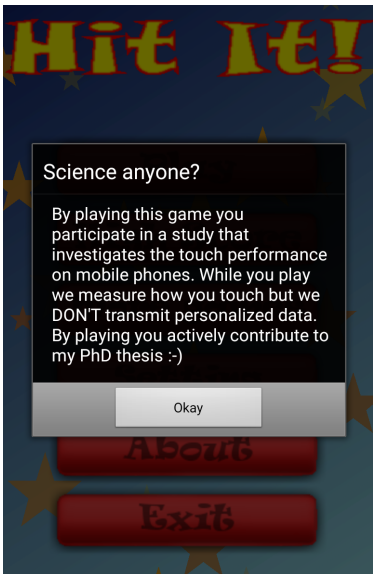


Figure 3.2. *The consent popup for the Android app Hit It [19]*

and information has anything to do with the overall performance of the participants in the study. For the rest of the report, the crowd group who receives the consent has been dubbed crowd-plus, whereas the other group is simply called crowd.

3.1.3 Data Logging

For data logging we have to examine each individual level and gather the appropriate information. This means that if we are interested in the time it takes the users to complete certain tasks, and then the time has to be logged and stored for later analysis.

Each time a level is added, the data logging has to be revisited and new information for that level has to be added. Some of the different kinds of data that can be logged for touch and tablets are for example time, score, amount of interactions (touches), where the touch points occur et cetera. In the following sections, the data logging will be described for each level in order to fit the needs.

3.1.4 Game Elements

Before we start investigating each of the levels that we want to implement, we need to examine some of the game elements that are needed for the overall game and not just the individual levels.

We have chosen to use some of the game elements from the research investigated in Section 2.2. From the research we concluded that gamifying an existing product takes much more effort than just adding a score to the product. However, as we still want users to be able to keep track of their progress and compare it to others, we have chosen to include a score and leader boards to our app as well. The leader boards can further add to the three principles of gamification, more precisely relatedness, as a community might be created around the app, which could improve the playability and make the users want to play the game again, thereby increasing the replay-value of the game [33, 35, 37]. To

further increase the relatedness for the game, we have chosen to create a main character and a story to the game. This means that for each playable level in the app, the main character has to have a significant role in the level. To further develop on this, a short story section can be included between each level in the final version of the app to make it more appealing to the players.

When examining competence throughout the game we will use the flow-theory [6] to keep increasing the difficulty of the levels, while at the same time make sure that the levels are not impossible to beat. In the following sections all of these elements, as well as other game elements used in each individual level will be described.

3.2 Levels: HCI User Studies

In the following section we will describe the design of the levels that we have made for the prototype (or alpha-version) of the game. The prototype will consist of two levels, both using the premises of gathering knowledge for HCI, and both could be tested under normal circumstances without the gamification element added to it. The point of each designed and implemented level is to gather information about a specific HCI subject using touch interaction (on either mobile phone or tablet), and comparing the data gathered from the crowd against the data gathered from an experimental setup.

3.2.1 Level 1: Drop

The first level in our application is based on an existing Android game called Drop [31]. A screenshot from the game can be seen in Figure 3.3. Drop is a simple game, where the user has to control a ball using either the gyroscope, thereby tilting the phone or tablet from left to right in order to make the ball move in either direction, or creating a touch-point by pressing somewhere on the screen and having the ball move towards that point. The floor will then move at an increasing speed throughout the game and it is now the users task to stay alive for as long as possible, to get as many points as possible. This means that the time played is equal to the points generated. Furthermore, the user can collect stars throughout the game which add points to the score as well.

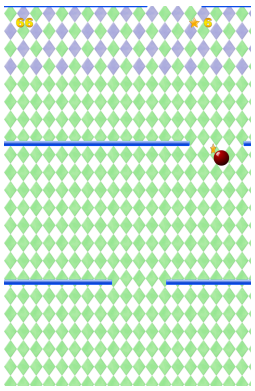


Figure 3.3. *A screenshot from the Android application Drop. The user controls the ball through the holes to earn points, for as long as possible.*

For the design of our version of Drop we have chosen to create a user study around the same elements as created for Drop. We want to create a game where we can test the performance of touch vs. tilt in a game very similar to Drop, where the performance of the users is the time/score that they are able to achieve using both types of input. Both levels will be tested on both a crowd (through Google Play), and in an experimental setup, where we control all the variables. This means that the data logging needed for this particular level is the time/score that the users are able to achieve, and the input method that they are currently using. For the score we will only implement the score based on the length (time) that they play the game and thereby remove the star points that are in Drop.

In order to follow the gamification principles, the ball will be replaced by our main character, which will have several lives in the game (3-5 depending on internal testing) in order to create repetitions of completing the level and thereby giving us more data. This is done while at the same time giving the users a story to follow and keep them occupied for a longer period of time.

A sketch of the first level can be seen in Figure 3.4. For our implementation we will also use an increasing speed for the game to relate to the flow-theory. Everything else will be based on the design of Drop, as presented above.

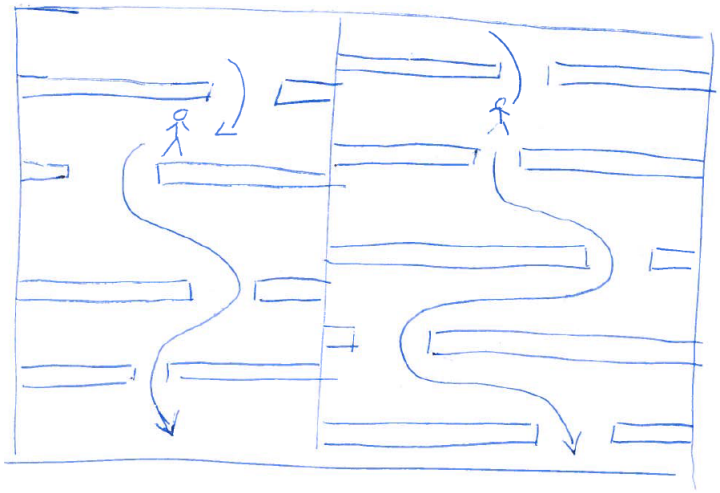


Figure 3.4. *A sketch showing the design of the first level.*

3.2.2 Level 2: Device Human Resolution (DHR)

The second and final level for our prototype of our application will be based on an existing user experiment created by Bérard et al. [4]. In their experiment they examined how devices in general are better and can be more precise than the users who are using them. Therefore they tested three different devices on 18 participants to find the resolution for each device for the humans, and the size of objects that users were able to select, with a low error margin, when interacting with each of the devices. The interface used in their experiment can be seen in Figure 3.5.

In the interface the user had to move the pointer to the starting area, press a button

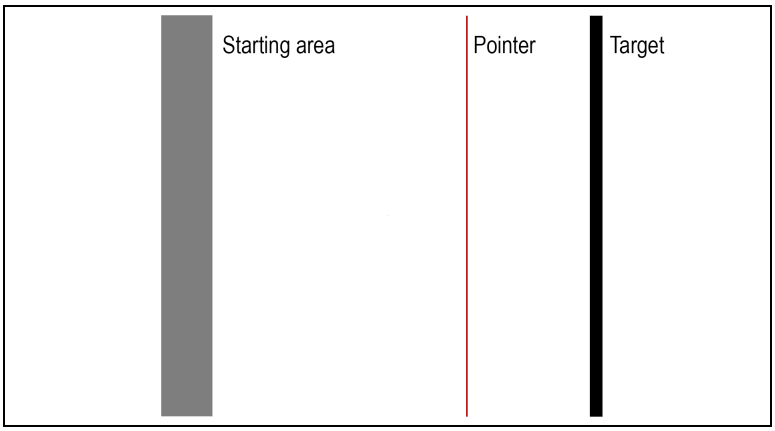


Figure 3.5. *An illustration of the interface used in the Device Human Resolution experiment. The users had to move the pointer to the starting area and then to the target to complete a task.*

which creates the target 1000 pixels from the current position of the pointer. The user then had to move the pointer to the target and when they were inside the target area press a button. This task was completed seven times, with seven different target sizes that where always presented in descending order of the target size, meaning the targets got thinner and thinner. The size of the smallest target was 4 pixels wide, which is the same width as the pointer. If the user pressed the button while not inside the target area with the pointer, it would count as an error, and the user then had to move the pointer to the starting area to try with the same target size. If the user made three errors for the same target size, they would move on to the next target size (a smaller target) and have three attempts to complete that task. The users had to repeat the seven target selections 20 times, which meant a minimum of 140 target selections (excluding errors) [4].

For the design, the first choice we made about this level was that it had to be the second level for the app. This is done to make sure that the interaction from this level does not influence the preference towards one of the interaction methods in the first level. Secondly, we decided to remove the pointer from the original implementation, as it was mainly there to illustrate a cursor for the users. When the interface is created on a touch surface, the user does not need the cursor as he/she can just click directly on the starting field, lift and move the finger, and then again click directly on the target. We did however consider one option where we had a visual pointer move from the touch-point in the starting area towards the target at a fixed speed. This idea was however discarded, as the task would change too much from the original experiment, and it would now be an experiment of doing a touch-point at the right time instead of finding the DHR for the users.

For the gamification of the experiment we have chosen to create a scene where a monster needs to be shot down in order to move on to the next level. A sketch of the scene can be seen in Figure 3.6. The main character is again in the scene, in order to create a story behind this character and to keep a clear goal when adding levels.

The point of the game is that the user has three shots (or attempts) to shoot down one size of a monster's shield. If one shield size is hit, the shield becomes smaller and the user then has three new shots to hit the shield. The shield will get smaller and smaller, and as

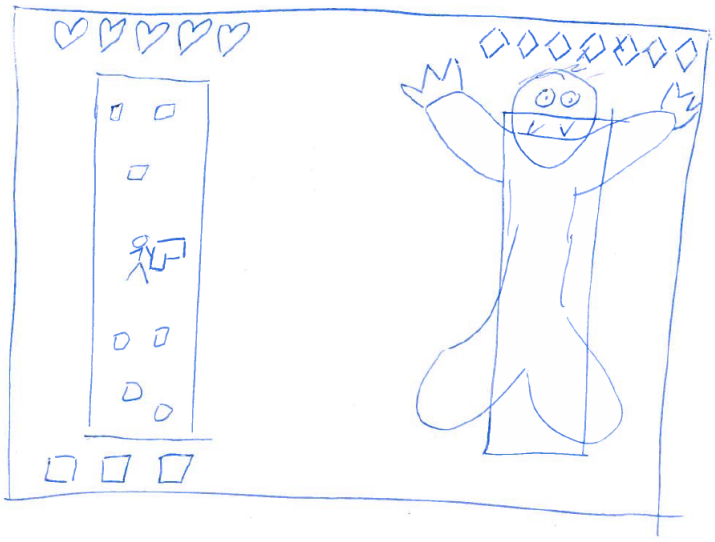


Figure 3.6. *A sketch showing the design of the second level.*

soon as seven sizes of the shield have been destroyed, the monster dies. If one monster is killed, a new spawns and the task starts over with all seven sizes of the shield intact. This is also illustrated in Figure 3.6 with the shields and shots in the top of the image. Lastly, if the user uses three shots that miss the shield, the monster will hit the user, causing him to lose a life, but still destroying one size of the monsters shield. The user will have a set number of lives (gathered from an internal test), and if all lives are lost the level is over. This means that the experiment will be repeated for as long as the user is able to stay alive, which will cause him to get a higher score, and us to get more data from that user.

The size of the shields will be based on the original DHR experiment, where Fitts' IDs were used and the target sizes were presented as ticks, where the smallest was one tick and the widest was 32 ticks. Each tick was equal to four pixels which means that that smallest target is four pixels wide and the largest is 128 pixels wide. A sketch of the shields used in our app can be seen in Figure 3.7. The middle of each shield will be created in accordance to the original experiment, where they used a distance of 1000 pixels from the touch-point on the starting area on a 1920x1080 display.

For the data logging during this level, we need to track the time between pressing the starting area and touching the target. Furthermore, we need to track the amount of errors that users make during each of the different target sizes, as both of these variables were tracked and used for analysis during the original DHR experiment.

One thing that we have to consider during our implementation is the fact that we are using touch, whereas the original experiment used three different devices. A complication when using touch compared to other input methods is the fat finger problem [2, 34]. When using the mouse, users are able to select elements that are very small, however, when interacting through touch the user's finger is the primary interaction tool. When touching a device a relatively larger area of the finger comes into contact with the surface, meaning that if the size of the fingertip is wider than the virtual object they are touching, then the system might not register the touch-point inside the object. This is especially an issue because

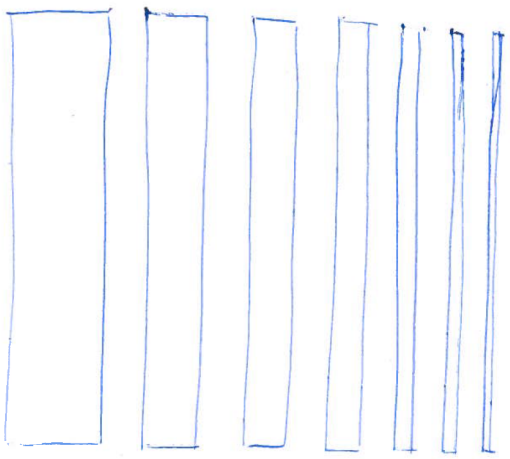


Figure 3.7. *A sketch showing the width of the shields from the thinnest to the widest. The shields are always presented in descending order.*

all existing touch platforms use a single point within this area to do the hit testing. We therefore need to address if the fat finger problem is too big of an issue when the users have to interact with the smaller target sizes.

Chapter 4

Implementation

In this section, it will be possible to find descriptions and flowcharts of the creation of "Minigames with Roboto". Development of the game happened in Unity 3D version 5. Graphics either is from the Unity asset store or self-created. The section is divided into categories of main menu, the first level "Drop", the second level "Wall destroyer", video tutorials and lastly marked release of the app "Minigames with Roboto".

4.1 Menu

This section, describes how the menu is build and how it communicate with the server as soon as the user starts the game. The user is presented with an app icon in his app launcher before launching the game. See Figure 4.1.

Figure 4.1. *Image showing the app icon used for the game.*

This icon should be easily recognizable by the user, and somewhat describe the application. Therefore, we decided that the icon should include, besides the name of the game, the main character "Roboto" and the biggest thread in the first level of the game. When the game has been launched, the user is presented with the game menu: See Figure 4.2

The game menu let the user go to different destination depending on his desire. However, in the background the game check if the user already have an ID or if it should receive an ID from the server. If the user do not have an ID, the game connects to the server to receive an ID. Further if the ID is an odd number, the user is presented with a consent window telling the user that he is testing the game and we as researchers are gathering data, while he is playing. See Figure 4.3.

