Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

Hugo Solheim Klævold

Digiplay: An analysis of metrics for evaluating public installations

Master's thesis in MTDT
Supervisor: Yngve Dahl
June 2020

NTNU
Norwegian University of Science and Technology

Hugo Solheim Klævold

Digiplay: An analysis of metrics for evaluating public installations

Master's thesis in MTDT
Supervisor: Yngve Dahl
June 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

Sammendrag

Det moderne liv har ført til at vi i økende grad blir stillesittende. En måte vi kan prøve å bekjempe dette er ved å skape installasjoner som oppfordrer folk til å ta et avbrekk fra hverdagen til å være litt mer aktiv. For denne oppgaven har det blitt laget nettopp en slik installasjon med i form av en spillinstallasjon med to ulike moduser, der målet har vært å se på hvordan vi kan finne gode måledata for slike installasjoner. Gjennom en blanding av både kvalitative og kvantitative målemetoder har et sett med retningslinjer får å finne slike måledata blitt laget, inkludert hva aspekt med installasjonen som er nyttige å se på.

I tillegg så er det vist et tydelig grunnlag for å ta i bruk måledata ved at disse både lett viste problemområder av installasjonen, men også presis informasjon som ingen kvalitativ metode var nær med å avdekke. Dette inkluderer blant annet at spillere som spilte alene hadde en responstid som var i gjennomsnitt 350 ms verre enn for de som spilte mot noen andre, som gjør dem nesten 50% tregere. De som spilte mot andre hadde også større problemer med å lære mekanikkene der 39% av alle flerspillerspill så en spiller feile å håndtere en mekanikk som mindre enn 2% av enkeltspillerene feilet på. Alt i alt gir konklusjonen et veldig klart inntrykk av at rent kvalitative metoder ikke alene kan gi et godt bilde av interaksjoner med slike installasjoner.

Abstract

Modern life is becoming increasingly sedentary, and one way to combat this is to try to create installations that encourage people to become more active in their day to day lives. For this thesis one such installation has been created with two player modes, with the goal to look at how we can make good metrics to evaluate such public installations. Through a mix of qualitative and quantitative methods, guidelines for making such metrics have been made. This includes a look at what aspects of the installation one should consider when doing so.

Furthermore, it has shown the value of using metrics by readily identifying problem areas of the design, while also revealing precise information beyond what the qualitative methods could see. This includes showing that with the exact same mechanics players who played alone had an average of 350 ms worse reaction time than those that played against each other. Those that used the multiplayer also had a harder time learning the mechanics, with 39% of multiplayer game sessions having players failing at the very same mechanic less than 2% of the singleplayers fail at. Overall, the conclusion gives the very clear impression that qualitative methods do not stand well alone for judging these installation.

Preface

This thesis marks the end of a long and wonderful study. Interaction design and how we end up using technology have always been something that I found engaging. Being given this opportunity to work on a cross-study task where we both get to design and make an actual product, before trying out how people end up using it and then finally taking our own dive into what we learned from doing so has been a fantastic experience.

For the having been fantastic coworkers that made doing this project a joy, I must thank Ingrid, Sofie and Henrik. The installation we made and the data we gathered was a team effort that not one of us could have managed alone. Thanks must also be given to the supervisors Yngve and Ole Andreas for being incredibly helpful and positive all the way. By the end of this I feel like I have gotten a more well rounded education than I ever would have thought years ago, having had the opportunity to work with such excellent designers. And it feels nice to breathe in the fresh air as a great 5 years of time at this university come to a close.

Contents

List of Figures

List of Tables

1 Introduction

Technology is rapidly changing the way we are living our lives, as the rate at which we create new innovations snowballs and we become more and more reliant on it for all aspects of our lives. More and more of what surround us can now gather data, which is saved cheaper and more efficiently than ever and can communicate among each other[1]. But among all this progress humanity see themselves becoming more sedentary, as technology surrounds us we start anchoring ourselves to one spot[2]. With the world in our pocket there is little motivation to move aside from between our pre-planned sedentary spots.

This style of living is showing its impact in peoples life quality and the increased health service cost to care for them[3]. But the tech that encourage us to stay still do not need be the final nail in the coffin of our collective health. More and more seek out digital support for either starting or staying active. It shows that there is a potential for using technology to tip the balance of our daily lives back towards physical activity, even if only briefly. The question is then how do we design something that do encourage this, and how can we judge if what has been designed is well functioning.

1.1 Topics

The main topic of this thesis is on the usage of metrics themselves to judge a public installations, with a focus on those that encourage physical activity. This includes a dive into what the metrics ended up telling, and a look at how such metrics can be designed. The data gathered will be looked at from an interaction and design perspective, and more qualitative data will be used to help either dispel or support the insights the metrics claim to show.

1.2 Research problem and question

This thesis address the following problem:

"How can we make good metrics to measure designs of public installations?"

The most important contributions of are:

1. Analysis of the user behaviour found for the installation used by the thesis and what that means for the design
2. Guidelines that intend to show how to create good metrics, and what aspects such metrics might cover

1.3 Thesis composition

Chapter 2 brings us into the background covering the installation used to gather data made in autumn project, and relevant theory. Chapter 3 explains the methods of data gathering. Chapter 4 goes through the various results and analyse these to put them into a broader context, with chapter 5 taking the analysis and tries to establish what we have learned, what we can use this for and where we can go from here. Lastly, chapter 6 gives the final conclusion.

2 Background, preliminary work and theory

2.1 Design considerations for installation used in data gathering

With the study being part of a grander study on how to use technology to motivate physical activity in people's daily routines, preliminary work was done in autumn specialization project with the goal being to create a public installation to use as a vehicle for data gathering during the spring master. This installation was done as a cross-study project with students from design and computer science with a focus on interaction design. As the participants would use the installation to look at different aspects, it needed to be designed in such a fashion that it would satisfy the data gathering needs of all parties. This was in addition to being physically strenuous to use in order to fulfil the grander study's ambitions. In an attempt to achieve these goals, a game installation requiring physical movement with multiple modes was made.

All the methods of data gathering was done in the context of this installation. While it has gone through several iterations and usability testing phases, for the purpose of this thesis only the latest iteration is relevant for understanding what the the data has been gathered from. The modes chosen for the game was eventually boiled down to one competitive multiplayer mode, and one single player mode.

The design, both in hardware and software, followed a principle of modular design and architecture. This was to minimise time to push out changes, reduce possible code debt and make each component easy to reuse[4]. Doing so not only made it realistic to push out multiple game modes, but also made it far easier to be able to look at how users interacted with specific elements of the installation.

2.2 The physical installation

The physical installation itself consist of four independent columns each containing four buttons that are evenly spaced on the face on the column, with one column shown in figure 1. These buttons contain 24 lights in a circle going around the inside of the button. Each of these lights can be individually changed. The button itself is 3D printed and connects to a microswitch that sends a signal when pressed in. Button composition can be seen in figure2 For either mode, only two columns are in use for each player. The other two are either not in use, or used by the other player. These columns are installed a suitable distance above the ground with some distance between each other, requiring the player to find a way to move between them. The modularity allows this distance to vary.

Figure 1: Close up look of one of the columns

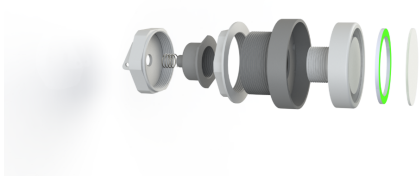


Figure 2: Composition of the button

The columns are detachable, but installation is set up is connected to a central microcontroller that takes care of the logic of the installation. The microcontroller fit into a box with power source, and the box has two poles elevating above it another box containing the scoreboard. This controller also connects to two other controllers, one that solely control sound, and one that solely controls the scoreboard. This was done both due to input and wire management, but also to maintain modularity of the installation. The wiring can be seen in figure 3, and the setup is shown later on in figure 6.

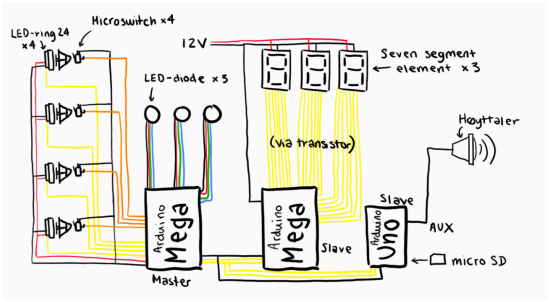


Figure 3: The wiring of the installation elements

The scoreboard consists of four seven-segment displays, with either two placed on front and back of the box for the competitive multiplayer, or only three used on the front of the box for the single player mode. In addition to these seven segment displays, both the front and the back has 3 LED lights below the scoreboard. These are used to give the player an indicator of the progress of the game. In competitive mode, these light up in the color of the player that won a set, with a set victory being “first to 9 points”. In the single player mode, all three are initially lit up as blue, and one gets turned off if the player lose lives.

2.3 The game logic of the installation

For both modes there are some common elements. The game is started by pressing the button that pulsates with green lights. Once pressed, there is a short animation with accompanying sound that shows each button light up in tune to the sound. This serves to not only give the player a short moment to prepare, but also make the player aware of what elements will be interactable during the the game.

When the game is started, the buttons have two possible active states. In the first they start with the circle light strip being uniformly green, which then turns off the lights along the strip in a clockwise fashion. By doing so it informs the player the time left to interact with the button. This button awards points in both game modes. The other active state has the circle light strip glow uniformly red, and staying so for the duration it is meant to be active. As pressing this button removes some points, having it behave this way signals it having a different functionality. Each possible active state (green or red) has a weighting. When it comes to which button is activated, the game logic selects a new button from the set of inactive ones at set intervals. These buttons work independently from each other, and have same weighting when it come to being selected for being the next to activate. Pressing either button play a sound that depends on mode of the button. Once the game is over, there is a small ending animation with sound where all buttons turn either red (singleplayer) or in the colour of the victor (multiplayer), before the installation resets the game to the pre-start glowing button.

For singleplayer, the goal is to stay in the game as long as possible to gain points. Missing a green button means you lose a life, and losing three lives means game over. To challenge the player to physical exercise and to gently introduce the player to the mechanics and the flow of the game, the game is broken apart into several “levels”. The existence of these levels are hidden from the player, and ramps up the challenge as the player progresses. Below is the flow of the levels. If not specified, then an aspect is the same as level above.

- Level 1 (0-5 points):
  - Only green buttons
  - One button activate at a time.
  - 2500 ms between new buttons becoming active.
  - 2400 ms to press a button before you lose a life.
  - 100 ms between each light gets turned off on a green button’s countdown.
- Level 2 (6-20 points):

  - Two buttons activate at a time.
  - 10% chance a button will become a red button.
- Level 3 (21-40 points):
  - 2000 ms between new buttons becoming active.
  - 1920 ms to press a button before you lose a life.
  - 80 ms between each light gets turned off on a green button's countdown.
- Level 4 (41-90 points):
  - Three buttons activate at a time.
- Level 5 (91-250 points):
  - 1500 ms between new buttons becoming active.
  - 1440 ms to press buttons.
  - 60 ms between lights get turned off on cooldown.
- Level 6 (250-998):
  - Four buttons activate at a time.
- Victory and game end (999 points)

For multiplayer, the players are competing against each other to be the first to press each green button. First player to reach 9 points wins a set, with there being 3 sets. The game is played until completion of all 3 sets regardless of victories of the first two sets. Players are designated as "green" or "blue" which depending on which pillars they use, which only affects the set lights and the color of the victory animation upon winning a set. For the purpose of these tests, the columns were placed on each side of a pillar, having the players face each other with the scoreboard in between them. The buttons to be activated are the same for each player, being mirrored on the columns that are on each pillar. This means that if a button activated on a players left side column, it activates the equivalent button on the other players right side column (being the same pillar when they face each other).

Like in single player, there are changes in intensity for each set to both introduce players at a decent pace, but also to keep it engaging and strenuous. The other player here however add another layer of challenge to overcome, that must be purely provided by the game in single player. Below is the progression of intensity, with amount of buttons being what one player sees. The actual amount of buttons in play is double as each player gets mirrored active. If not specified, then an aspect is the same as set above.

- Set 1:
  - Only green buttons
  - One button activate at a time.
  - 3500 ms between new buttons becoming active.
  - 2400 ms to press a button before you lose a life.
  - 100 ms between each light gets turned off on a green button's countdown.
- Set 2:
  - Two buttons activate at a time.
  - 10% chance a button will become a red button.
- Set 3:
  - 3000 ms between new buttons becoming active.
  - 1920 ms to press a button before you lose a life.
  - 80 ms between each light gets turned off on a green button's countdown.

2.4 What is metrics?

By definition, a metric is data measurements of a specific characteristic of what you intend to measure[5]. Or to boil it down, it is data points with a meaning. In software, the design of metrics tend to focus on performance and efficiency. Businesses themselves continue to increase the scale of the metrics that they gather, originally starting out with a focus financial health and moving on to customer satisfaction and engagement[6][7]. A lot of modern privacy concern relates specifically to big businesses trying to design metrics that learn about their customers wants and patterns of use. They want their customers to stay, spend a lot of time, and engage with not only their products but those of their partners. Metrics are perhaps the greatest tool of the data scientist to talk about why something is happening. They are the tool of improvement, as Klubeck says in his book Metrics: How to Improve Business Results[8].

2.5 How to develop metrics

Knowing what metrics are is near useless if we cannot find a good way to select metrics. Over the years there have been many strategies laid out to make metrics easier to develop. A common method is the usage of SMART criteria, which is a relatively old concept related to effectively setting goals[9]. The letters stand for Specific, Measurable, Achievable, Relevant and Time-bound. It is a fantastic general system for setting objectives, which shows in its heavy use in the four decades the concept has been known[10]. The idea to use it as a tool for picking metrics falls somewhat apart rather quickly. Metrics are by their nature specific and measurable, though it can be helpful to remind yourself of what is possible to pick metrics from. The first point to truly aid in design is "Achievable", as regardless of interest or usefulness a metric that can't be measured with the tools you have or can obtain must be discarded. Similarly, a metric has to be Relevant. There is an excessive amount of things that truly can be measured, and there is both limited time and resources to process everything. Time-bound aspect is once more not as relevant, other than picking when to start and end measurements. Its a neat system, but flawed when looking at pure metric design.

Other systems out there are IPA (Important, Potential Improvement, Authority)[11], TIE (Trackable, Important, Explainable) and many more, but while useful metric evaluation tools they do not do a really useful job at designing them to begin with. Here we return to Klubecks book on Metrics[8] where he has a chapter on designing metrics, looking at the process of going from the system to the metrics for it. Here he goes through the motions of asking questions about what is the end goal that we are trying to figure out something out about, and then looking at the avenues available. What would a possible metric tell, what ranges would it go through? More so than anything else, the examples he gives shows a clear tendency that everything has to be evaluated on a case by case basis. What is important regardless is that metrics are clear in what they represent and what they are measured in.

More over, he establishes that metrics are not always true but are indicators, and do require interpretation to be useful. Metrics do not stand on their own. It mentions the complement between the quantitative approach that metrics use and qualitative data, but details little beyond it.

2.6 Competitiveness

As the game installation will have two modes where one of these pits players against eachother, it would be remiss to not look at what studies show about the effect of competition on the participants. In a study from 2003 on bicycle racers and fitness exercisers that competitiveness was positively correlated to motivation for both groups, and in the case of the bicyclers higher level of adherence to the activity[12]. Another study done in Sweden looked at the amount of steps done in the workplace based on if there was introduced a contest to the amount of steps done[13]. The result was that the subjects that had the contest did about 10% more steps than the control group. Even though both game modes used for this thesis' installation do feature some element of challenge, the multiplayer mode will have a more direct element of competition and this could very well change the results.

3 Methodology

3.1 Data log file

To obtain data, several metrics were chosen to be saved into a log file on an SD-card upon completion of any game. The only input from the player is button presses, so the metrics had to be designed around this limitation. All the metrics are anonymous, as they record only how the game was played, and nothing of who. For any case of picking metrics, privacy will always remain a concern as it become challenging to be able to get useful relations between data without using or even accidentally creating identifying data. The usage of arduino mega and uno also meant that internal timestamps needed additional accessories that was not obtainable in time for testing, again making the data more anonymous, but also makes it far harder to judge the installation by time of the day. Thus it loses the ability to accurately judge what time of day the installation is most or least successful on.

With helpful background in how to judge metrics, but little in the process of creating them as discussed in section 2.5, a quick method of brainstorming metrics was devised. Below is the list of questions that were asked, going in strict order:

1. **What is possible to measure?**
   Here limitations was established, inputs and outputs were listed, and a look at what internal systems could be latched onto to give usable data. This question was meant to throw a wide net to be shrunk by others.

2. **What is reasonable to measure**
   Was there data that was possible to measure, but would take too much work to get working or take too long to process into usable data. The project had timetables that needed to be met.

3. **What is relevant to measure**
   Not all data that pass the first two questions would be data that would tell anything useful. While some data have the possibility to tell far more than expected, there is absolutely data that would contribute nothing more than noise. All the possible metrics were thus reviewed with an installation that encourage physical activity in mind.

From here, the metrics taken for this thesis were developed with a few data groups in mind, namely metrics that relate to:

- **Dimensions of the installation**
  This in particular is very relevant as the design need to balance accessibility vs physical challenge.

- **Engagement of installation**
  How long do players use, what encourage their use, and repeat engagement.

- **Difficulty curve**
  Does it challenge the player the way intended, do players get overwhelmed, does the challenge ramp up too fast or never reach a desired level and much more.

- **Understanding of mechanics and player relations.**
  How did players use mechanics, and what kind of players did they play against?

From these, a log entry was made for single and multiplayer. These are the core used for the results and discussion later in the thesis. An example of the single player log data entry is in figure 5, while an example of multiplayer log entry is in figure 4. Both have data from a developer test run, with explanation.

```
Session: 2, started 108596    (Game session since start, and time in ms since then)
Time: 78911                   (Time the game lasted in milliseconds)
Score: [9-0,9-0,9-0]          (Score of each side during the sets)
Winner: [1,1,1]               (Winner of each set)
Avg P1: 1296                  (Average reaction time in ms for player 1)
Fastest P1: 682, button: 5    (Fastest reaction time, and which button for player 1)
Slowest P1: 2087, button: 0   (Slowest reaction time, and which button for player 1)
Avg P2: 0                     (Average reaction time in ms for player 2)
Fastest P2: 0, button: N/A    (Fastest reaction time, and which button for player 2)
Slowest P2: 0, button: N/A    (Slowest reaction time, and which button for player 2)
Red buttons pressed: 2        (How many times were red buttons pressed)
Red time: [45599,57770]       (Time in session when red was pressed)
```

Figure 4: Log entry for multiplayer

```
Session: 2, started 28065     (Game session since start, and time in ms since then)
Time played: 85477            (Time the game lasted in milliseconds)
Score: 45                     (Score on game completion)
Level reached: 4              (Level of difficulty achieved)
Average response: 1341        (Average reaction time in ms)
Fastest response: 708, button 5 (Fastest reaction time in ms, and which button)
Slowest response: 1904, button 6  (Slowest reaction time in ms, and which button)
Red buttons pressed: 0              (How many times were red buttons pressed)
Red time (ms): []                   (Time in session when red was pressed)
Life lost at (ms): [85477,70027,29802] (Time in session when life was lost)
Life loss buttons: [5,0,4]          (Buttons causing life loss)
```

Figure 5: Log entry for singleplayer

3.2 Questionnaire

Each participant was asked to fill out a quick questionnaire if they tried the installation at least once. This questionnaire recorded the players gender and age. To increase participation in the questionnaire it was designed to be very quick to fill. The questionnaire itself can be found in Appendice A, being written in norwegian as the players would very likely be norwegian.