Figure 4.2. *The menu screen as shown to the users.*

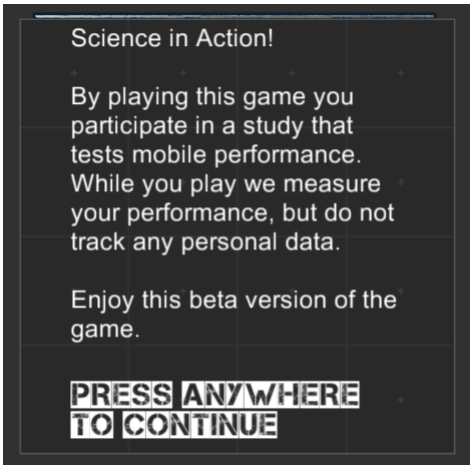


Figure 4.3. *The consent window, which was shown to all users with an uneven ID (also called crowd-plus). The window was presented on top of the menu.*

When the user is done playing the game a similar window will open and redirect the user to an online questionnaire. In Figure 4.4 a flow chart illustrating the process that users go through when first opening the game menu.

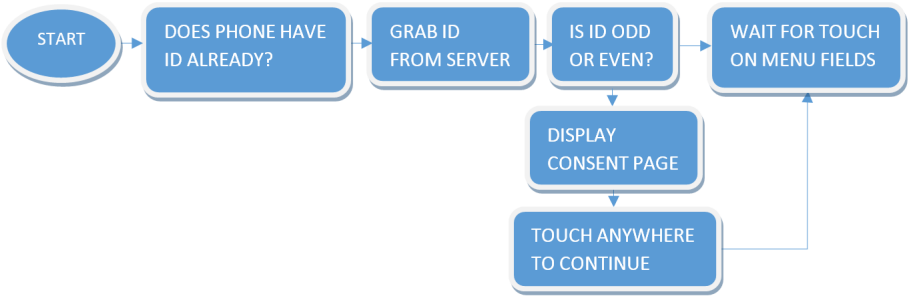


Figure 4.4. *A flow-chart illustrating the implementation and working process of the app and menu.*

4.2 Level 1 - Drop

This section describes map generator, how the character is controlled and lastly, an explanation how the sound- and score system works.

4.2.1 Map Generator

During the development of the level, the map iterated from one type of static map to an automatic generated map. The first iteration was a static map with an actual ending. Therefore, the difficulty level of the map were determined to kill the user before the user reached the end of the map. This meant the map were somewhat short and a small pool of test participants reached almost same depth before dying. Therefore, time of death did not vary. In the second iteration of the game, the map automatically extents when the user is reaching certain depths. This means that the user in theory can play forever.

The map was not randomly generated in order to maintain control of the perceived experience for pretest of Touch control versus Tilt control. For the final experiment, the map was identical to the pretest map, and not randomized, because the level should be the same for the crowd and the lab subjects.

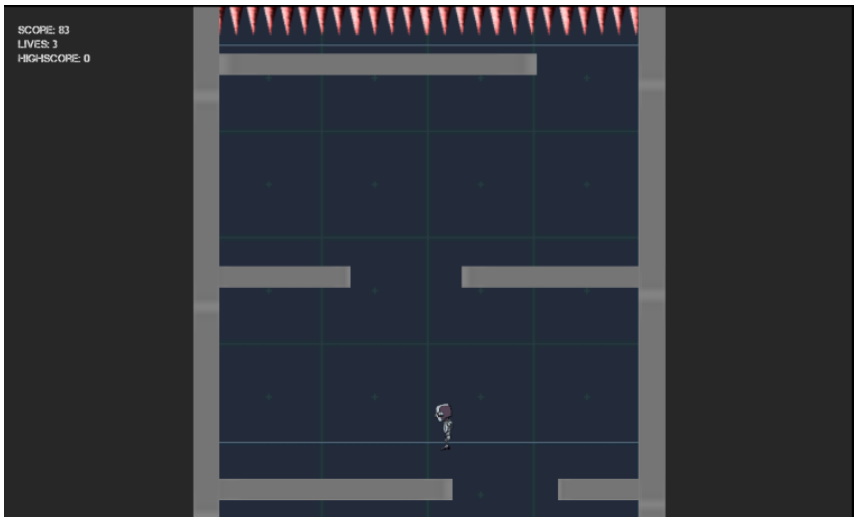


Figure 4.5. *The map and scene used in the first level Drop.*

A part of the map can be seen in Figure 4.5. The texture for the map is in a dark grey theme, and follow the characters robotic theme. The spikes in the top is the deathly bar that will end the game if the character touches the spikes. The spikes is colored red to indicate danger. The score system in the top left corner show current score, lives and high score.

A flow chart of the map generator for Drop can be seen in Figure 4.6.

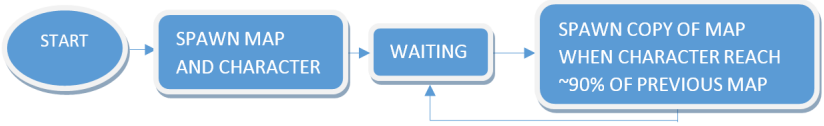


Figure 4.6. *A flow-chart showing the implementation of the map generator for the first level Drop.*

4.2.2 Character and Camera Control

The camera for the drop game is controlling the level of difficulty in the game. If the user is moving faster than the camera, the cameras speed will increase until it catch up with the user. However, if the user is too slow the camera will move beyond the user and eventually the user will touch the spikes. If the user touches the spikes, the user will lose a life and the level will restart. Before the user starts moving, the camera is in standstill until the user is ready. A flow chart illustrating the camera controls for Drop can be seen in Figure 4.7.

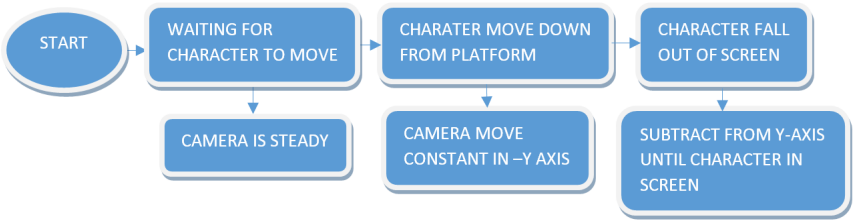


Figure 4.7. *A flow-chart illustrating the implementation and process of the movement of the main camera for the Drop level.*

For the pretest two type of input control were developed, touch points and tilt input. The touch control will add a speed in the positive or negative direction depending on left side or right side touch point on the screen. When the screen receive a touch point on either side, the character will further flip to the touched side and the character will shift from an idle animation to a walking animation.

For tilt control, the amount of degrees the phone was turned had a direct impact on the amount of speed the character had. In an earlier design, the character when falling used gravity this had some implications that the character was able to be stuck in walls further that at some point the camera will move faster than the character is falling which automatically led the user to death. For the finale design, the user will always fall with a constant speed, which solved the problem of getting stuck in walls as well. Further, the constant speed will increase slowly over time, so the speed always is higher than the camera moving speed. The running speed of the character also increased over time. The controls of the character in Drop can be seen in Figure 4.8.

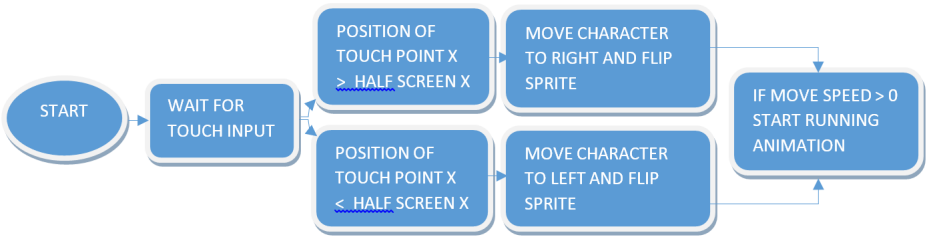


Figure 4.8. *A flow-chart showing the movement of the main character for the Drop level.*

4.2.3 Sound

In the Drop level the sounds used is minimalistic meaning there is only a death sound and background music called “Snowdaze” by Airtone. The death sound plays as soon the

character touches the spikes. The background music plays when the level start, and the music is continuous, meaning when the music ends it will start over throughout the level. A flow chart illustrating the implementation of the sound during Drop can be seen in Figure 4.9.

Figure 4.9. *A flow-chart showing the sound system for the Drop level.*

4.2.4 Score

For the drop level, the score system is simple. The score is directly linked to the y-axis of the character. This mean when the character is standing on the starting area the character will have a y coordinate of zero. When the character jump down to the first floor his y coordinate will go to minus four, which automatically converts to positive numbers only and therefore, receive a score of four. When the character reach e.g. a y coordinate of minus 800 the score will be 800 and so forth. A flow chart showing the implementation of the score can be seen in Figure 4.10

Figure 4.10. *A flow-chart illustrating how the score system functions during the Drop level.*

For the Drop level, the server will only receive data when the user die in the level. The data the server receive is ID of the user, the level name (used for pretest in particular to see if they played Drop with touch or Tilt) and for every entry, their score and time survived was recorded.

4.3 Level 2 - Wall Destroyer (DHR)

In this section, a description of the second level, called “Wall Destroyer” is available. It describes how the user drags from the starting area to the target area. An explanation of the structuring of the map, the animations within the level, the sound- and score system and the data logging.

4.3.1 Drag

The user have to touch the starting area (The green field), when the user is touching the area the target will spawn 24 mm away from starting area. When the area is visible, should the user drag his finger from the initial starting point to the target area and then release the finger from the screen. When the participant is dragging across the screen there is invisible object following the finger that functions as a hitbox, when the box totally covers

the target area and the finger is not touching the screen. Will the system check if the user hit or missed the target. A flow-chart of the drag can be seen in Figure 4.11.

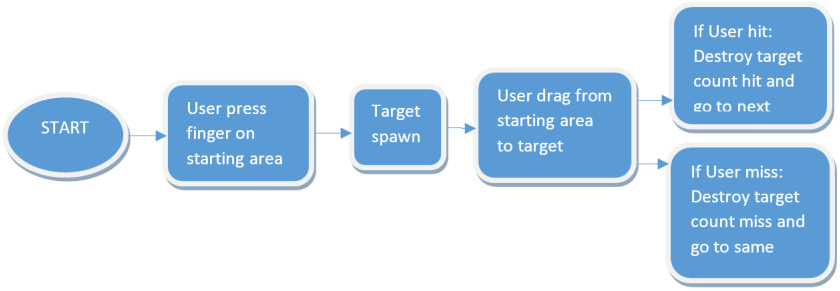


Figure 4.11. *Flowchart showing the dragging function used for the Wall Destroyer level.*

4.3.2 Map

As mentioned before the map consist of a starting area (green field) and a target area (grey wall). The target area gets smaller and smaller the more the user proceed for each killed wall segment. The background is a static grey background in the theme of "Roboto". See Figure 4.12

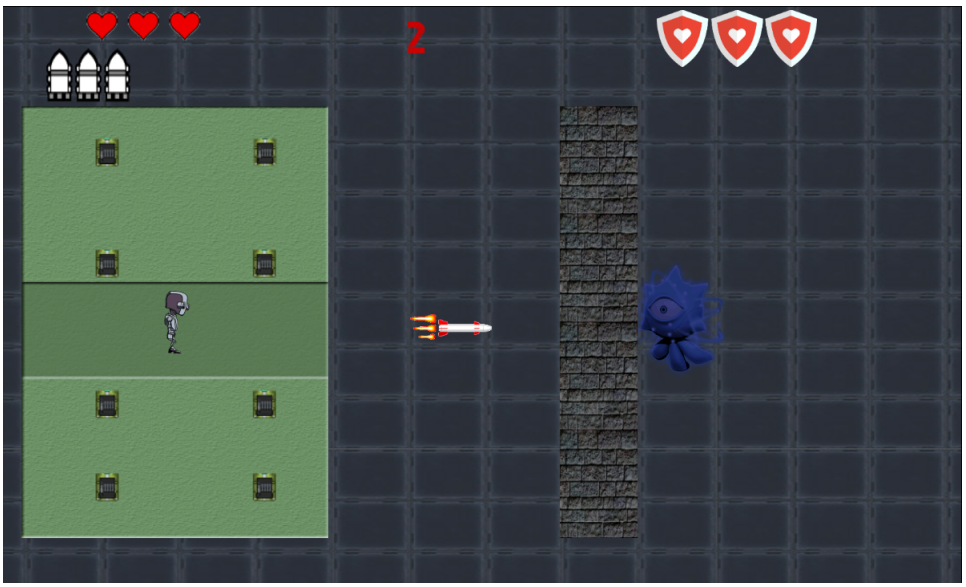


Figure 4.12. *Image showing the map used for the Wall Destroyer level.*

4.3.3 Animations

There is several animations in the wall destroyer level. When hitting the target area, the character fires a missile against the wall and the wall will disappear. Further when the missile hit the wall, an explosion animation will play. However, if the user miss the wall three times, the monster will shoot a missile (different kind) against the user. When the missile hit the user, an explosion animation will play. See Figure 4.13

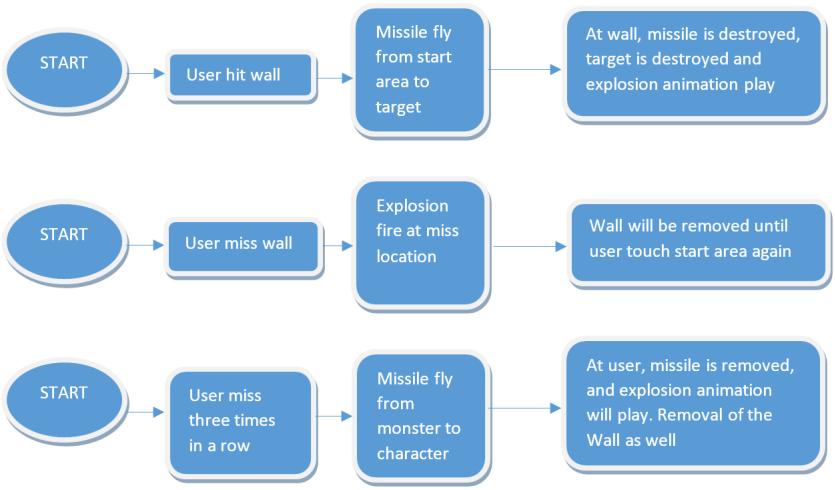


Figure 4.13. *Image showing the map used for the Wall Destroyer level.*

4.3.4 Sound

The wall destroyer level use the same background music as used in Drop level. However, compared to drop level, more feedback that is audible is giving to the user. When hitting the target an explosion sound plays (When wall collide with missile). When the player loss a life (When user collide with missile) the explosion sound is played and as well as the players death sound. When the monster lose all the shields, an alien death sound will play. Lastly, when the user miss a target a "dunk" sound will play that indicate a miss. See Figure 4.14.