It asked players to rank whether they agreed with statements on how they found and experienced the various elements of the game, like intensity, length or usefulness of sound effects. It also asked how important they found these to be, and how enjoyable they found the experience on a scale from "very important" to "not at all".

The purpose of the questionnaire was to allow players to give a low effort way to say how interaction with the installation felt. By having this the hope was to minimise personal bias that reviewing data from a purely outside observer perspective would give, in particular when taking the role of observer while also being the designer.

3.3 Interviews

To get more in depth feedback from participants, a small portion of players from both the singleplayer and the multiplayer participants were asked to be part of a moderate length interview which was recorded. It had looser semi-structured form, touching upon several topics on how the participant interacted with the installation. While less relevant to the topic of this thesis compared to the others making use of this data, the ability to provide more detailed statements to support the conclusions removes the need to make guesses on behaviour shown in the metrics. An interview guide (in norwegian) can be found in appendice C.

These interviews were transcribed and anonymized, and the audio files is not to be kept post thesis due to privacy considerations from recognition of the voice in the files.

3.4 Observations

Two tools were used to aid in observation of player behavior. Firstly, a camera to film game sessions as long as the players signed a consent form for this. The consent form can be found in appendice E in norwegian. The consent form can be found in appendices. The video files is to be destroyed destroyed and not be kept long term due to privacy considerations, and video faces has been blurred in any usage of the video.

Secondly, an observation form was made to make it quick to note what was expected to

be the predominant points of interest, while still leaving room for making notes on the unexpected. These could have been translated to more direct data points, however due to time pressure and the amount of data already gathered they were rather included as help to review filmed sessions for points of interest. This was in particular as the data points from observations had a significant overlap with both questionnaire and interviews.

The observation form focused on the three aspect of player attitude and signaling, the movement of the player, and questions about the usability. One was made for each mode, and both can be found in Appendice B for multiplayer and appendice C for singleplayer.

3.5 Test plan, location and time

Data gathering was set for a period of two weeks, with 3 days each week for testing. The weeks set for these were the last week of february and first week of march. The first week would be dedicated to multiplayer, and the second week to single player. End goal was at least 50 players for each of the two modes, with as many as possible of these answering the questionnaire and being observed with observation form and camera. Ideally the goal was also about 20 interviews, evenly split across both modes. For the data log file, players would be allowed to play several times if there was no new players waiting to try, giving a significant quantity of test runs for the most objective measurement that it would sorely need to be able to say anything of significance. The time to process all this data meant that test period would likely not benefit from an increase in length.

If these goals could be achieved, the hope was that the solid and sizable mix of both quantitative and qualitative data would be a good basis not only for discussing the design of the installation, but also how to study such designs. The installation would be placed indoors at Gløshaugen at NTNU on the same location for all 6 days. The height and distance between the columns would also be the same. Due to location, the odds would be that the vast majority of participants would be students from the late teens to mid twenties for the entire test period. A look at the setup can be seen in 6.

Figure 6: Setup of installation in test locale for single player, only using two of the column.

4 Results

4.1 Data gathered from test phase

The goals for quantity of results solidly surpassed expectations, with the amount of participants during the test period going beyond the goals set. Each mode had over 60 unique players. This even with some technical issues meaning loss of some multiplayer log entries, and the crowd meant that it was too hectic to be able to consistently get all the data that was desired from every participants. Most participants joined in the 15 minute period between classes, with relatively low amount of participation for the remaining part of the hour.

- 33 log entries from multiplayer, with data from each player
- 138 log entries from singleplayer, as several players wanted second and even third round.
- 60 observation sheets for singleplayer, and 30 for multiplayer. Equal amount of camera footage clips from each mode.
- 60 questionnaires from each mode.
- 21 interviews, split evenly from each mode with 11 for multiplayer and 10 for single.

The focus on the results will be from the view of the metrics in the data log entries. As both game modes was tested for the same amount of time, the 33 log entries of multiplayer vs the 138 log entries for single player mode show a distinct increase of people trying the single player mode. Do again note that a handful of log entries did not get recorded for multiplayer due to an early technical issue caused, and that multiplayer featured two players each game. Thus the gap between the two is less than what it initially appears as.

4.2 Observations

Most players spent some time taking in the installation and preparing themselves with some stretching and some testing of the buttons before starting. Few questioned how to start, and for multiplayer the player whose side the start button was on waited for acknowledgement of the other player before starting. As the introductory animation played, players sprung quite readily into the action with some wondering if the game had already begun at very high intensity. The sound effect made almost none of the players interact with the installation during the animation however, as was intended. In one of the multiplayer games one player managed to hurt themselves, and in another multiplayer game one player managed to put in enough force to almost trip the entire setup. These were issues not seen in singleplayer, even though it had more games. Players mentioned they thought games took about 3 minutes, which they felt fit perfectly into the time they had available.

From the data from the forms, the vast majority was classified as "curious" or "excited" for both players. Players chose to communicate during the game and laughter was common, being much more common in multiplayer games. For movement, players ended up using whole body as a necessity, as those tall had to reach quite a bit down while those short very readily started jumping to quickly reach the top buttons. No player could immediately easily reach the more difficult combinations. Once the game started, there was little time spent questioning how it worked, with the only notable questioning was that some players though the game ended earlier than it did during multiplayer. These however readily sprung back into playing to finish once they saw the buttons activate.

An almost uniform tactic players did adopt was standing in between the two pillars, using their side vision to pick out when buttons activated. Those that did not do this quickly learned from those that did and kept this strategy throughout. This allowed them to have clear view of their opponents for multiplayer too. Taller players usually stretched their arms out, with shorter players being more willing to leave the middle position. Some of the players choose to jump to reach buttons even when not needed to reach the buttons faster. A common thread is also visible player celebration at winning or disappointment at not quite making it, and some games had players shaking hands afterwards. Some players expressed worry once they found out there were red buttons in addition to the green, as they had colourblindness. This happened for both modes.

4.3 Interviews

While interviewing players, a few aspects of the way they had experienced the game quickly became a common factor. Firstly, once the game started the reason they kept engaging was the competitive factor. For multiplayer they wanted to beat the other player, for single player to get as far as possible. That took over their focus. Second of all, players did not expect the difficulty increase, but brought up that they enjoyed it a lot. It kept the game fresh throughout their session. Players of the multiplayer mode did however mention that it was more sudden than they anticipated, so even though they enjoyed it they wanted more introduction to the mechanics.

Most importantly, players said that by the end they did not feel tired, but they felt pumped, warm and like they got to use their body. Players were split between if they wanted it to be harder, as some thought you could have gotten some more motion into it, but others said they barely managed to make it between columns in the end. By the end of their session they felt more ready to continue with their day and some said it was nice to feel a bit out of breath.

Below is one translated interview from a multiplayer session, with "I" representing interviewer, and "P" representing player interviewed.

I: How was the your first impression?
P: The installation looked professional, more so than I would have expected. I saw people moving a lot, and thought it looked fun
I: What do you think now that you have tried it?
P: It was a lot of fun, you did not really think about how you used your body while playing, but afterwards you felt out of breath.
I: So you how did you find the level of intensity?
P: It was an alright level of activity, but it was a nice way to use your body without it doing something very tiring you know? So it was a lot of fun.
I: Anything you liked a lot?
P: I liked that there was multiple rounds, and that they varied in how you had to play.
I: How did that style of difficulty affect you?
P: It was nice to have an intro round where there was no red buttons and the like, just to get into the groove of the game. Helped you to get a feel for it.
I: How did you feel when before you came over to try the game?
P: Sleepy, tired, just had the feeling that today would be a really long day.
I: And now?
P: I would say I woke up more from jumping around, it sounds silly but I really felt that way. Made you go "woo!"

I: Was that something that came gradually or.. ?
P: I felt that it was something that happened right away, by the second button you were really motivated to play. I'm not sure how to explain it.
I: You mentioned intro round and how you liked the intro round. How did you feel about the learning curve?
P: Was decently easy to understand, red is bad green is good you know?
I: Anything that helped to aid understanding?
P: The lights were brilliant but perhaps some more sound, I heard it but....
I: So more sound?
P: Yeah definitely, some big sound effects would be nice.
I: What did make you want to try it out?
P: It looked very cool, with the sounds and the colours and the people here, and it looked different.
I: When you started, what made you want to play?
P: To win, I wanted to beat the guy I played with.

4.4 Questionnaire results

	Totally disagree	Somewhat disagree	Neither disagree or agree	Somewhat agree	Totally agree	No answer
I found the duration of the game to be too short	22,8	24,6	26,3	21,9	4,4	0
I found the intensity of the game too high	36,8	29,8	26,3	4,4	1,8	0,9
I found the level of difficulty suitable	1,8	4,4	5,2	32,5	56,1	0
I feel like I got to use my body	0,9	0,9	1,8	28	68,4	0
I would categoriz this game as play	0,9	1,8	2,6	21,9	71,9	0,9
The sound effects made the game easier to understand	14	15,8	35,1	16,7	12,3	6,1
The lights made the game easier to understand	1,8	0,9	5,2	18,4	72,8	0,9
I consider this activity fun	0,9	0	0	15,8	83,3	0
I would recommend this to friends	0,9	0	5,3	22,8	71	0

Figure 7: Questionnaire statement results - Results in percentage

What we see immediately from this result is that the majority of the players felt that both the game duration and the game intensity was fitting. Most players thought they spent about 3-5 minutes playing from interviews and statements noted during observations. There is still shown that there is an audience for a longer lasting installation, though far less so for a less intensive game. An interesting point here is that while 93.8% did not find the intensity too high, less than 8% of players considered the difficulty to be not suitable, and less than 2% considered it completely unsuitable. Since "unsuitable" would encompass both too easy and too hard, this suggest that the difficulty was in a "sweet spot" for the vast majority of players that took part.

When asked if they got to use their body, if they considered the game to be play, the game to be fun, or something they would recommend, the players filled out overwhelmingly positive responses. Less than 2% expressed that they felt their body did not get use, less than 3% would not consider this a form of play, and a grand 99.1% considered the game fun. From this we can reasonably say that the installation managed the goal of incentivising physical activity, with the players self reporting an enjoyable time that they themselves felt were active. The fact that so many considered it play could indicate that playfulness either is a large motivating factor, or that such installations inherently appeal to people that look for playful distractions in their day to day life. As 93,8% of players report that would willingly recommend it to their friends, the word of mouth could increase the amount of people being

active beyond just random passersby. This were observed to take place during the test period as players ended up returning later with friends that had not tried it out.

Lastly here we can see that of the two ways of signalling to players the progress of the game, lights were significantly more important than sound. This in part was due to a technical issue meaning half a test day had to be done without functional sound, but even when sound was present players reported not consciously noticing it in interviews. This happened even if it was audible to observers that were further away from the installation. In addition the players who had no sound very inclined to give "no answer". Do note that the installation itself was more light-heavy in usage, as it could be played without the audio response but not without the visual response.

	Not at all	A little	Some	A lot	Very much	No answer
Collaboration	36,8	8,8	7	4,4	1,8	41,2
Competition	0	1,8	6,1	28	60,5	3,5
Sound	24,6	26,3	24,6	14	2,6	7,9
Lights	0	4,4	17,5	41,2	34,2	2,6
Body movement	0	2,6	8,8	50	36,8	1,8
Intensity	0,9	2,6	14,9	32,5	47,4	1,8
Concentration requirement	0	0,9	4,4	33,3	59,6	1,8
Unordinary activity	0	4,4	14,9	29,8	49,1	1,8

Figure 8: Questionnaire motivation results - Results in percentage

The points about sounds versus lights shows even clearer here when asked what motivated players. Sound was one of the least self-reported motivating factors, while lights scored pretty decently. No player reported that light had no factor compared to almost a quarter doing so for sound. Players did highly report it being an activity out the ordinary as a motivating factor, so the question is whether the novelty aspect would make or break the long term usage of the installation.

Most players perceived no collaborative aspect as expected as the game was either single or competitive multiplayer, though 22.8% still mentioned it as having some motivational effort. Some players were observed to strategize and support each other while playing while not partaking, so this unplanned player behavior could be the source of these answers. Competition showed itself as the strongest motivator for players according to the questionnaire, which was compounded through it being brought up quickly during the interviews as why they even tried it out. Notable is that no player put competition as a non-factor regardless of mode played. Both the intensity and the concentration requirement were things players reported as noticeable engaging factors, with no 78,9% and 92.9% reporting it as either "a lot" or "very much" respectively. Internal testing of the game before the test period had been reported as too easy and less engaging, so the upped intensity for the test

period to the intensity given in section 2.3 seem to have given players a distinctly better experience.

4.5 Engagement of the installation

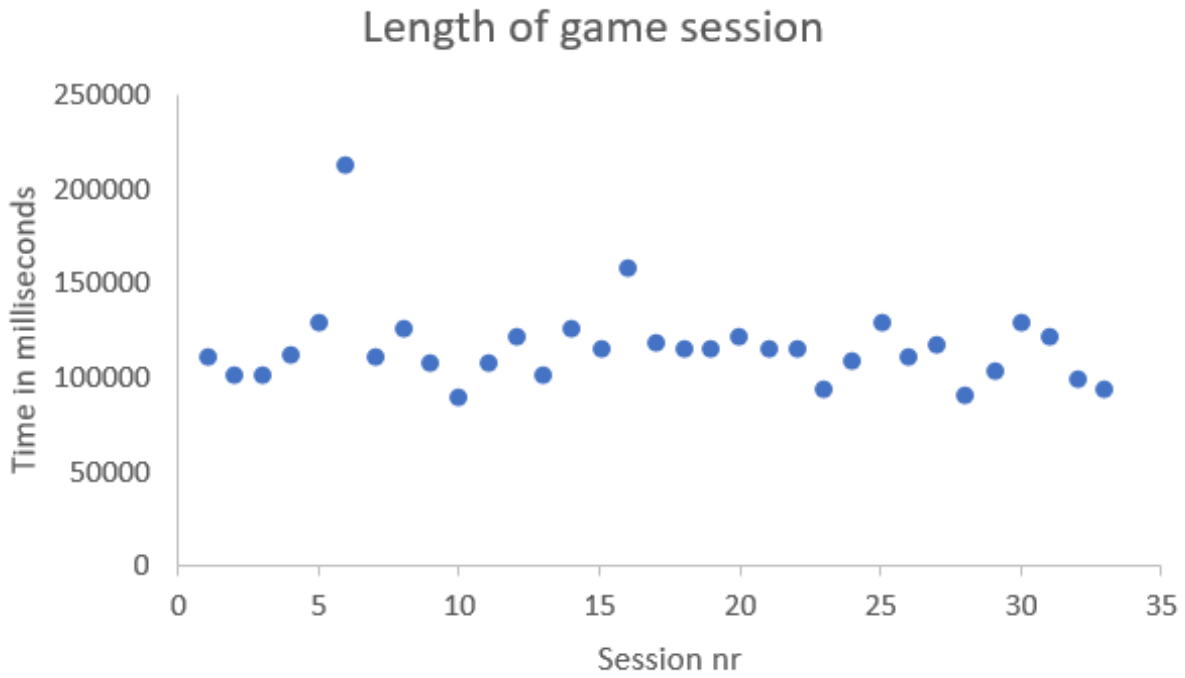


Figure 9: Game session length competitive multiplayer

The session time for the multiplayer games were relatively uniform in length as seen in 9, with an average of 1 minute 56 seconds to play from start to finish. This is notable in that players and observers perceived the game to be over a minute longer than it actually was. As the game only finished whenever all three sets were won, the game could at earliest be finished at 27 points, and at most last to 51 points. As long as one player would not dominate the other the time could be very reasonably adjusted by either lowering or increasing speed of buttons. The outliers in long game time is not due to the other games being dominated, but rather that in that specific game players took a small break in between sets. The game has no mechanics to force you to immediately engage in multiplayer, and these players took advantage of that.