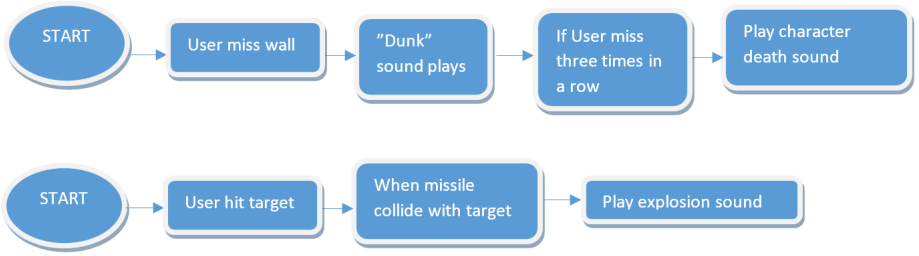


Figure 4.14. *Flowchart showing how sound playback work for wall destroyer level.*

4.3.5 Score and Lives

The scoring system for this level, count how many times the user destroy all seven shields (top right corner) or as explained to the user the number of monster killed (Shown in top center). The lives system work in two fold, for every target the user have three tries, indicated by ammo icons. Besides the ammo, the user have five lives indicated by hearts. If the user miss the same target three times (Use all his ammo), he will lose a heart. The Ammo system functions as retries for the user in order to maintain a high data amount, e.g. in case of a user will miss different target five times in a row, which end the game and will give almost non-data. See Figure 4.15.

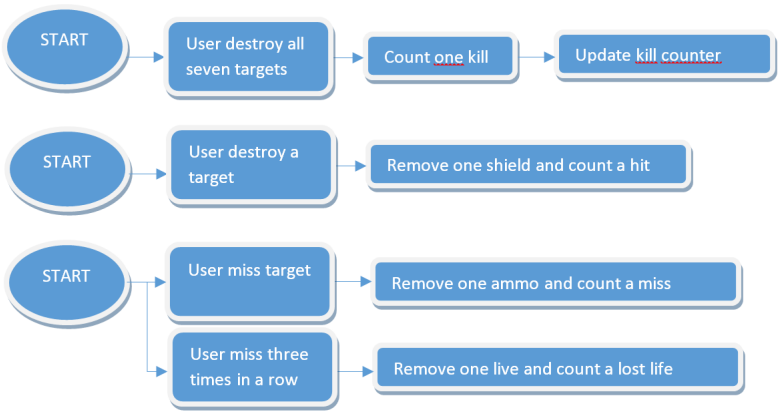


Figure 4.15. *Flowchart showing how score system work for wall destroyer level.*

4.3.6 Data logging

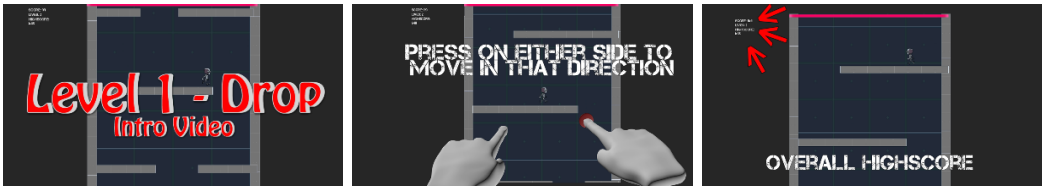
Logging of several types of data run in the background. Since the experiment is the original DHR experiment, the same variables is measured or logged [4]. This mean logging when users hit a target with movement time in seconds from starting area to the target area. If the user, miss a target it is logged as a failure and how long it took. Further amount of tries on each wall is recorded this mean it know if a user only tried once or twice on each wall segment. Naturally, the ID is uploaded to identify the user on the server.

4.4 Video Tutorials

Videos explain what the user should do to complete the different levels. This was to ensure users complete the level in the same way and to limit the amount of error prone data. In this section, an explanation follows for how the tutorial videos was created.

4.4.1 Drop

Before the level start, a short tutorial video presented to the user, it show how to play the level. This video show the user how to control the character, which type of object he should be aware of (the red bar). Further, the video describe the current score, amount of lives and the high score. The video last 24 seconds. It shows a user playing the map while text and arrows explain the above mentioned. Three screenshots from the video can be seen in Figure 4.16



(a) Intro (b) Touch Interaction (c) Scoring system

Figure 4.16. *Screenshots from the tutorial video for Drop.*

4.4.2 Wall Destroyer (DHR)

Right after the Drop level the tutorial video for the DHR level will start playing. The video last 31 seconds and explains dragging a finger from the green area to the target area enable to destroy targets. Further, it tell about the lives and ammo system and how target destroyed affects the monsters shields. Lastly, it show how the current score work when killing a monster. Therefore, just as in the Drop level a user is playing the level for 31 seconds in the video while text and arrows explain the above mentioned. In Figure 4.17 and 4.18, six screen shots can be seen from the tutorial video for DHR.

(a) Intro (b) Drag (c) Walls

Figure 4.17. *Screenshots from the tutorial video for Wall Destroyer (DHR).*

(a) Bullets (b) Lives (c) Score

Figure 4.18. *More screenshots from the tutorial video for Wall Destroyer (DHR).*

Chapter 5

Preliminary Test

In this chapter we describe our preliminary test, which was conducted before the final experiment. The test was primarily a technical test, which meant that we were looking for bugs in the system as well as improvements for both levels, before the release to the crowd. Furthermore, we wanted to reduce the amount of data gathered in the final study for the Drop level, by gathering data on both controls (i.e. tilt and touch) and analyzing which control method was the preferred and most efficient by the users, and removing the least efficient one for the final study.

5.1 Participants and apparatus

6 voluntary participants were recruited from the university campus. The participants were all male students in the age range of 24 to 27 years old ($M = 24.83$, $SD = 1.17$). All participants reported to be daily users of touch devices, and with the exception of one all reported to play computer games for an extended period a week ($M = 18.83$, $SD = 11.23$), essentially making them gamers with experience in this line of work.

The test was conducted on a LG Nexus 4 smartphone, with a 4.7 inch display and a 768x1280 resolution. Participants were allowed to hold the phone as they pleased to achieve the most comfortable interaction pose. Some participants also put the phone on the table in front of them.

5.2 Procedure

Before the test, each participant was given a short demographic questionnaire (age, profession, touch interaction, and gaming habits) followed by a short introduction to the test. In the introduction to the test, the main purpose was also revealed to the participants in order to make them look for bugs in the game and be able to evaluate the application from a technical point of view. After the introduction the participants were presented with the first level of the app, namely Drop. The order of the controls for Drop was counterbalanced, meaning that three of the participants would start with tilt, and the other three would start with touch. All participants understood that the goal was to stay alive for as long as possible. When a participant had completed both controls for Drop, they moved on to the second level, DHR, where they had to destroy the monster as

many times as possible. After completion of the DHR level, we conducted a short semi-structured interview in order to further get the participants' opinions on various parts of the application.

The questions asked during the semi-structured interview were:

- How did the interactions in Drop feel?
- Was the game too easy / hard?
- Did you encounter any bugs?
- Was the sounds / background music distracting?
- Which control method did you prefer (tilt vs. touch)? And why?

The same questions were asked for DHR, with a change to the last question where they were asked if they used one or two fingers for the interaction instead of the control method.

For Drop, we measured the time for both control schemes. For DHR, we measured the time it took to destroy each wall, as well as the amount of errors that users made for each wall, during the level. The entirety of the test took approximately 15 minutes.

5.3 Expectations

For Drop, we expect the survival time to be somewhat similar for both controls, however, we expect touch to outperform tilt, as users have more experience with using touch than tilt. Furthermore, the control scheme for touch enables usage in every setting, whereas tilt needs more special conditions to function the best. An example of this can for example be playing with the phone lying on a table. You can still play the touch version, however, in order to play the tilt version you need to have the phone in free space, thereby limiting the control scheme. Because of this, we also expect that the users will prefer the touch control over the tilt.

For DHR, we expect the results to be too good to actually be comparable to the original DHR experiment. The reason for this is that we realized right before we ran the test that users were able to use two hands while interacting with the game, which basically makes Fitts' law obsolete. Therefore, we chose to change this for the final study, as described in Chapter 4, and made the interaction a drag from the starting area to the targets instead of pointing. This also resembles the way the interaction is done in the original paper more, as the mouse, freespace device etc. also drags the cursor from the starting area to the target and not just points at it. This all means that the results from the DHR part of the test are of less importance, and the part is mainly included in the test to find bugs and improvements to the level before releasing the application.

5.4 Results

In this section we present our findings and results from the pretest. The results have been divided into three sections, namely Drop, DHR, and Qualitative Results. The chapter is ended with a partial discussion and conclusion to summarize the results shown here, and

what implications the results from the results will have on the application for the final user study.

5.4.1 Level 1 - Drop

For Drop we investigated the time that users were able to stay alive in the game. This means the higher the time the better, as it means that users were able to stay alive for a longer period of time with that control method. Through a paired t-test we found that the control scheme significantly affected the time ($t(17) = 3.0988$, $p = 0.007$, $r = 0.6$). Examining the mean values and the box-plot seen in Figure 5.1 for the time values, we see that Touch ($M = 126$, $SD = 11.55$) is higher than tilt ($M = 118.5$, $SD = 6.37$).

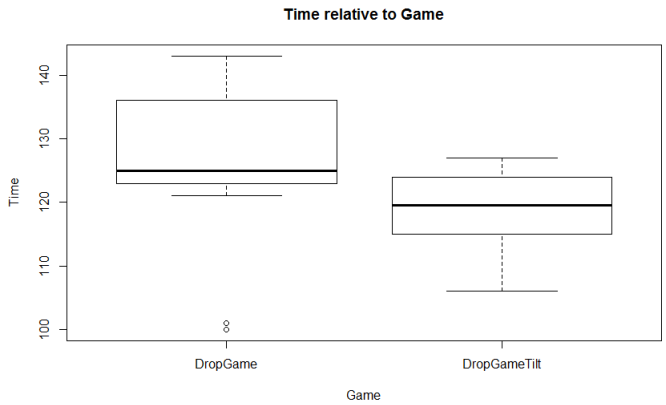


Figure 5.1. *Image showing a box-plot for the first level Drop, with Game type (control scheme) on the x-axis and the time on the y-axis.*

Examining the time data even further, we can find differences between the individual users. Using a Friedman ranked sum test, we found that there was overall no significant differences between the time data of the users ($\chi^2(5) = 8.86$, $p = 0.11$), however the post-hoc of the Friedman test shows that there was a significant difference between some of the individual users. The data from the post-hoc can be seen in Table 5.1.

ID	Rank
1	A
2	A
3	AB
4	BC
5	BC
6	C

Table 5.1. *Table showing the rankings from the Friedman Ranked Sum post-hoc. The ranks show no significant difference if the letters are the same.*

A box-plot illustrating the distribution of the data can be seen in Figure 5.2.

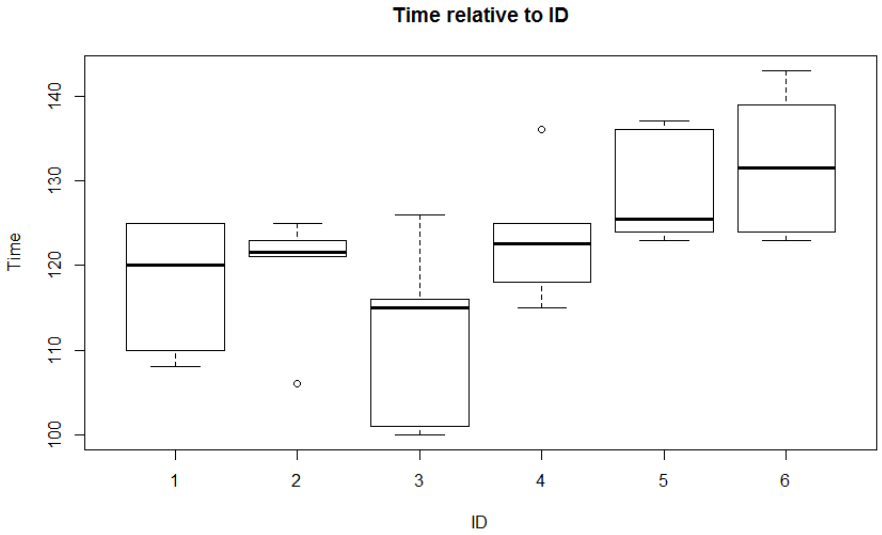


Figure 5.2. *Image showing a box-plot for the first level Drop, with user ID on the x-axis and the time on the y-axis.*

5.4.2 Level 2 - DHR

For DHR, we investigated the time between hitting the starting area and the different targets. Furthermore, we examined the amount of errors that users made when hitting the targets. Users were allowed to make three errors before moving on to the next target, as this was similar to the original experiment. As mentioned earlier, the experiment cannot be compared to the original experiment as there was actually no travel distance because the users were able to use two hands, one to hit the starting area, and the other was used to hit the target. This means that the time data will be too good to be compared to the data from the original DHR experiment, however, we still show and analyze the results to make comparisons to the results in the final user study. The results from the DHR part of the test can be seen in Table 5.2, where the overall slope value is 0.06.

Target Size (mm)	Fitts' ID	Mean Time (sec)	Failure Rate (%)	Mean Slope
32	1.32	0.88	2.74	
24	1.58	0.49	0	1.59
16	2	0.4	0	0.49
8	2.81	0.36	0.70	0.14
4	3.70	0.44	4.73	-1.13
2	4.64	0.55	23.33	-1.57
1	5.61	0.92	54.47	

Table 5.2. *Table showing the values of the DHR part of the pretest.*

The time data for the DHR part of the pretest can be seen in Figure 5.3.