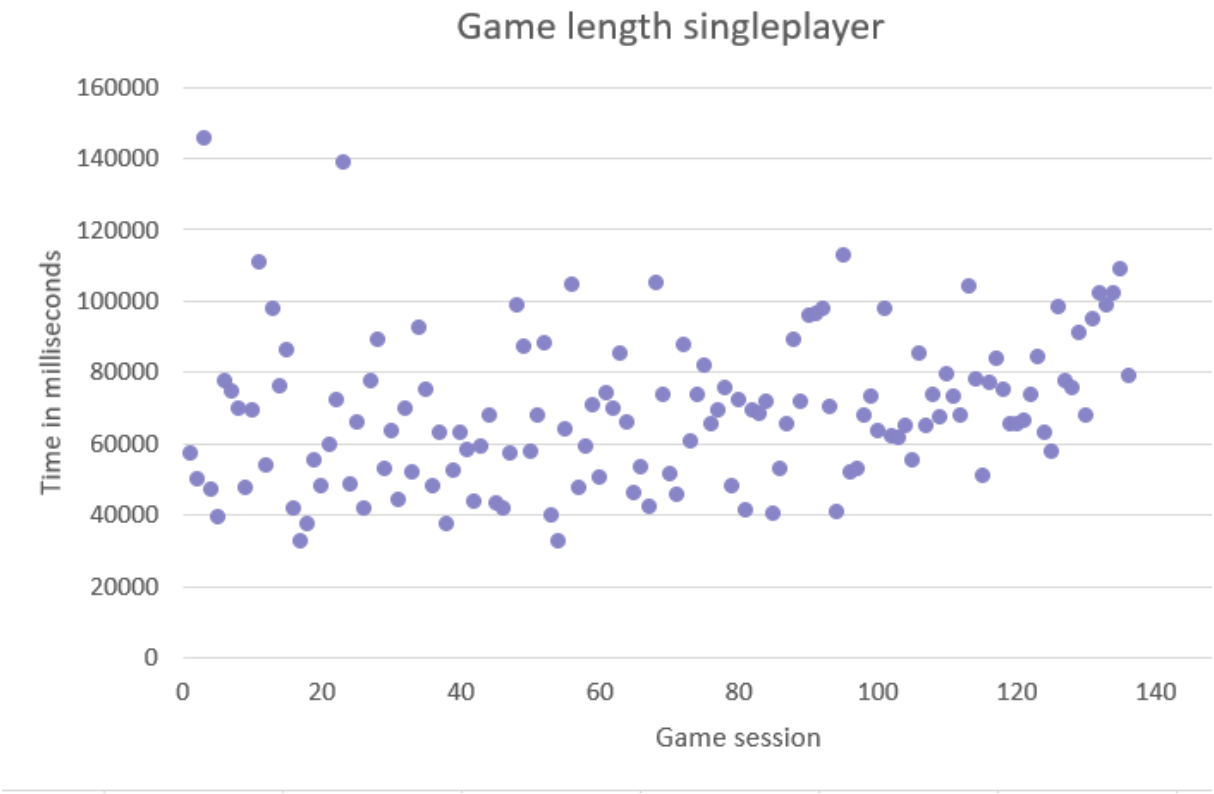


Figure 10: Game session length singleplayer

Singleplayer session varied much more wildly in the time spent as shown in 10, as your time in game was directly impacted by performance. The average game length was 69 seconds, dramatically lower than players themselves perceived at about 3 minutes. Even with the wild spread, you can see a general upwards trend of players lasting longer as players come back for more rounds to try for better score, or learn from observing each other. This is particularly seen in the final cluster. The lowest scores were often the first session in between classes. One thing to note is that there were more unique players and repeat players for singleplayer, even though multiplayer took two players at a time. Question then become if the drop-in/drop-out at any time, combined with the game length being shorter lowered the threshold to take part enough to increase both player count and replayability.

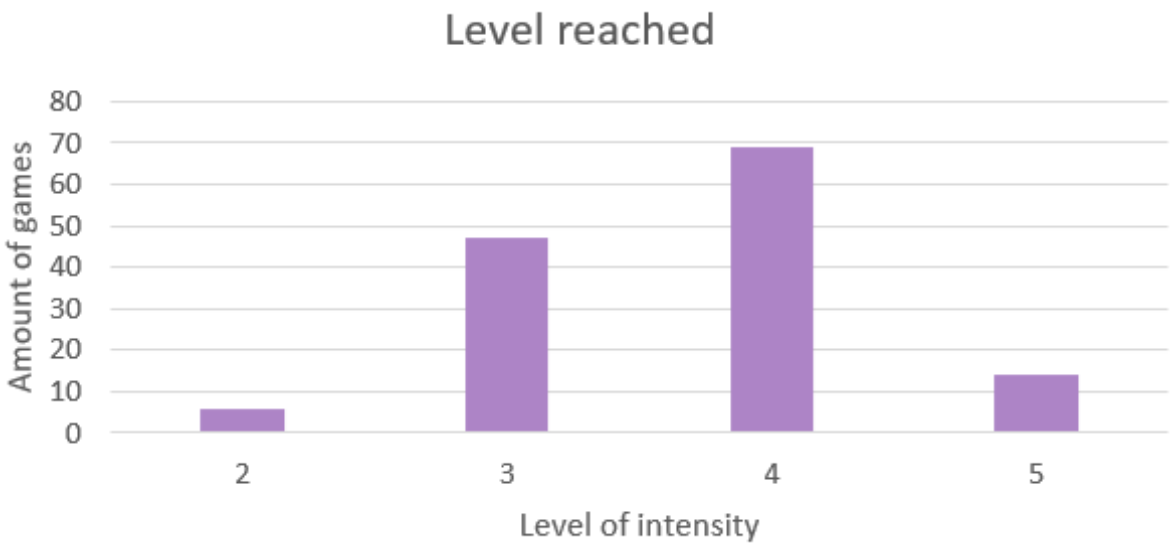


Figure 11: Level of difficulty reached

All players ended up experiencing multiple levels of intensity. Due to game length these intensity levels stayed long enough to not overwhelm but not so long as to overstay their welcome, as was self reported in figure 7. What we can see in figure 11 is level 4 was a good baseline for what challenges the average player, with half the games ending at this difficulty. While few reached the last level in the graph, the last levels was intended to present a challenge to the best performers and thus served its role as none reached level 6, which would significantly increase game time. What is also noticeable is that level 1 and level 2 served their purpose well for introducing the players to how mechanics works without any instructions, as no player immediately lost, and only a handful of the 138 games had someone losing by level 2.

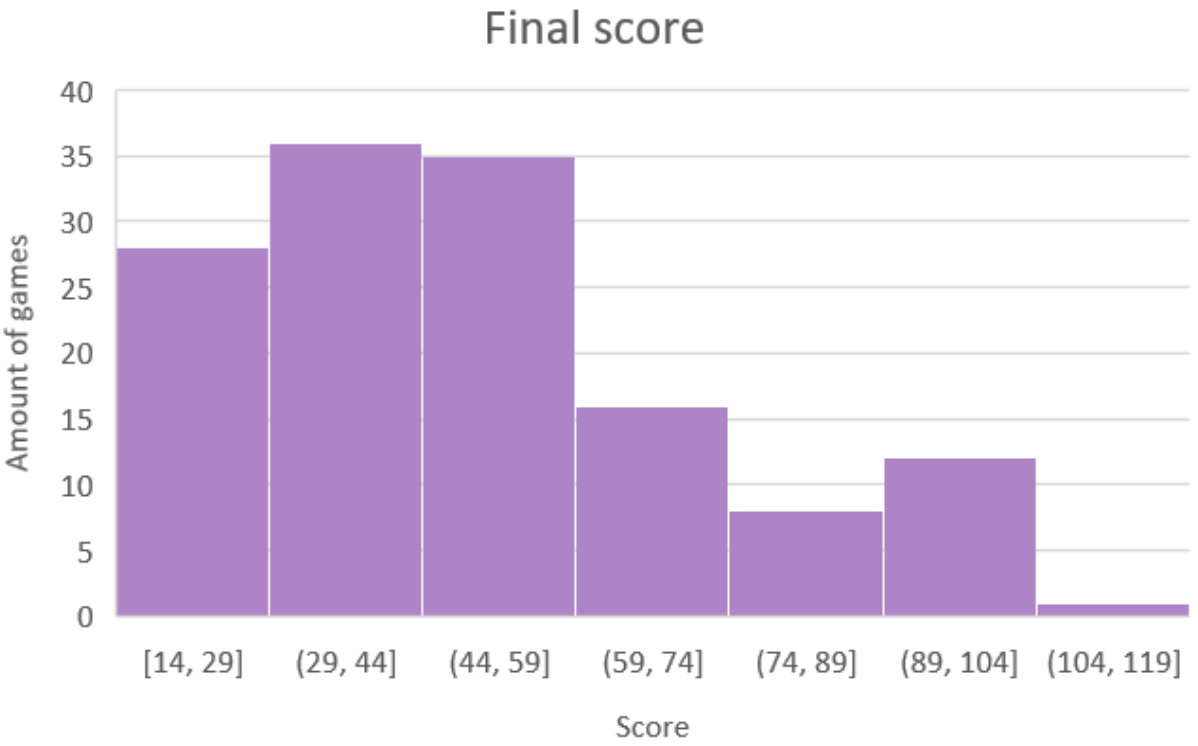


Figure 12: Score reached singleplayer

In figure 12 we can see that all players that lost in level 2 did so in the latter half, as it lasts from 5 to 20 points. This is reflected for level 3 as well, lasting from 21 to 40 points. Thus these phases do not start with too high of a challenge. Level 4 and level 5 has more players dropping off in the first part of these phases, with level 4 having about half the people reaching that level of difficulty losing by the halfway point to next level. Level 5 has a much more significant dropoff, and thus would be the biggest reconsideration for adjusting the difficulty.

4.6 Dimensions of the installation

For the dimensions of the installation used, the big and almost sole consideration was the accessibility of the buttons. The columns were placed at a distance from each other and at a certain height above ground. Due to the distance between them, the installation needed to be judged not only whether players could reach the buttons considering the different heights of the players while simultaneously retaining a movement challenge, but also whether the players could perceive the buttons when they activated.

One metric used to judge this was the fastest button press of the game from time of activation, and the singleplayer result can be seen in figure 13. Meanwhile player 1 and player 2 of multiplayer is in figure 14 and figure 15 respectively. To no surprise, players favoured their right hand column as most players are right handed. This was less pronounced in singleplayer than in than multiplayer. The top buttons were never favoured by player 2, and player 1 never favoured it on their right hand column. What is interesting to see here is that player 1's left hand top button is more favoured than the button below, even though it should be more accessible based on height. The figure with the most inputs, being singleplayer in figure 13, had a distribution that showed that all buttons were favoured for a least some amount of players, even those buttons meant to be a challenge. With all buttons being fast for some players this indicate that spotting the button is less of an issue than reaching the button. This is compounded with players not losing early on in singleplayer. Thus the top buttons, and to a lesser degree bottom buttons, are identified as potential problem in the dimension of the installation. Height chosen to attach column at is then a more likely factor of accessibility issues than width between columns.

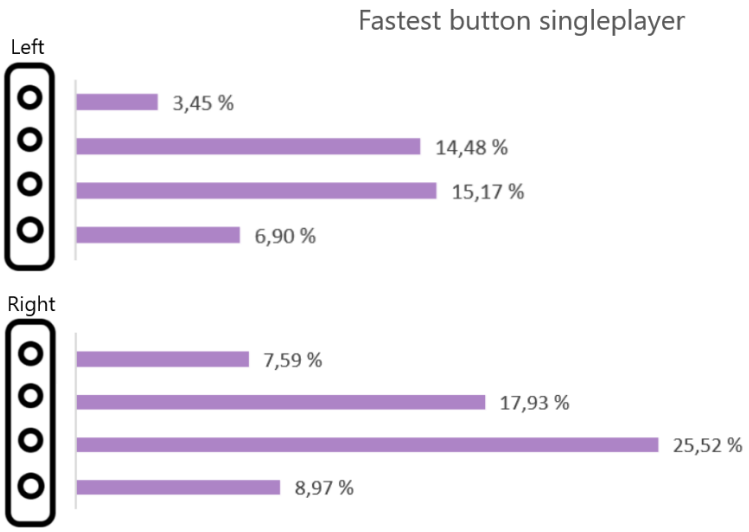


Figure 13: Fastest buttons for singleplayer

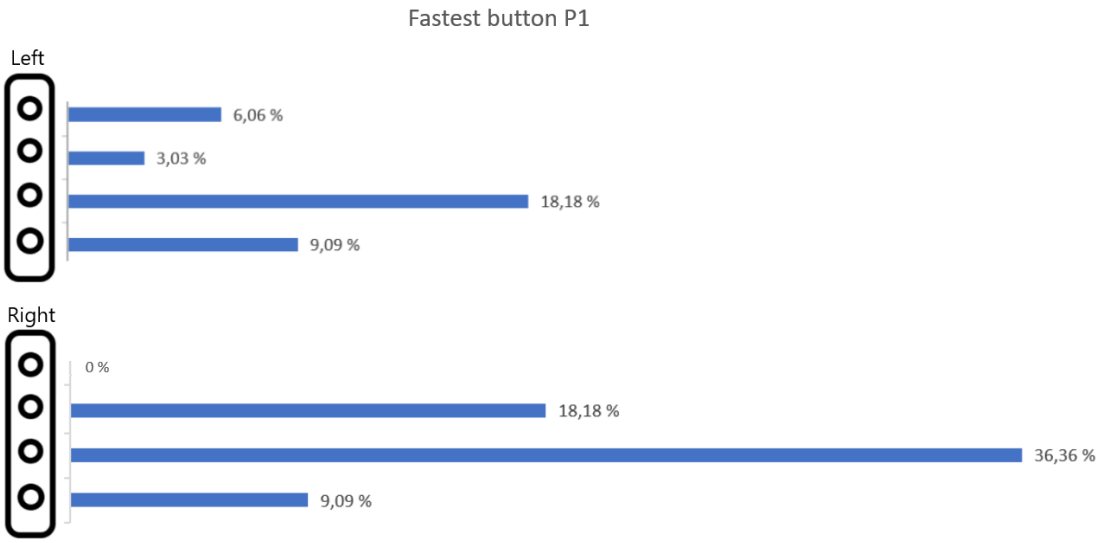


Figure 14: Fastest buttons for player 1 multiplayer

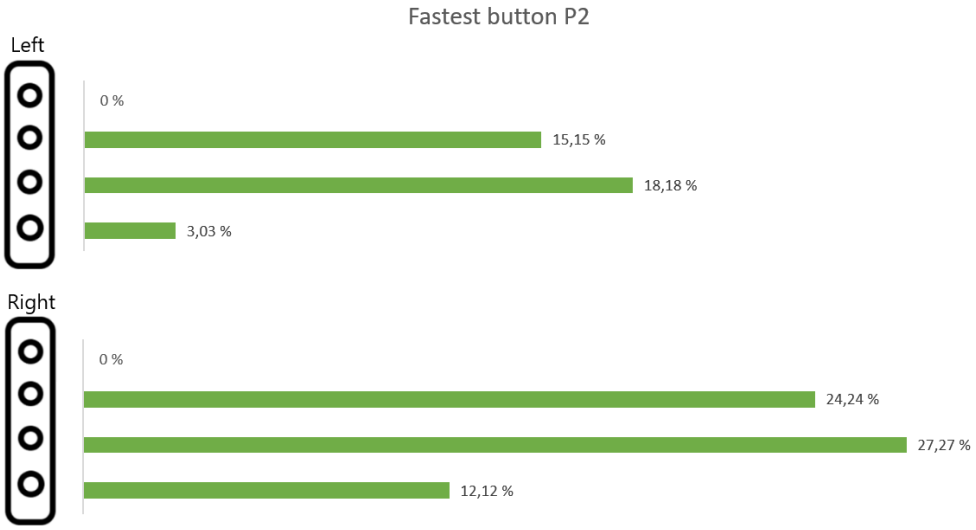


Figure 15: Fastest buttons for player 2 multiplayer

Onward to the slowest buttons, we see results from the fastest buttons reinforced. Single-player shown in figure 16 has players struggling the most with the top buttons on either column, and overall more with the left column. Still interesting is that a sizeable amount of players are slowest on buttons that are considered accessible, which would be unlikely to happen if any one button was far too much of a challenge to reach. In multiplayer shown in figure 17 and figure 18 the bottom button end up as a significant challenge, with exception for player 2 having a much easier time with their bottom right button. Here one of the observations noted during testing can be pulled in about shorter players. These players due to necessity end up jumping to reach top button, taking enough time that it is likely to be worse in such statistics while fulfilling the installation goal of encouraging physical activity. Meanwhile larger players have to bend down, and were more common in multiplayer testing. Lastly, as multiple buttons become active in either mode, both the top and bottom button will necessarily be further away from the other buttons that activate. This makes it so they are more likely to become "the slowest button" if the buttons closer to average players hand height is active at the same time.

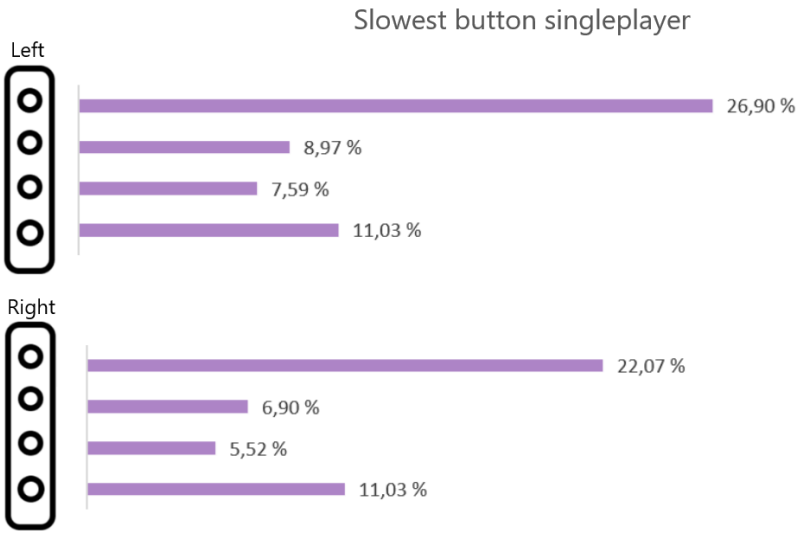


Figure 16: Slowest buttons for singleplayer

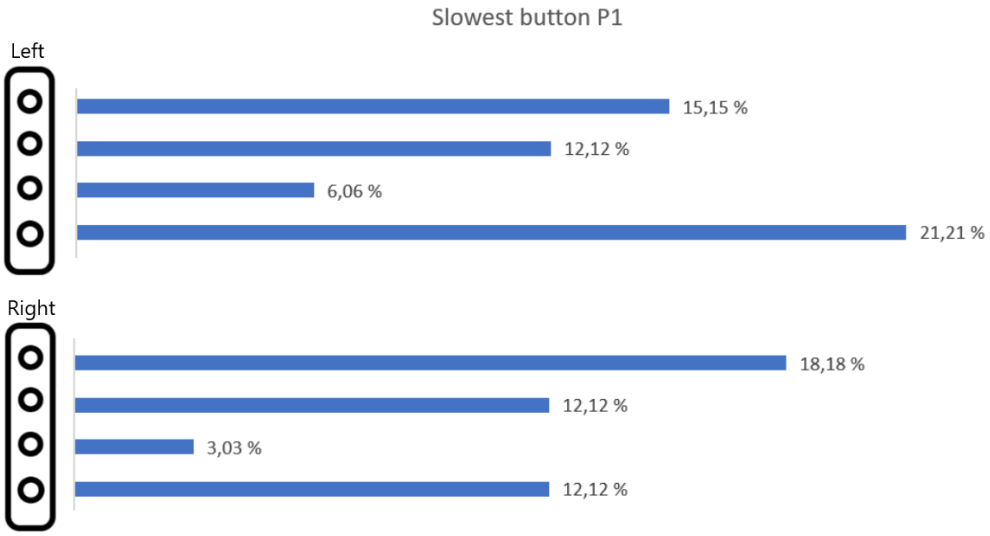


Figure 17: Slowest buttons for player 1 multiplayer

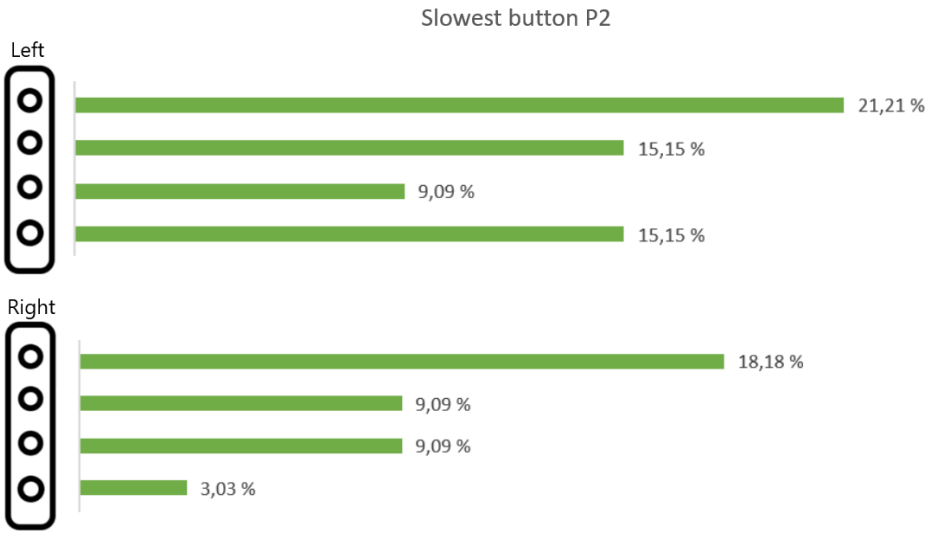


Figure 18: Slowest buttons for player 2 multiplayer

Lastly, we have the buttons that caused players to lose one of their three lives in single player shown in figure 19. These results are in line with the previous results, and is based on the highest quantity of inputs with 414 life losses. This solidifies top button and bottom button as inputs that should be looked at, while not being represented often enough to likely go beyond challenging input into game breaking territory. This is consistent with the questionnaire result that difficulty was suitable, and that the body movement was both noticed and worked as a motivating factor. Thus top and bottom button representation could be an indication of what is going right, and not what is going wrong.