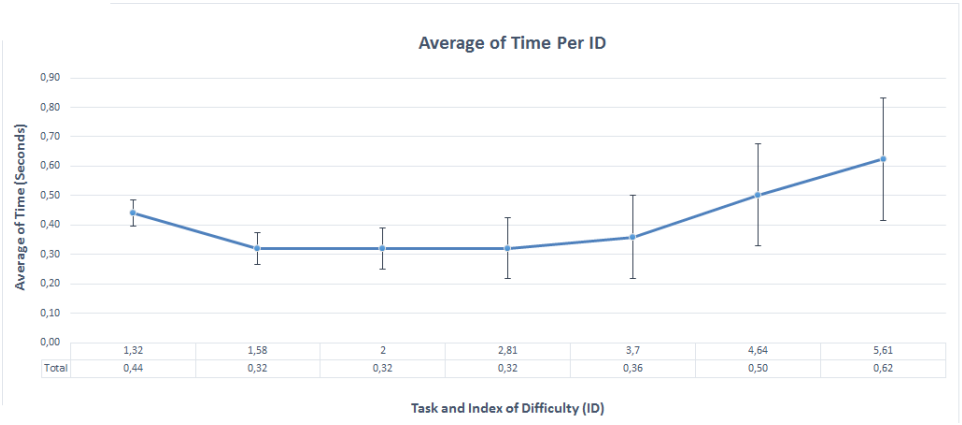


Figure 5.3. *The average time used in each task difficulty for the DHR part of the pretest. As it can be seen from the standard error bars, the variability of the results increases at the highest difficulty level.*

In Figure 5.4 an image of the mean slopes for each point can be seen, as well as the overall slope and how much the mean slopes deviate from the overall slope.

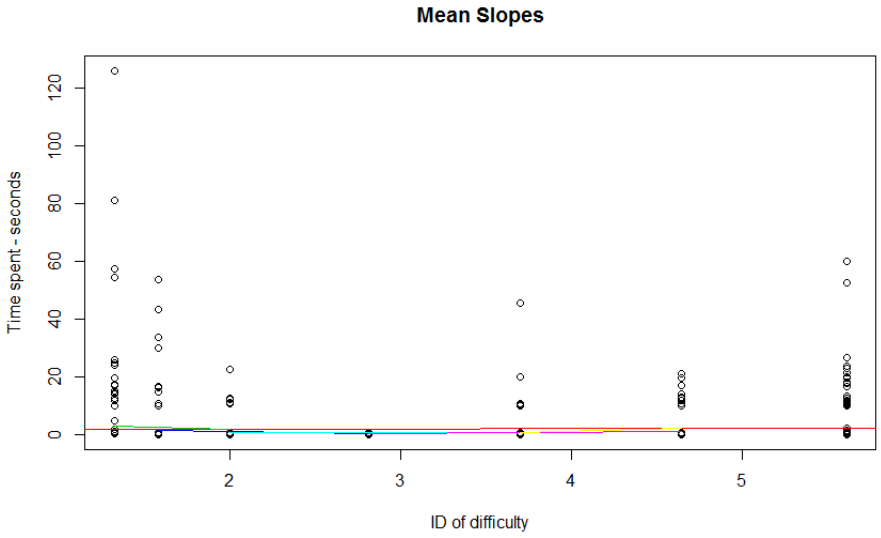


Figure 5.4. *Illustrates the mean slopes for each index of difficulty. Each mean slope was calculated using the above and below lying data point, therefore there are no mean slope for the first and last point. The red line in the slope is the overall data, whereas the remaining are the mean slopes for the data points. As it can be seen the individual mean slopes lies close to the overall mean slope.*

The percentage error data for the DHR part of the pretest can be seen in Figure 5.5. As the error data was not completely ruined by the setup of the pretest, we can still analyze the data. By looking at the graph we can see that there is a great increase (spike) in errors between Task 5 and 6 (Fitts' ID of 3.70 and 4.64) and the increase is even larger between Task 6 and 7 (Fitts' ID 4.64 and 5.61). From the graph we can also see that it becomes very difficult for users to select targets in Task 7, or with a width of 1mm.

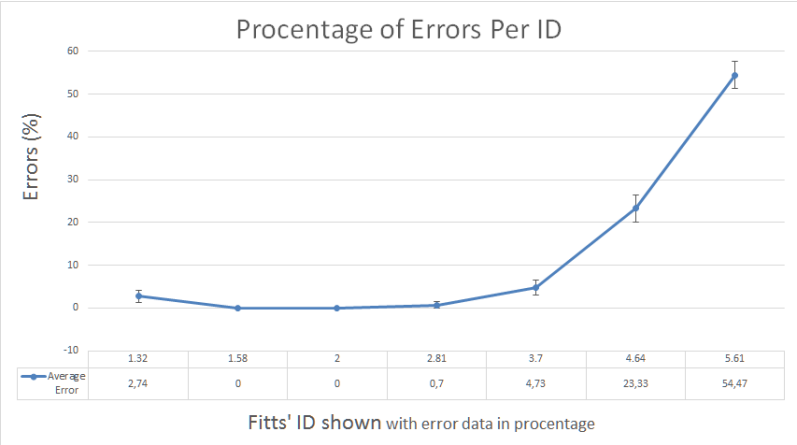


Figure 5.5. *Failure rate presented in percentages.*

Using a Friedman Ranked Sum test we can see that there are significant differences between tasks in terms of errors ($\chi^2(6) = 30.96$, $p < 0.001$). The post-hoc of the Friedman test ranks the data as seen in Table 5.3. The data is consistent with the analysis seen above, as we can see that the jump in errors happens between Task 5 and Task 6.

Task	Rank
Task 7	A
Task 6	B
Task 5	C
Task 1	CD
Task 4	CD
Task 2	D
Task 3	D

Table 5.3. *Table showing the rankings from the Friedman Ranked Sum post hoc. The ranks show no significant difference if the letters are the same.*

5.4.3 Qualitative Results

For the qualitative results we examine the answers to the questions from the semi-structured interview, as well as some of our observations during the test. For the preferred interaction method for Drop, 5 out of 6 participants (83%) preferred to use touch to control the character during the game.

Furthermore, for DHR we observed and users responded that all of them used two finger/hand interaction, which again proves that Fitts' law is not applicable for the pretest.

Chapter 6

Main Experiment

The main experiment for our project consisted of different parts. First of all, as described in the pretest in Section 5 we limited the Drop level to only include one control method, i.e. touch, in order to limit the amount of data gathered in the main experiment. Furthermore, the test was divided into testing on three different user groups and analyzing the results and comparing the groups.

6.1 Participants and apparatus

As mentioned above, the test was divided into three groups. Each of these groups had participants with different knowledge about the application to see if the results changed because of this information. As the information the participants would receive about the app was reduced, the amount of information we could gather about the participants was also reduced.

The first group consisted of 16 participants in a lab environment (hereafter known as the lab group). All the participants were volunteers and from the WEIRD population. They were in the age range of 22 to 26 years old ($M = 24$, $SD = 1.5$) and all of them were male. This group knew that they were participating in a study, where their results mattered and that their data would be recorded and used for later analysis.

The second group was the crowd group. It consisted of 19 participants. This group knew nothing about the experimental part of the application and was only presented with the app as a normal video game that they could complete. As we wanted to avoid all kinds of communication with this group we know nothing about the group as well. There was no interaction between us and the group during their testing. We know nothing about the environment they were testing in either.

The third and final group was what we have named the "crowd-plus" group. It consisted of 14 participants under the same condition as the crowd, namely that they were downloading an app and testing in their own environment and in their own time. The difference between the second and third group however, is that the crowd-plus group was informed that they were participating in a user study before playing the application for the first time, and was also requested to fill out a short questionnaire after they had completed the two levels. The participants were in the age range of 22 to 57 years old (($M = 28.16$, $SD = 9.52$) and

four of them were female.

For the lab, the app was played on a LG Nexus 4 smartphone, with a 4.7 inch display and a 768x1280 resolution. Similar to the pretest, participants were allowed to hold the phone as they pleased and interact with the game in the most comfortable position, as this made it comparable to the environments of the two crowd groups.

For the two crowds the participant would be using their own device to complete the experiment. We received data from the Google developer console after the testing period between 20th of April 2015 and 1st of May 2015. The data revealed that the app was installed 49 times, but we have only received data from 33 users. The data we have about our two crowds: 46.93% used Android 5.0 or newer, while 53.07% were using Android 4.4 or older, with 2.3.3 as the oldest. We know that 44.92% used unknown devices, 34.68% used a Samsung device, 10.2% used Sony devices, 6.12% used LG, and lastly 4.08% used HTC. For countries, we know 81.63% come from Denmark and 10.20% is from USA. The remaining 8.17% come from other countries.

6.2 Procedure

The procedure for the lab part of the experiment was similar to that of the pretest. The participants would first be given a short demographic questionnaire. After they had filled out the questionnaire, they were given a very short introduction to the test. There was no information about the actual test in the introduction, in order to make the data and user learning as comparable to the crowd data as possible. After the introduction, the smartphone was given to the user, who was asked to start the game and follow the instructions. There were introductory videos for both levels of the application for both the crowd and the lab. Each video was about 30 seconds long in order to try and introduce the participants to the two levels. After the video was complete, the users had to first play the Drop level, followed by the intro video for DHR and then the DHR level. After completing both levels in the lab, the users were given a short questionnaire. As mentioned above, the crowd version of the app would present half the users with a note that they were participating in a study before they played the game (this was the crowd-plus group). These users were also redirected through a link to a questionnaire after completing the DHR level.

6.3 Expectations

As mentioned earlier, the goal for the experiment is to examine if there are significant differences in performance between testing on users in a controlled lab environment versus testing on users in an uncontrollable environment (i.e. the users own environment). Furthermore, we want to examine if there is a difference between informing a crowd that they are participating in a study compared to not providing them with any information on the matter, and making them believe that they are just playing a game on their touch device. One thing that should be noted before examining the results is that we are not able to say if the differences in results are due to the differences in the environment or due to user differences, or a combination of the two.

For Drop we expect the survival time to be significantly higher for the users in the lab environment than the users from both crowd groups. This is primarily due to the overall hypothesis for the experiment, namely that data acquired in a lab environment will be better than the data acquired from a public user environment. Furthermore, we expect that the crowd-plus group will have a significantly higher survival time than the crowd group, due to the overall expectation that users will try harder if they know that they are participating in a study.

For DHR we expect a similar trend as seen in Drop. We expect both the time and errors used to be best for the lab group, followed by the crowd-plus group, with the crowd group getting the worst score for both variables. However, we do expect the DHR to be similar for all three groups, with the overall DHR for touch to be comparable, but worse than the DHRs found for the three devices in the original experiment by Bérard et al. [4].

6.4 Results

As mentioned above, the results have been divided into the three different participant groups, which thereby enables us to analyze and evaluate the groups individually and together. For all tests we have used a significance level of $\alpha = 0.05$.

6.4.1 Drop

For Drop we investigated the time that the users were able to stay alive in the game. This means that the higher the survival time, the better.

Through a one-way Analysis of Variance for Independent Samples test (ANOVA), we found that the groups significantly affected the survival time ($F_{2,286} = 23.48$, $p \ll 0.001$). The results is shown in a box-plot in Figure 6.1. Using a TukeyHSD post-hoc test, we find that there are significant differences between all groups, with lab being significantly better than the others ($p \ll 0.001$), while crowd-plus is significantly better than crowd ($p = 0.02$)

However, after we ran the test, we examined the data more closely and found a lot of very low time entries for both the crowd and crowd-plus data, which skewered the data in that direction. In order to get a closer comparison to the lab results, we decided to examine the results again after removing all instances below a time of 70. An ANOVA for the new data set still shows an overall significant difference between the groups (($F_{2,187} = 7.60$, $p < 0.001$). A box-plot illustrating the results can be seen in Figure 6.2. Running a TukeyHSD post-hoc test shows that there was no significant difference between the lab and crowd-plus groups for the new results ($p = 0.95$). Furthermore, there was still a significant difference between lab and crowd ($p < 0.01$) and also between crowd-plus and crowd ($p < 0.01$).

6.4.2 DHR

For the results of DHR we examine the groups independently to find the DHR for each group. If nothing else is stated, the used test method is ANOVA with a TukeyHSD as post-hoc. For the completion time we analyzed the deviation of data from the Fitts' model

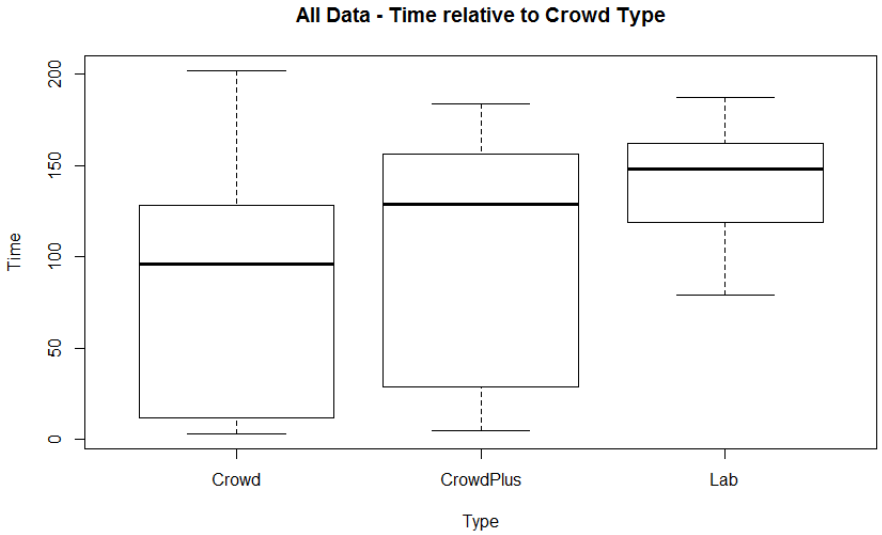


Figure 6.1. *Image showing a box-plot for the first level Drop, with group type shown on the x-axis and the survival time on the y-axis.*

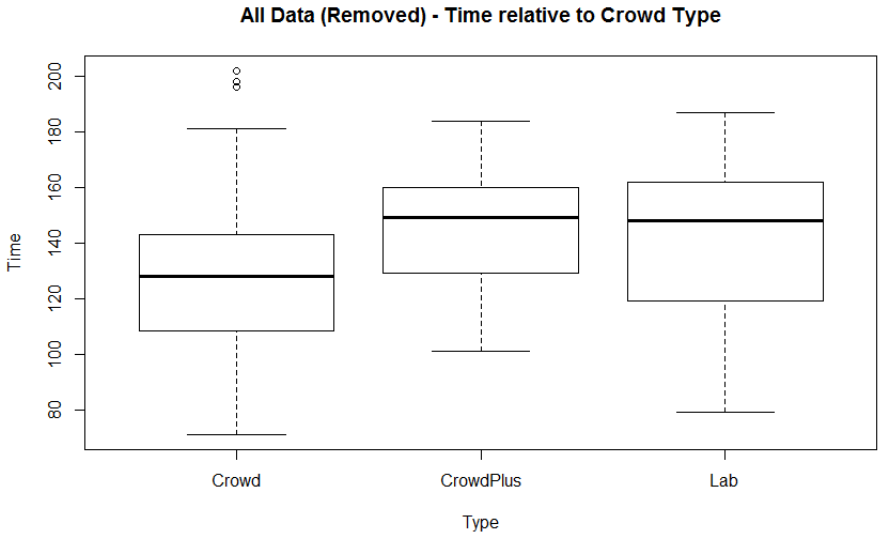


Figure 6.2. *Image showing a box-plot for the first level Drop, where all times below 70 has been removed, with group type shown on the x-axis and the survival time on the y-axis.*

prediction, and calculated a linear regression for the whole dataset, and in subsets of three successive indices of difficulty each. This means that for the first and last ID there is a missing slope value due to them not having an ID on both sides.

6.4.2.1 Lab

For the lab group we found that there was an overall significant difference in completion time between the seven tasks ($F_{6,1767} = 156.16$, $p \ll 0.001$). We created a mean slope for the time used for the lab group in each of the tasks. The mean slopes can be seen in

Figure 6.3.

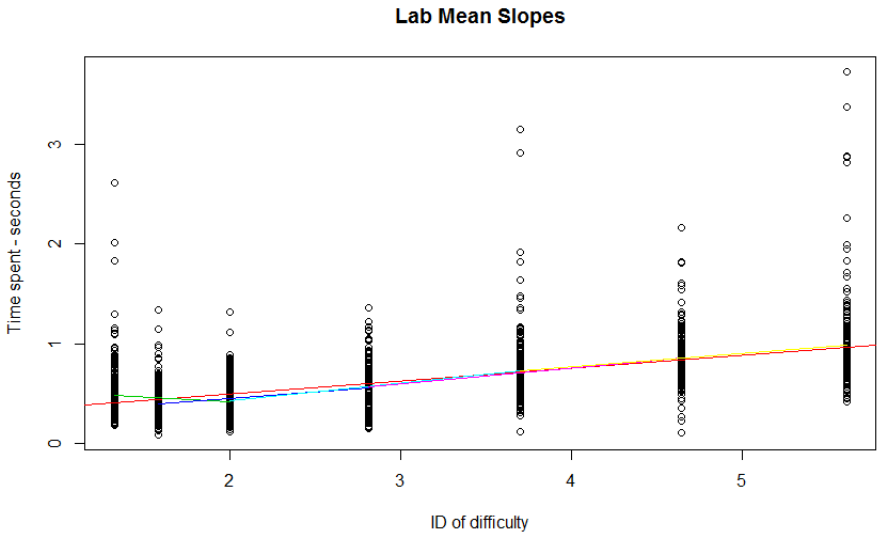


Figure 6.3. *The mean slopes for the time data for the Lab group.*

Analyzing the slope data, we find that no slope significantly deviated from the overall slope (0.09). Therefore we cannot say much about the DHR from the time data, which means that we have to examine the error data for the lab group. The distribution of the error data for the lab group can be seen in Figure 6.4

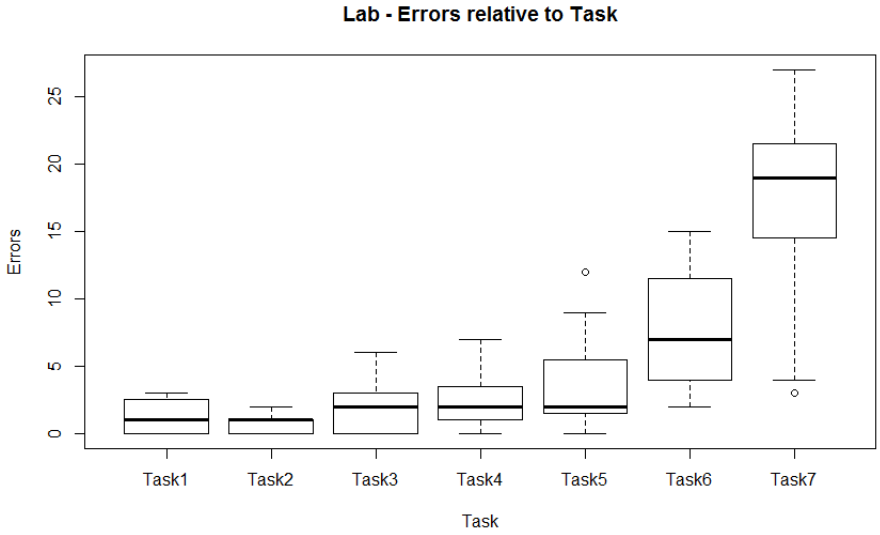


Figure 6.4. *Mean error data for each task in the lab group.*

Using a Friedman Ranked Sum test we found that there was an overall significant difference between the tasks ($\chi^2(6) = 63.98$, $p \ll 0.001$). Using the post-hoc of the Friedman test we find that a significant increase in errors happens around Task 5, which is similar to what is shown in the graph in Figure 6.4, where we can see that a small spike happens around this task. The ranks for the post-hoc of the Friedman test can be seen in Table 6.1.

Task	Rank
Task 7	A
Task 6	B
Task 5	C
Task 4	C
Task 3	CD
Task 1	DE
Task 2	E

Table 6.1. *Table showing the rankings from the Friedman Ranked Sum post hoc for the lab group. The ranks show no significant difference if the letters are the same.*

All the values for the lab group can be seen in Table 6.2. The mean slopes are calculated between the previous point and the point after. This is why there are no slope values for the first and the last IDs.

Target Size (mm)	Fitts' ID	Mean Time (sec)	Failure Rate (percent)	Mean Slope
32	1.32	0.55	6.64	
24	1.58	0.40	4.66	0.10
16	2	0.44	10.27	0.11
8	2.81	0.55	13.44	0.12
4	3.70	0.74	17.67	0.09
2	4.64	0.83	32.98	-0.13
1	5.61	0.99	56.76	

Table 6.2. *Table showing the values for the lab participant group in the main experiment.*

6.4.2.2 Crowd

For the crowd group we found that there was an overall significant difference in completion time between the seven tasks ($F_{6,962} = 3.58$, $p < 0.01$). We again created a mean slope for the time used for the crowd group in each of the tasks. The mean slope can be seen in Figure 6.5.

Analyzing the slope data, we again find that no slope significantly deviated from the overall slope (-0.0004). Again we cannot say enough from the time data due to this, and therefore we examine the error data for the crowd group as well. The distribution of the error data for the crowd group can be seen in Figure 6.6.

Using a Friedman Ranked Sum test we found that there was no overall significant difference between the tasks ($\chi^2(6) = 10.80$, $p = 0.09$). Using the post-hoc of the Friedman test we find that there is almost a spike between Task 3 and Task 4, but it is not big enough to be significant. This is similar to the results shown in Figure 6.6, as we can see that most of the data seem to lie closely on the same line. The ranks for the post-hoc of the Friedman test can be seen in Table 6.3.

All the values for the crowd group can be seen in Table 6.4. The mean slopes are calculated between the previous point and the point after. This is why there are no slope values for the first and the last IDs.

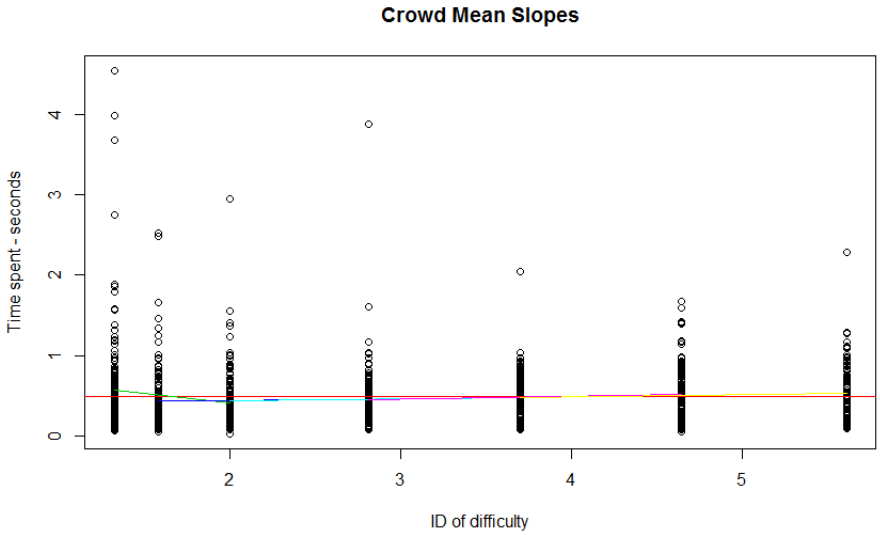


Figure 6.5. *The mean slopes for the time data for the Crowd group.*

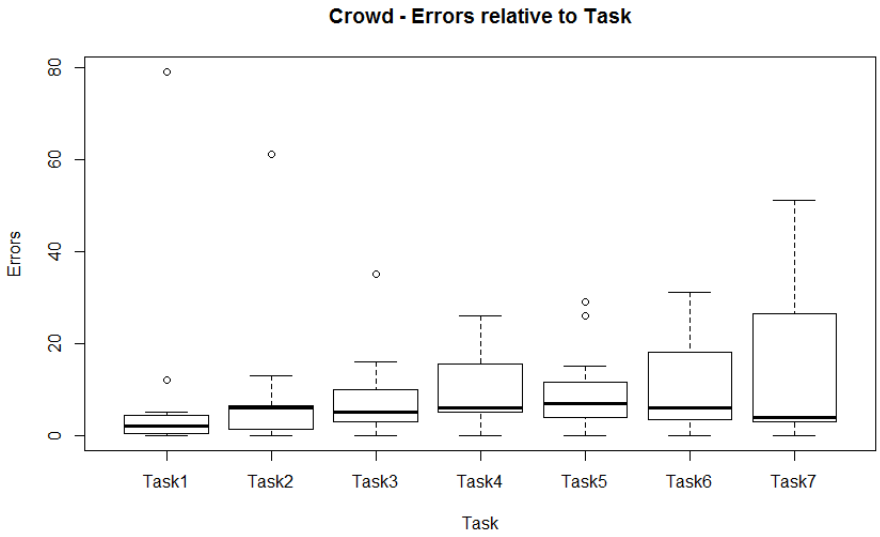


Figure 6.6. *Mean error data for each task in the crowd group.*

6.4.2.3 Crowd-Plus

For the crowd-plus group we found that there was an overall significant difference in completion time between the seven tasks ($F_{6,840} = 27.46$, $p \ll 0.001$). We again created a mean slope for the time used for the crowd group in each of the tasks. The mean slope can be seen in Figure 6.7.

Analyzing the slope data, we again find that no slope significantly deviated from the overall slope (0.09). Again we cannot say enough from the time data due to this, and therefore we examine the error data for the crowd-plus group as well. The distribution of the error data for the crowd-plus group can be seen in Figure 6.8.

Task	Rank
Task 5	A
Task 4	A
Task 6	A
Task 7	A
Task 3	AB
Task 2	AB
Task 1	B

Table 6.3. *Table showing the rankings from the Friedman Ranked Sum post hoc for the crowd group. The ranks show no significant difference if the letters are the same.*

Target Size (mm)	Fitts' ID	Mean Time (sec)	Failure Rate (percent)	Mean Slope
32	1.32	0.61	40.45	
24	1.58	0.44	41.47	0.03
16	2	0.44	37.92	0.04
8	2.81	0.45	44.49	0.02
4	3.70	0.47	44.03	0.01
2	4.64	0.52	49.80	-0.24
1	5.61	0.52	60.23	

Table 6.4. *Table showing the values for the crowd participants in the experiment.*

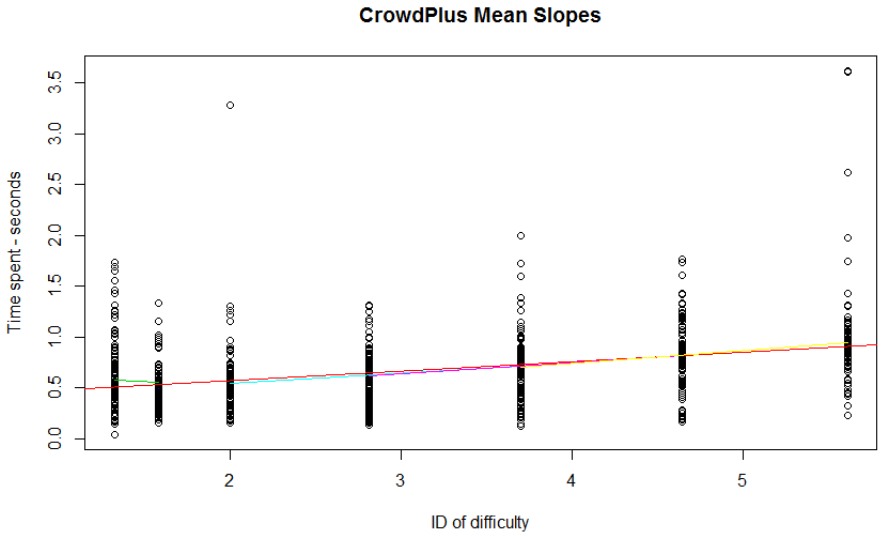


Figure 6.7. *The mean slopes for the time data for the Crowd-Plus group.*

Using a Friedman Ranked Sum test we found that there was an overall significant difference between the tasks ($\chi^2(6) = 23.47$, $p < 0.01$). Using the post-hoc of the Friedman test we find that there is a significant increase in errors between Task 4 and Task 5. This is similar to the results shown in Figure 6.8, as we can see a small spike happening between the two tasks. The ranks for the post-hoc of the Friedman test can be seen in Table 6.5.