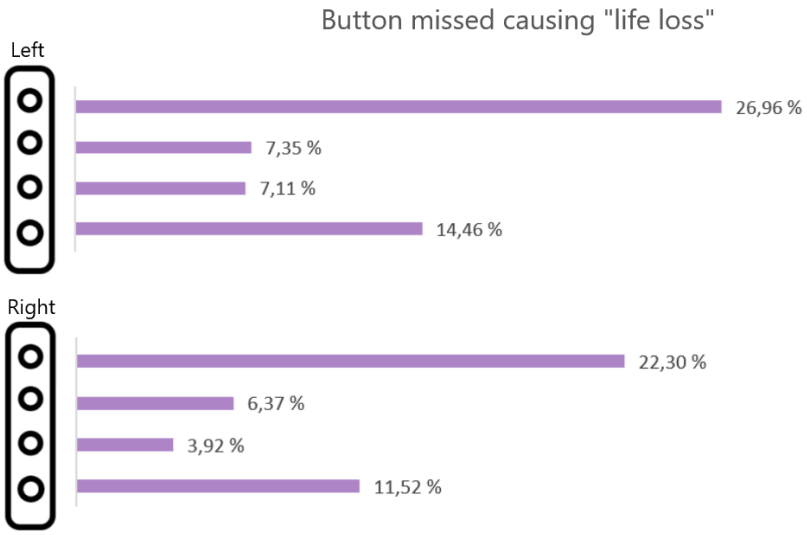


Figure 19: Buttons missed in singleplayer causing "life loss"

4.7 Difficulty curve

Now, the question for difficulty for single player mostly revolves around how hard it is to avoid losing one of their three lives. From timestamps graphs showing how far into the game a life was lost can be seen in figure 20 for the first life, and in figure 21 for the second life of the session. The average point of first life loss is at around 61% of the game completed, with it being decently uncommon to lose life early, or in the very final stage of the game. This indicate a low confusion of mechanics as these are introduced early, and rather the challenge itself becoming high enough that players lose the life.

More over, the average point of second life loss is at 83% completed, which is a decent midway point between first life loss and game over at third life loss (always being at 100% completed). 39% of the game session pass between the first failure and the last of the game, indicating that difficulty does not ramp up too dramatically that player gets overwhelmed. This is good for avoiding the feeling of unfair difficulty while reinforcing the feeling that you could have made if you just tried a little better. Though it has to be noted that in 33 of the 138 games the second life was lost within the last 4% of the game completion. This can either be due to failing often beget more failure, or that level 5 ramps up the difficulty too much and being a significant amount of these games, as indicated in 12. Since multiple buttons do activate at once, it is possible to lose all lives in the same second, but this never happened in any game.

From observations there were a lot of crowd motivation towards the end of the singleplayer testing, as players wanted to be the one that managed to get the furthest into the hard part and go beyond 100 points. Thus, a "too hard" part might non-intuitively increase engagement, as long as it is not presented too early. It also means that best performers are rewarded, while at the same being more entertaining to watch and makes it less embarrassing to fail earlier as even the best can’t last too long. So a high threshold of ending difficulty could paradoxically lessen threshold for trying.

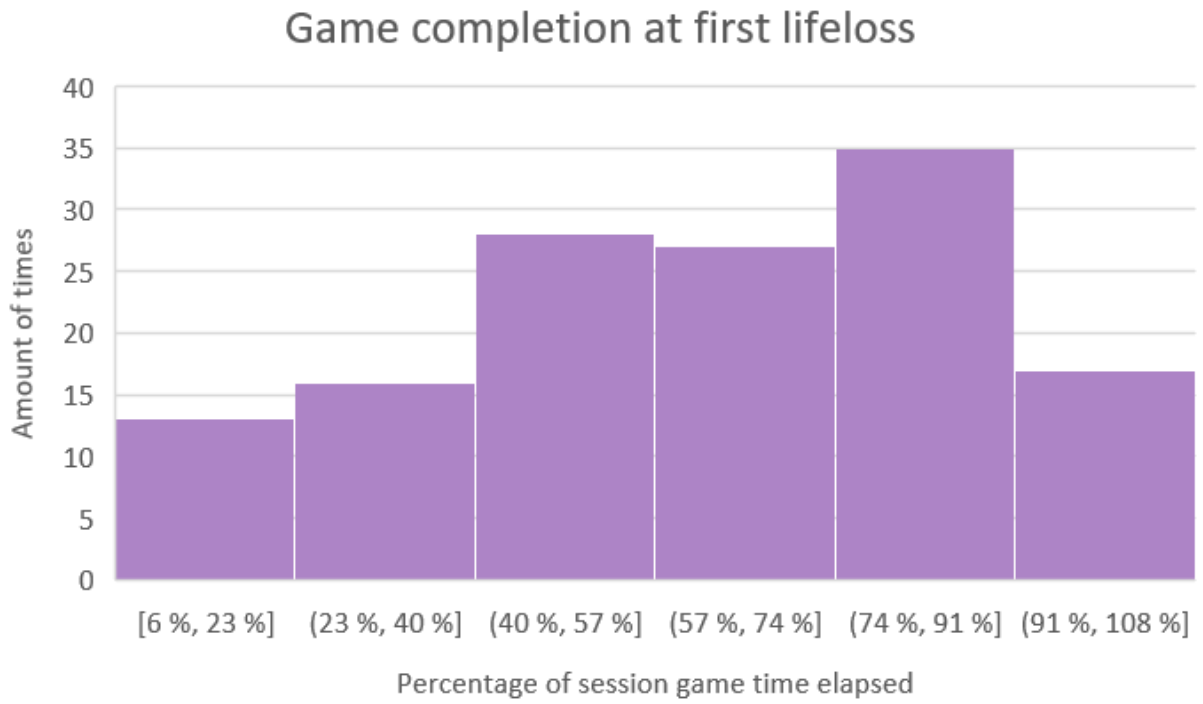


Figure 20: Percentage of game completed once first life was lost

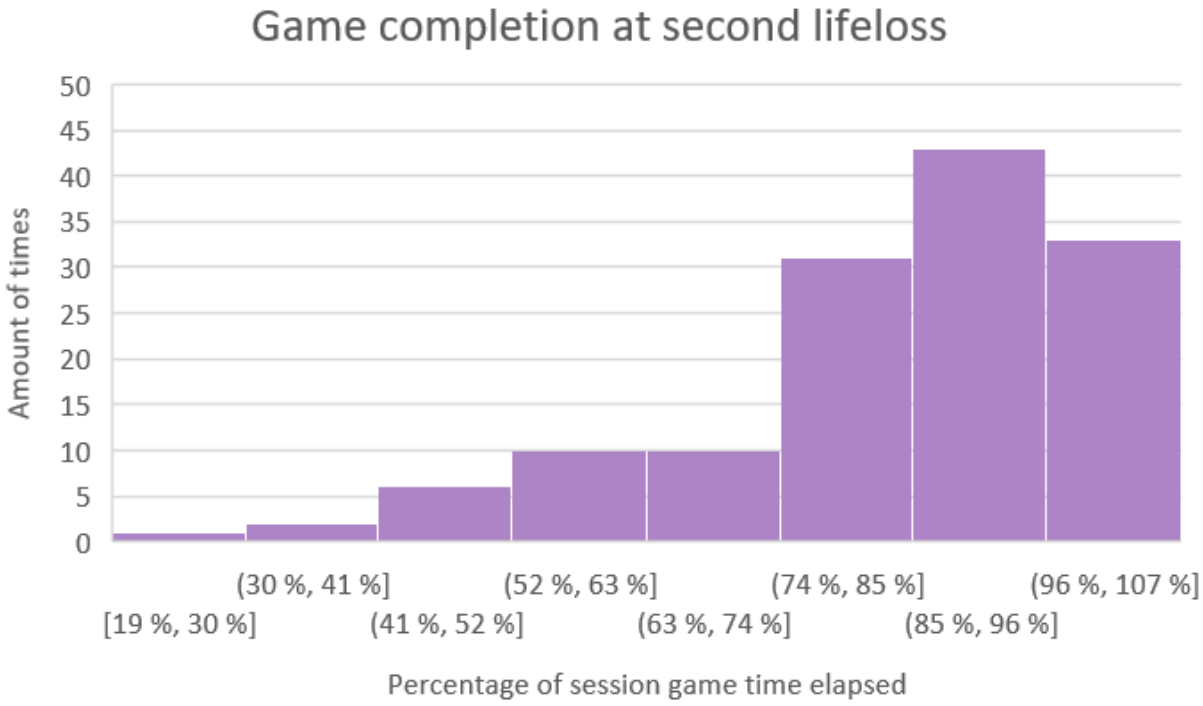


Figure 21: Amount of game completed once second life was lost

In figure 22 you can see the fastest reaction across all single player games, while figure 23 is the fastest reaction for both players in multiplayer. What is interesting to see is that there

is a significant amount of games with low to no variance between player 1 and player 2. In table 1 you can see the average fastest reaction time of the players.