All the values for the crowd-plus group can be seen in Table 6.6. The mean slopes are calculated between the previous point and the point after. This is why there are no slope

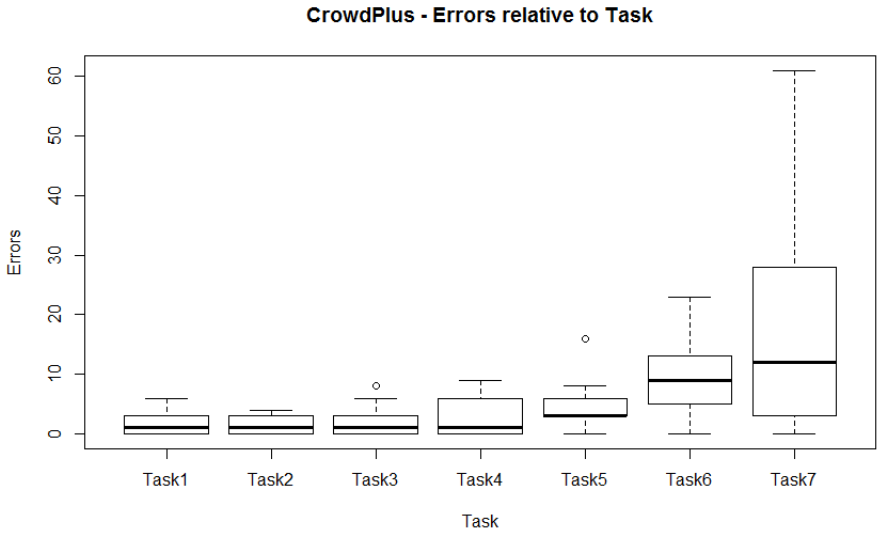


Figure 6.8. *Mean error data for each task in the Crowd-Plus group.*

Task	Rank
Task 6	A
Task 5	A
Task 7	A
Task 4	B
Task 3	B
Task 1	B
Task 2	B

Table 6.5. *Table showing the rankings from the Friedman Ranked Sum post hoc for the Crowd-Plus group. The ranks show no significant difference if the letters are the same.*

values for the first and the last IDs.

Target Size (mm)	Fitts' ID	Mean Time (sec)	Failure Rate (percent)	Mean Slope
32	1.32	0.62	10.89	
24	1.58	0.48	10.27	0.13
16	2	0.54	14.00	0.11
8	2.81	0.62	15.69	0.10
4	3.70	0.71	27.22	0.11
2	4.64	0.82	41.41	-0.09
1	5.61	0.95	64.66	

Table 6.6. *Table showing the values for the crowd-plus participant group of the experiment.*

6.4.2.4 All Data

Now that all the data has been compared individually, we look at the overall comparisons between participant groups. After that we find the overall DHR and Fitts' ID for all the participant data put together.

Group comparisons

Examining the overall time for each type we see that there is an overall significant difference between the participants groups in terms of time ($F_{2,3587} = 57.99$, $p \ll 0.001$). Using the TukeyHSD post-hoc we see that there is no significant difference in time between crowd-plus and lab ($p = 0.09$), however, there are significant differences between lab and crowd ($p \ll 0.001$) and crowd-plus and crowd ($p \ll 0.001$). This data has also been visualized in a boxplot in Figure 6.9, where we see a similar trend as the one from the post-hoc.

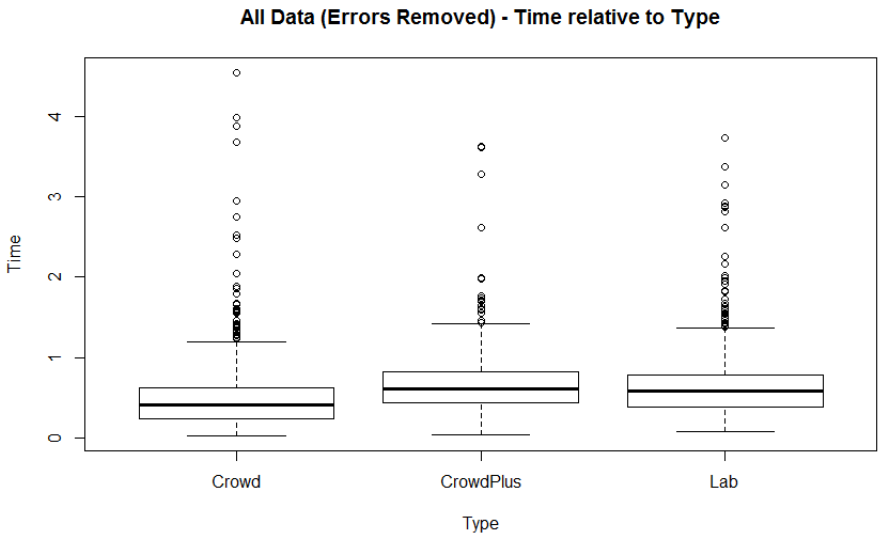


Figure 6.9. *Boxplot showing comparative time distribution for each participant group.*

For the error data an ANOVA was used, which showed that there was an overall significant difference between the different participant groups ($F_{2,249} = 7.17$, $p < 0.01$). Using a TukeyHSD post-hoc test we see that there is a (small) significant difference between crowd and crowd-plus ($p = 0.02$). However, for crowd-plus and lab there is no significant difference in the amount of errors made ($p = 0.87$). Finally there is a significant difference between lab and crowd, where the crowd participant group makes significantly more errors than the lab ($p < 0.01$). A boxplot can be seen in Figure 6.10, for a visualization of the data.

Touch DHR

To find the overall DHR for touch we first examine the time data together. An ANOVA shows that there was an overall significant difference between the seven tasks ($F_{6,3583} = 98.39$, $p \ll 0.001$). Using a TukeyHSD post-hoc test on the data we find that there are no significant difference between Task 1 and 4 ($p = 0.82$), and Task 2 and 3 ($p = 0.62$), with significant differences between all other tasks.

Analyzing the slope data for the overall DHR we find that once again the mean slopes did not significantly deviate from the overall slope (0.09).

Using a Friedman Ranked Sum test on the error data we find that there is an overall significant difference in errors between the seven tasks ($\chi^2(6) = 78.61$, $p \ll 0.001$). Using a TukeyHSD post-hoc test on the data we find that there are significant differences in

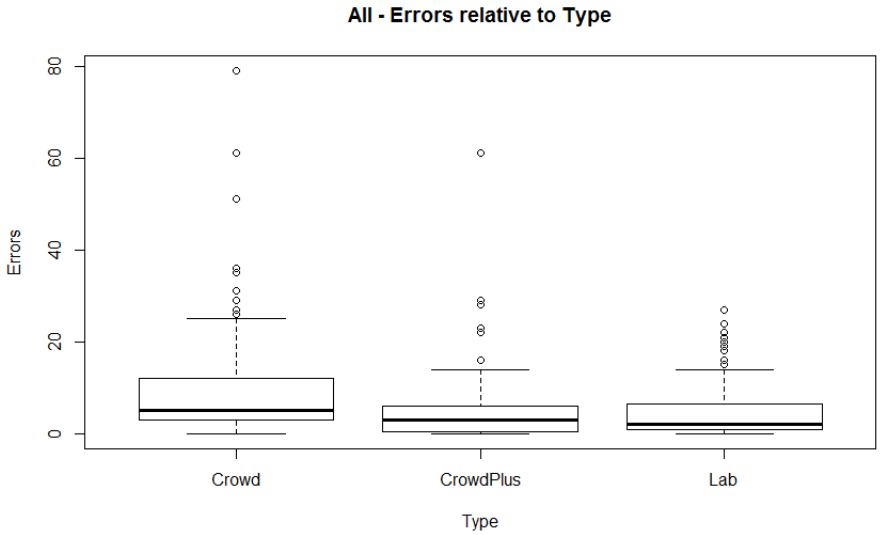


Figure 6.10. *Boxplot showing comparative error distribution for each participant group.*

errors between most tasks, however, most of the tasks are also not significantly different from its prior task. The results from the post-hoc test can be seen in Table 6.7.

Task	Rank
Task 7	A
Task 6	AB
Task 5	BC
Task 4	CD
Task 3	DE
Task 2	E
Task 1	E

Table 6.7. *Table showing the rankings from the Friedman Ranked Sum post hoc for all the participant groups together. The ranks show no significant difference if the letters are the same.*

As it is still somewhat hard to say from the data when the significant increase in data occurs, in order to determine the DHR for touch, we examine the data in a graph, which can be seen in Figure 6.11. From the graph we can see that a large increase (spike) in errors seems to occur between Task 5 and 6, whereas yet another spike occurs between Task 6 and 7.

All the values for the gathered participant groups can be seen in Table 6.8. The mean slopes are calculated between the previous point and the point after. This is why there are no slope values for the first and the last IDs.

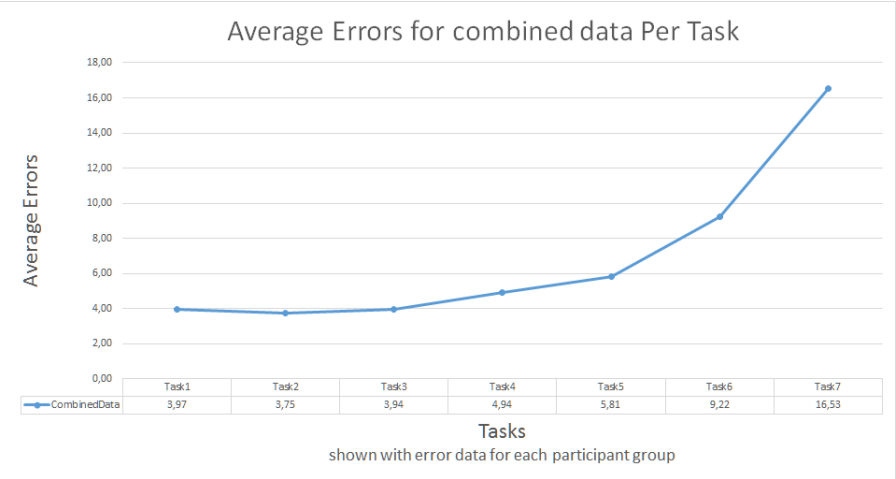


Figure 6.11. *Graph showing the average errors used for the combination of the participant groups.*

Target Size (mm)	Fitts' ID	Mean Time (sec)	Failure Rate (percent)	Mean Slope
32	1.32	0.58	20.43	
24	1.58	0.43	19.77	0.10
16	2	0.47	20.82	0.11
8	2.81	0.54	25.00	0.12
4	3.70	0.66	28.67	0.09
2	4.64	0.75	40.19	-0.13
1	5.61	0.86	59.64	

Table 6.8. *Table showing the values for the gathered participant group (all data) for the experiment. The values are used to calculate an overall DHR for touch devices.*

6.4.3 Extended Data Analysis

Now that we have found the results for both levels in the game, we extend some of the results even further. When examining the above graphs and the data used for the gathering of the results, we can see some tendencies. The results for Drop might be skewered a bit for the Crowd due to a lot of time results in the lower end, and the DHR results might also be skewered a bit because of some participants giving very bad results, which could again skewer the data in the negative direction, especially for the Crowd participant group.

6.4.3.1 Drop

As mentioned above, for Drop we examine the results for the test when all overall outliers have been removed. Before we tried to remove all data below a time of 70, however, this might not even be sufficient enough. Therefore, we have chosen to examine the results if we only include the highest survival time for each participant. This means that for each participant we will only have one of the minimum three (because of the three lives) time scores included. For the 'only highest' survival time we ran an ANOVA test, which showed that there was no overall significant difference between the results ($F_{2,38} = 0.205$, $p = 0.82$). Using a TukeyHSD post-hoc test we found there was no significant difference

between crowd and crowd-plus (p = 0.87), the same for crowd and lab (0.99), and also for lab and crowd-plus (p = 0.81). The results can be seen in a box-plot in Figure 6.12.

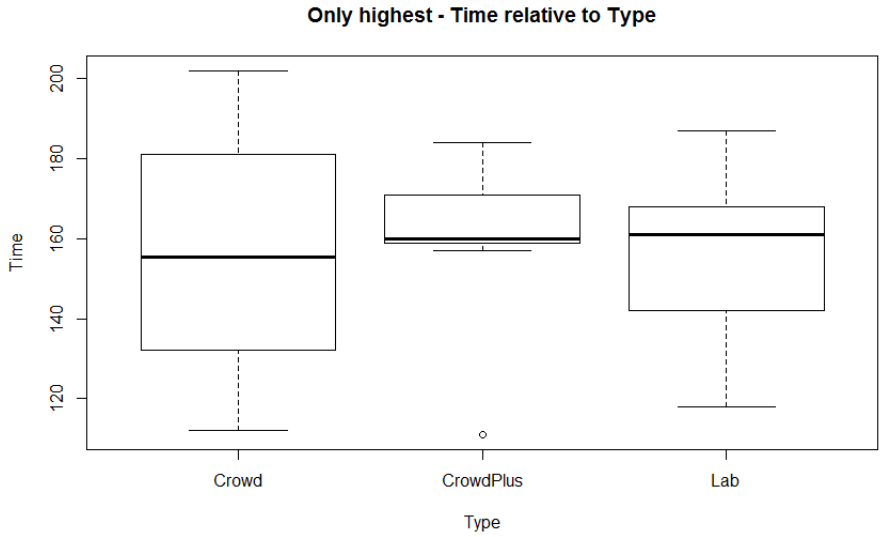


Figure 6.12. *Boxplot showing comparative survival time distribution for each participant group for Drop.*

The results show that when only considering the highest result for each participant there is actually no difference between the groups. This means that if only examining the best achievable performance for all participants, there is no difference between the groups, however, the previous results shows that there are significant differences between groups when all data is included, and also when only the very low survival times for each group is removed. This indicates that the participants from all groups might be able to perform at the same level, however, the participants in especially the crowd group chose to either not complete the level all three times, or simply only had a high performance in one of their trials rather than keeping the survival times more consistent, as seen in for example the lab participant group.

6.4.3.2 DHR

In order to examine DHR in a different way, we again looked at the data and saw a big difference in the amount of repetitions that the users in general used. We also saw that the participants who did a low amount of repetitions seemed to be considerably worse at completing the exercise. Therefore we examine if removing all participants below the average amount of repetitions for each participant group might show different results than what we initially found. This thereby only includes the data from the participants that we are sure actually understood the required task for the DHR level. The amount of repetitions for each group can be seen in Table 6.9. This means that for crowd all participants who did under 7 repetitions were removed, leaving only 7 participants. For crowd-plus it was all under 9 reps, leaving only 6 participants. For lab it was all under 16 reps, leaving only 8 participants. For the remaining results we will only use the data from these participants.

Participant Group	Avg Repetitions	Standard Deviation
Crowd	7.68	9.87
CrowdPlus	9.43	9.34
Lab	16.56	10.41

Table 6.9. *Average amount of repetitions and the standard deviation for each participant group.*

For DHR we again first examine time. An ANOVA shows that there is an overall significant difference between the different participant groups ($F_{2,2856} = 92.206$, $p \ll 0.001$). A TukeyHSD post-hoc test shows that there are significant differences between all groups ($p \ll 0.001$). A box-plot illustrating the data can be seen in Figure 6.13.

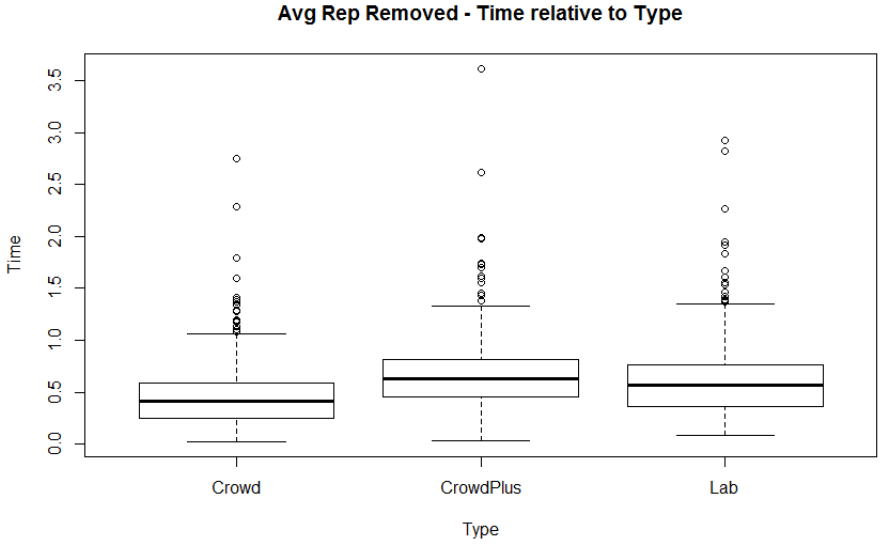


Figure 6.13. *Boxplot showing comparative time distribution for each participant group for DHR.*

Examining the mean slope, we once again find that no slope deviated significantly from the overall slope (0.09). The mean slope for all the data combined can be seen in Figure 6.14.

Again we now examine the errors for DHR in order to see if they say anything significant. First of all we find if there are differences between the three participant groups. Through an ANOVA we find that there was no overall significant difference between the three groups ($F_{2,144} = 0.7941$, $p = 0.454$). Using a TukeyHSD post-hoc test we find no significant difference between the individual groups as well. For crowd compared to crowd-plus we found a p-value of 0.565, for lab and crowd a p-value of 0.990, and for lab compared to crowd-plus gave a p-value of 0.466. The results from the group comparisons of the data with the participants under average amount of repetitions removed, can be seen in Figure 6.15.

To find the DHR for touch when the participants below the average amount of repetitions has been removed, we examine the tasks for the overall data to see if there is a spike in the errors in the data. Through an ANOVA we found that there was an overall significant difference in the amount of errors for the tasks ($F_{6,140} = 26.29$, $p \ll 0.001$). Using a TukeyHSD post-hoc test we find that there is a significant difference between Task 7 and

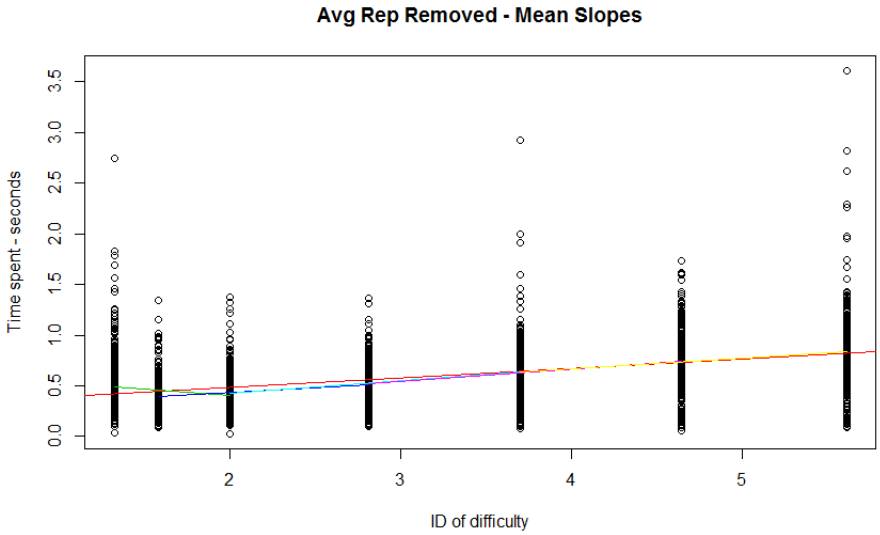


Figure 6.14. *The mean slope and subslopes for all the data combined. No average slope deviated significantly from the overall slope.*

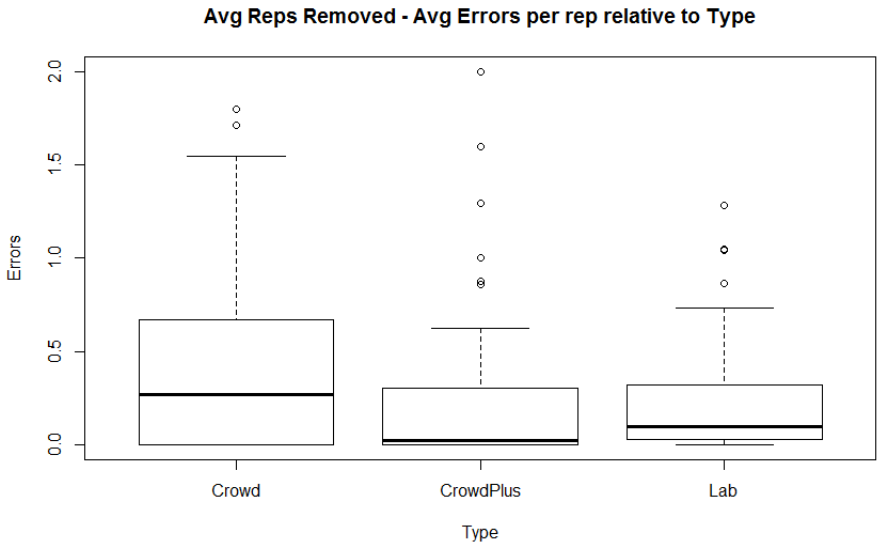


Figure 6.15. *A boxplot illustrating the error data for the different participant groups.*

all other tasks. Task 6 was significantly different from the other tasks, except for task 1 ($p = 0.07$). Examining the box-plot in Figure 6.16 we see the same tendency, that a spike in the amount of errors happens between task 5 and 6, which means that the DHR for the overall touch performance error data is still between a Fitts' ID of 3.70 and 4.64.

In order to illustrate the overall differences between the groups for the amount of errors generated per repetition we furthermore created the graph that can be seen in Figure 6.17.

The overall data for the extended data analysis can be seen in Table 6.10.

The results above show the DHR for the data, when all the poor data has been removed.

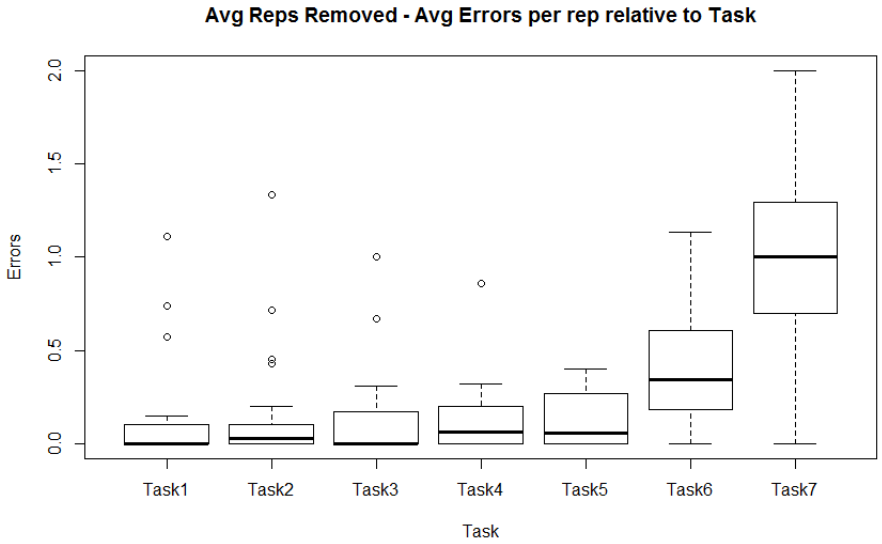


Figure 6.16. *A boxplot illustrating the error data for the different tasks with all the groups gathered.*

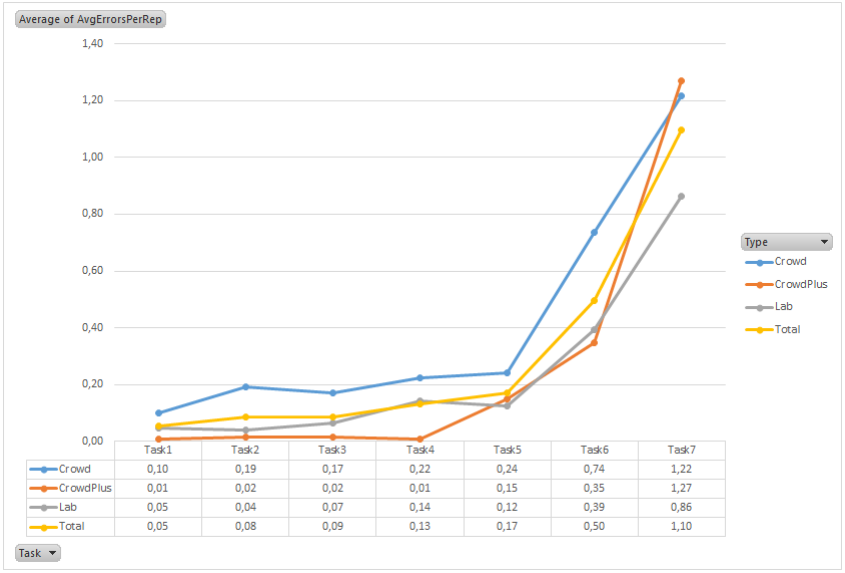


Figure 6.17. *Graphs illustrating the distribution of the error data for each of the participant groups, and also the overall average errors. The numbers below show the average amount of errors per repetition for each task.*

This means that only the participants who were able to complete several repetitions were included in the data. We can see that the average amount of repetitions for the lab is way higher than the other groups, which shows that the lab once again just seems to be superior in the understanding and completion of the tasks in the DHR level.

6.4.4 Qualitative Analysis

The participants in the lab and crowd-plus answered a questionnaire instantaneously after they participated in the experiment. We were not able to ask the crowd group, as they

Target Size (mm)	Fitts' ID	Mean Time (seconds)	Failure Rate (percent)	Mean Slope
32	1.32	0.53	11.88	
24	1.58	0.41	10.85	0.10
16	2	0.44	9.70	0.12
8	2.81	0.51	10.78	0.12
4	3.70	0.61	10.21	0.09
2	4.64	0.74	27.35	-0.12
1	5.61	0.84	52.08	

Table 6.10. *Table showing the data for final data distribution where participants below the average amount of repetitions has been removed.*

did not know they participated in an experiment. From the lab environment, we know about the particular environment they used, however in the crowds we know fairly little about them. Therefore, from the questionnaire in the crowd-plus we asked few questions regarding their environment.

6.4.4.1 Lab Experiment

16 participants from the local university campus participated. The voluntary participants were in the age range 22 to 26 years old (M = 24, SD = 1.5). All participants were male. All participants played computer games and 14 out of 16 used touch screen during the week and all participants played less than 11 hours per week on mobile devices with three not playing at all. For a summary of the data, see Table 6.11.

Participant rating on hours per week using Touch device, hours per week playing Computer games and mobile games						
Hours per week	0	1-5	6-10	11-15	16-20	20+
Touch devices	2	2	7	4	0	1
Computer games	0	3	3	4	3	3
Mobile games	3	9	4	0	0	0

Table 6.11. *Table showing the distribution of answers for Touch device usage, amount of playing computer games and mobile games.*

For the experiment, they answered about their focus and motivation level while playing the mobile game, they should also state if they found replay value in the game they just played. First, they rated focus with 62.5% on four, 25% on 3 and the remaining rated 5. Several participants mentioned they saw the game as two different games and had probably rated differently if they should rate individually on both games with the drop level being the highest rated. Observations during the experiment showed that in the start of the first level they seemed to have less focus towards the screen compared to when the challenge increased. For the second level, their focus seem to vary, however, at the smaller targets all participants moved closer to the screen. For their motivation of playing mobile games, they were circulating more around the middle with ratings of 43.75% on 3, 37.5% on 4, 12.5% on 5 and 6.25% on 2. Several participants mentioned that for motivation of playing this game, they would probably rate the first level higher than the second level. See summary of data in Table 6.12.

Participants rating of focus and motivation during mobile games						
	0	1	2	3	4	5
Focus	0 (0%)	0 (0%)	0 (0%)	3 (25%)	10 (62.5%)	2 (12.5%)
Motivation	0 (0%)	0 (0%)	1 (6.25%)	7 (43.75%)	6 (37.5%)	2 (12.5%)

Table 6.12. *Table showing the distribution of answers for Focus and motivation level.*

The Participants further commented on the replay value of the game. They had to answer why / why not they found replay value in the game. Only a single participant did not find replay value at all in the games, his reason was it felt too repetitive. In total six participants mentioned repetitiveness as a reason for not playing again. However, beating own highscore, the intense feeling and challenge in the level was mentioned as reasons to try the game again. Several gave suggestions to improve the game for better replay value; some of these were power ups, global highscore, faster phase from start, more visual feedback and better sound feedback in the game.

Observations During the experiment, participants used three different grips while playing the game. The section divides in Drop level and wall destroyer level; because the level seem to change, the participants grip type.

Drop Level

The majority of the participants decided to hold the phone in a firm grip with both hands and used both thumbs for interaction, see A showing the firm grip used during the test in Figure 6.18. They did not have to be precise and therefore they focused on fast phased motions. Several tried to put the phone on the table but eventually decided to pick it up and hold it in a firm grip. A few left the phone on the table for the whole duration of the test, see B for grip while phone is on table in Figure 6.18.

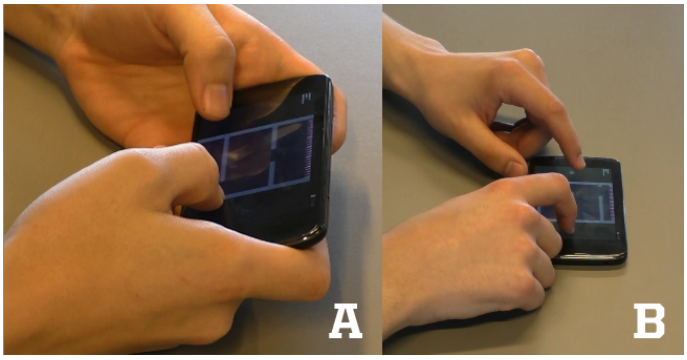


Figure 6.18. *Images showing participants holding the device in different grip types while playing Drop.*

Wall Destroyer Level

The participants mentioned this game as a precision game and therefore the majority of participant used their index finger. Several participants started with a firm grip but ended up either using one hand, see A, B, E in Figure 6.19 or placing the phone on the table,

see C, D in Figure 6.19. When going from the firm grip to the other also meant going from thumb to index finger. Some participants mentioned that their thumb was "too fat" for the task and they had to use their index finger and even some responded that their index finger was too large for the task as well. Further, participants tried during this level only to use the tip of the index finger to get the highest precision, therefore none participants used a flat finger on the screen. This observation were also found by Holz and Baudisch [20].

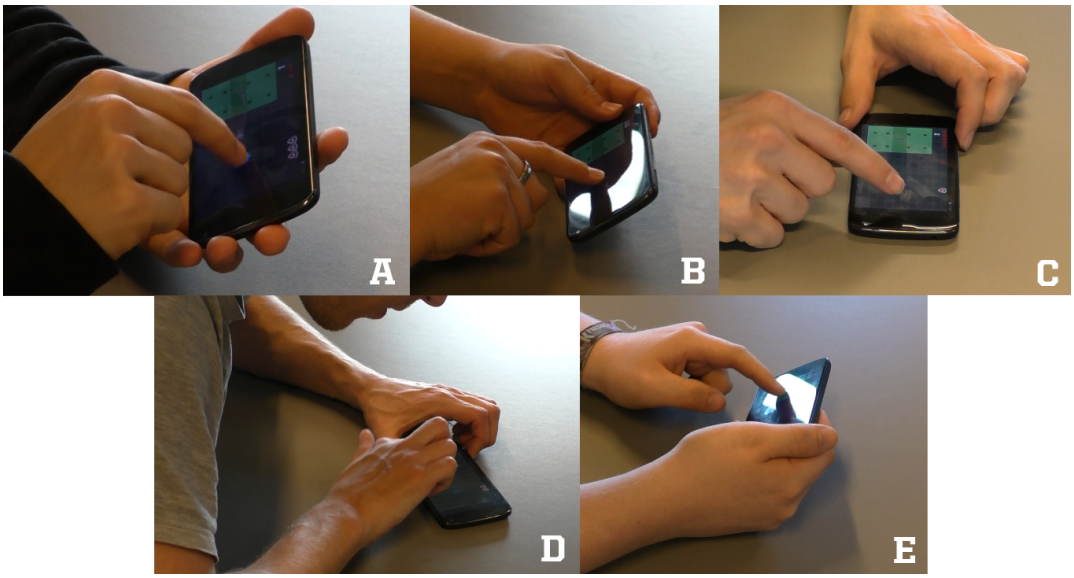


Figure 6.19. *Images showing participants holding the device in different grip types while playing Wall Destroyer*

6.4.4.2 Crowd-plus

12 participants in the crowd answered a questionnaire after playing the game. The participants were in the age range 22 - 57 years old (M = 28.17, SD = 0.8). Four female and eight male participated in the crowd. Only a single participant played the game in a public environment and the rest (11) answered private environment.

All participants were experienced touch devices users with a range from 1 - 20+ hours touch device usage per week with an average of 9.75 hours per week (SD = 0.67). The majority of participants played little to none computer games during a week with five not playing at all, and the rest on average played between 3.9 and 5.9 hours per week (SD = 0.53 and 0.57). For mobile gaming, our participants are not playing that much during a week. Only two participants play 6 - 10 hours during the week. On average they play between 1.5 and 4.2 hours during a week (SD = 0.18 and 0.3). See summary of data in Table 6.13.

About replay value, the participants were torn. Some did not find the games interesting, while other found them fun and challenging. Several wanted to beat their high scores and compete against other peoples high scores (mentioned as a feature request). A few mentioned they did not understand the second level, and this highly indicate the game have to iterate more before a finale release.

Gaming Habits						
	0	**1-5**	**6-10**	**11-15**	**16-20**	**20+**
Touch Devices	0	4	2	1	3	2
Computer Games	5	3	2	1	0	1
Mobile Games	4	6	2	0	0	0

Table 6.13. *Table showing the distribution of answers for Touch device usage, amount of playing computer games and mobile games.*

For noise level, the one in public had a noise level of four, while only a single in the private environment had answered four as well. Most in the private environment answered two or one indicating them sitting in a quiet environment. When answering whether they were in movement or being stationary. The majority answered one for stationary. While two participants answered three including the one in public environment.

The participants were somewhat focused while playing the game, with seven answering above 4 for focused and five answered below 2 for unfocused. However, their engagement level while playing were high with only two participants answering below 4. This indicate that the participant probably easily could get distracted while playing the game, but was engaged to keep playing. See summary of data in Table 6.14.

Participant Rating					
	1	**2**	**3**	**4**	**5**
Noise Level	3	4	2	4	0
Movement Level	8	2	2	0	0
Focus Level	2	3	0	4	3
Engagement Level	1	1	0	5	5

Table 6.14. *Table showing the distribution of answers for Noise, Movement, Focus and engagement level.*

Chapter 7

Discussion

The results from the main experiment confirm our expectations that the lab setup for most touch tasks will outperform the crowd setups. As mentioned in the experiment we however cannot with a 100% certainty state if the differences in performance between the participant groups is due to the environment or due to the users participating in the tests. However, we have found that both of these factors can influence the touch performance and we have found that the WEIRD subjects (more specifically western university students) in a controlled lab environment with no environmental disturbance will have a higher performance than the subjects playing in their own environment. It should be noted that there was no significant performance difference for both time ($p = 0.09$) and errors ($p = 0.87$) in the DHR part of the experiment between the lab and crowd-plus groups. One of the issues with the experiment, is also the fact that we are examining if the inclusion of the WEIRD population has an effect on the data, however, from the demographic questionnaire included in the crowd-plus participant group, we can see that most of the participants there can also be classified as being part of the WEIRD population.

Furthermore, the results show that there is a significant increase in performance if participants are given a consent and informed that they are participating in a study, rather than just playing a game. A reason for this could be the same reason as why participants in a the lab environment are also performing so much better, namely that when knowing that their results matter and have an influence on the results from a test, the users will try their very best and in most cases increase their performance while interacting with the application. In all cases of the test both the lab and crowd-plus participant groups significantly outperformed the crowd group.

While investigating the performance differences between groups, we gamified the existing HCI test of finding a device's human resolution (DHR) for touch applications. The results show that a significant increase in errors happens around a Fitts' ID between 3.70 and 4.64 (task 5 and 6), and if testing is done with an even smaller target size (Fitts' ID of 5.61), then another significant spike in errors is to be expected. One thing that should be noted with the overall combination of the participant groups in order to find the DHR for touch is that the amount of errors that the crowd group made (even for the large target sizes) is so high that it skewers the overall data. However, in order to get an accurate measurement of the DHR we need to include all participants, as it measures the human

resolution and as we can see from the results, the accuracy and performance of humans is very different between individual users.

Lastly, we investigated if there were any issues related to gamifying existing HCI user studies in order to make users in the crowd participant group believe that they were only playing a game and not participating in a user study. We cannot with absolute certainty state that users thought they were only playing a game, however, as stated above the users who were informed that they were participating in a study had a significant increase in performance, which shows that the participants in the crowd group played the game as normal, but might have focused less on their own performance compared to the two other groups. We also saw that gamifying existing tests might quickly become tedious, as game elements, scores, lives et cetera. are simply not enough to make a game fun. With the feedback we got from the test, we can also see that even adding an overall story for the game in future iterations, might not change the fact that the DHR level of the game is simply not fun and engaging, but rather feels like forcing the users to do something for an extended period of time, which they do not even feel like doing.

Bibliography

[1] J. Arnett. The neglected 95%: Why American psychology needs to become less American. *Behavioral and brain sciences*, 63:602–614, 2008.

[2] O. K.-C. Au, C.-L. Tai, and H. Fu. Multitouch Gestures for Constrained Transformation of 3D Objects. *Computer Graphics Forum*, 31:651–660, 2012.

[3] N. Baumard and D. Sperber. Weird people, yes, but also weird experiments. *Behavioral and brain sciences*, 33, 2010.

[4] F. Bérard, G. Wang, and J. Cooperstock. On the limits of the human motor control precision: the search for a device's human resolution. *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction*, 6947:107–122, 2011.

[5] B. Costello and E. Edmonds. A study in Play, Pleasure and Interaction Design. *Proceedings of the 2007 conference on Designing pleasurable products and interfaces*, pages 76–91, 2007.

[6] M. Csíkszentmihályi. *Flow: The Psychology of Optimal Experience*. Springer, 1990.

[7] M. Csíkszentmihályi. *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*. Springer, 2014.

[8] S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From Game Design Elements to Gamefulness: Defining "Gamification". *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15, 2011.

[9] A. J. Elliot and C. S. Dweck. *Handbook of competence and motivation*. Guilford Press, 2005.

[10] D. R. Flatla, C. Gutwin, L. E. Backe, S. Bateman, and R. L. Mandryk. Calibrating Games: Making Calibration Tasks Enjoyable by Adding Motivating Game Elements. *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 403–412, 2011.

[11] S. Gächter. (Dis)advantages of student subjects: What is your research question? *Behavioral and brain sciences*, 33, 2010.

[12] Gartner. Gartner says by 2015, more than 50 percent of organizations that manage innovation processes will gamify those. `http://www.gartner.com/newsroom/id/1629214`. Accessed 25th February 2015.

[13] Google. Metrics and Grids. `http://developer.android.com/design/style/metrics-grid.html`. Accessed 10th March 2015.

[14] F. Groh. Gamification: State of the Art Definition and Utilization. *Proceedings of the 4th seminar on Research Trends in Media Informatics*, pages 39–46, 2012.

[15] J. Hamari, J. Koivisto, and H. Sarsa. Does Gamification Work? - A Literature Review of Empirical Studies on Gamification. *Proceedings of the 47th Hawaii International Conference on System Sciences*, pages 3025–3034, 2014.

[16] J. Heer and M. Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 203–212, 2010.

[17] J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33, 2010.

[18] N. Henze, M. Pielot, T. Schinke, and S. Boll. My App Is an Experiment: Experience from User Studies. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 3:71–91, 2011.

[19] N. Henze, E. Rukzio, and S. Boll. 100,000,000 Taps: Analysis and Improvement of Touch Performance in the Large. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 133–142, 2011.

[20] C. Holz and P. Baudisch. The Generalized Perceived Input Point Model and How to Double Touch Accuracy. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 581–590, 2010.

[21] M. Jones and G. Marsden. *Mobile Interaction Design*. John Wiley & Sons, Ltd, 2006.

[22] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing User Studies with Mechanical Turk. *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.

[23] S. M. Kolly, R. Wattenhofer, and S. Welten. A Personal Touch: Recognizing Users Based on Touch Screen Behavior. *Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones*, pages 1–5, 2012.

[24] H. Korhonen, M. Montola, and J. Arrasvouri. Understanding Playful User Experience Through Digital Games. *Proceedings of the International Conference on Designing Pleasurable Products and Interfaces*, pages 274–285, 2009.

[25] R. Koster. *Theory of fun for game design*. O'Reilly Media, Inc., 2013.

[26] M. Kranz, L. Murmann, and F. Michahelles. Research in the Large: Challenges for Large-Scale Mobile Application Research - A Case Study about NFC Adoption using Gamification via an App Store. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 5:45–61, 2013.

[27] A. Lucero and J. Arrasvouri. PLEX Cards: a source of inspiration when designing for playfulness. *Proceedings of the 3rd International Conference on Fun and Games*, pages 28–37, 2010.

[28] A. Lucero and J. Arrasvouri. The PLEX Cards and its techniques as sources of inspiration when designing for playfulness. *International Journal of Arts and Technology*, 6:22–43, 2013.

[29] A. Lucero, J. Holopainen, E. Ollila, R. Suomela, and E. Karapanos. The Playful Experiences (PLEX) Framework as a Guide for Expert Evaluation. *Proceedings of the 6th International Conference on Designing Pleasurable Products and Interfaces*, pages 221–230, 2013.

[30] B. A. Nosek, M. R. Banaji, and A. G. Greenwald. E-Research: Ethics, Security, Design, and Control in Psychological Research on the Internet. *Journal of Social Issues*, 58:161–176, 2002.

[31] Out of Pixels. Drop. `https://play.google.com/store/apps/details?id=com.infraredpixel.drop`. Accessed 17th March 2015.

[32] T. Sala. Game Design Theory Applied: The Flow Channel. `http://www.gamasutra.com/blogs/ToniSala/20131208/206535/Game_Design_Theory_Applied_The_Flow_Channel.php`. Accessed 1st April 2015.

[33] S. Swink. *Game Feel: A Game Designer's Guide to Virtual Sensation*. Morgan Kaufmann, 2009.

[34] D. Wigdor and D. Wixon. *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*. Morgan Kaufmann, 2006.

[35] Wikipedia. Replay-value. `http://en.wikipedia.org/wiki/Replay_value`. Accessed 17th March 2015.

[36] V. Williamson. On the Ethics of Crowd-Sourced Research, 2014. `http://scholar.harvard.edu/files/williamson/files/mturk_ps_081014.pdf`. Accessed 9th March 2015.

[37] M. J. P. Wolf. *Encyclopedia of Video Games: The Culture, Technology, and Art of Gaming*. Greenwood, 2012.

[38] G. Zichermann and C. Cunningham. *Gamification by design: Implementing game mechanics in web and mobile apps*. O'Reilly Media, Inc., 2011.