Singleplayer	568 ms
Player 1 multi	542 ms
Player 2 multi	517 ms

Table 1: Average of fastest reaction for each mode

Players have their fastest result being slightly better when playing multiplayer against each other, than when purely single player. Singleplayer is programmed to give less time before a button deactivates than multiplayer, and has the added consequence of leading to an early game over. Regardless, there is a 16 ms speed difference between Player 1 and single player, and a whole 51 ms between Player 2 and the average single player. Just the aspect of having to compete for the button with another player cause both players to perform better even without the other motivational factors singleplayer tries to bring to the table.

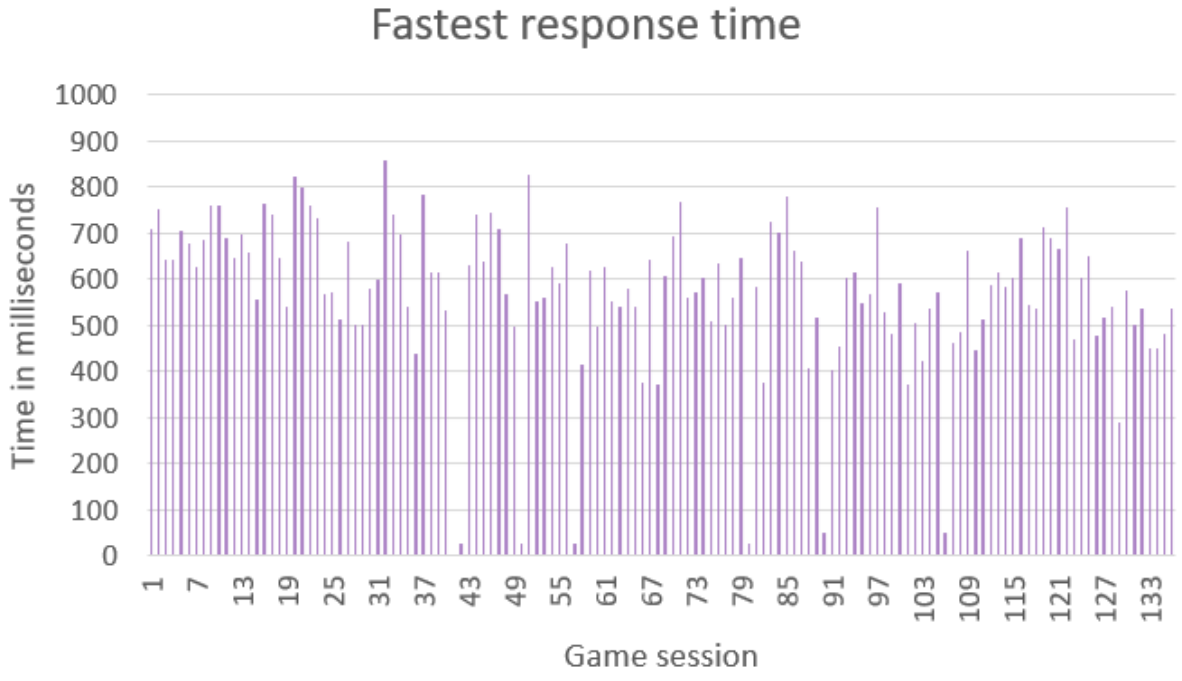


Figure 22: Fastest response in singleplayer

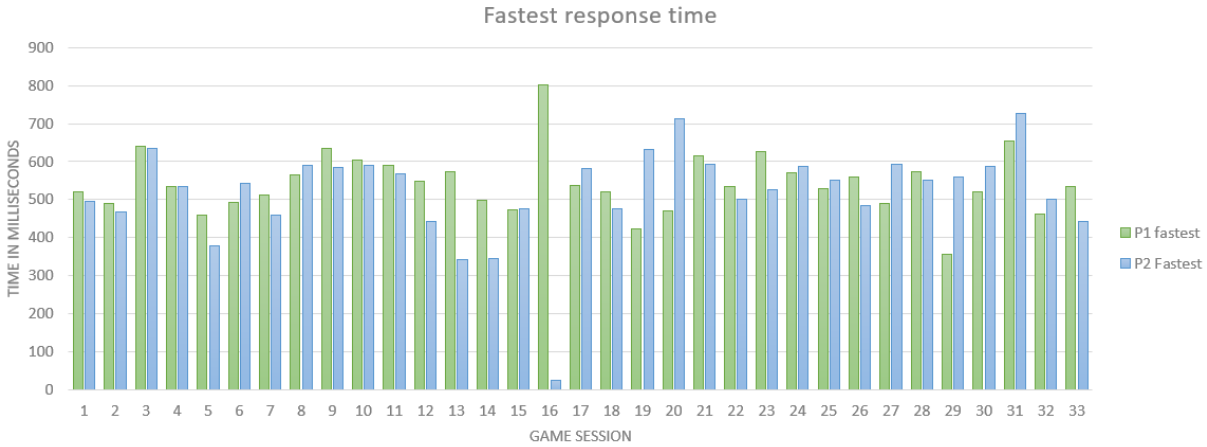


Figure 23: Fastest response in multiplayer

Slowest response hammers this difference home much harder, as figure 24 shows all singleplayers have their worst successful reaction above 1500ms, which is not only common to do better than but almost expected looking at the multiplayer players in 25. A minority of the multiplayer players even have their worst reaction below 1000 ms, significantly faster than even the best in singleplayer. In table 2 you can see the average of the slowest response time for each mode.

Singleplayer	2015 ms
Player 1 multi	1438 ms
Player 2 multi	1495 ms

Table 2: Average of fastest reaction for each mode

Singleplayer at their worst use 1/3 more time reacting to the inputs. The average worst for the singleplayers must also have happened early on, as past 20 points the needed reaction time is 1920 ms to not miss a button. What this show is that single player here give the players enough leeway initially that they can take their time, but your opponent in multiplayer afford no such thing. This means that multiplayer is throwing the player immediately more off into the deep end of the pool learning wise. This was something reflected in the interviews, as players of multiplayer mode mentioned not being quite ready for the increase in difficulty unlike those from singleplayer interviewed. Even so, more than 500 ms is a huge number.

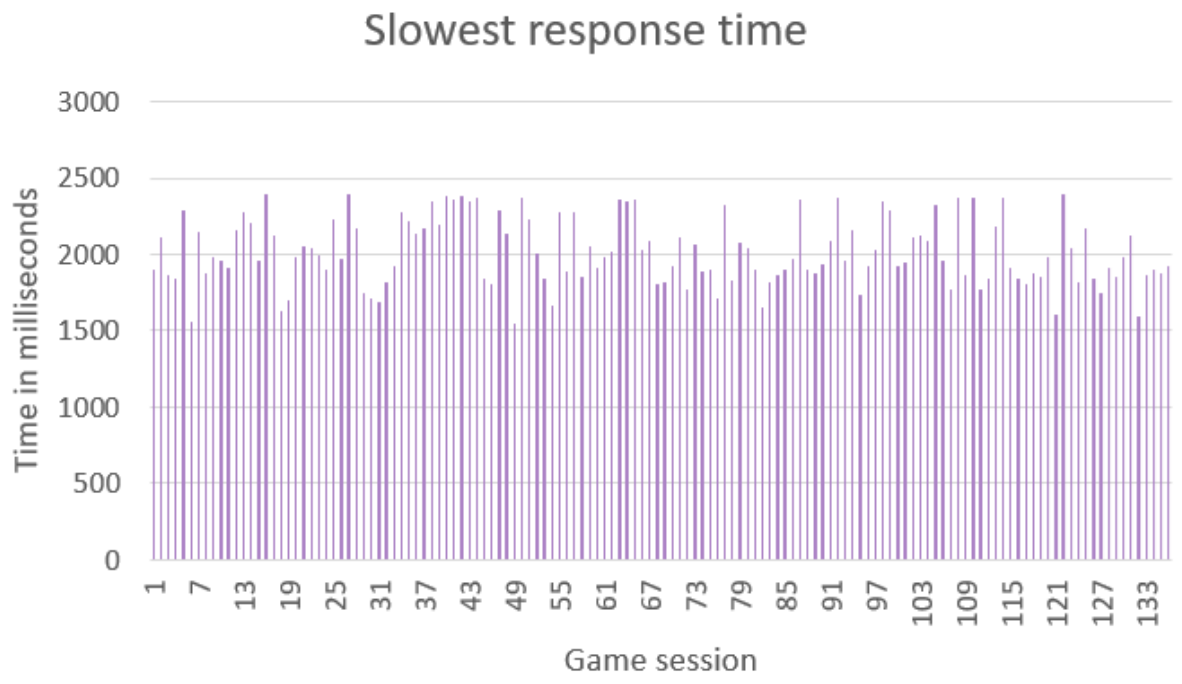


Figure 24: Slowest response that still awarded points in singleplayer

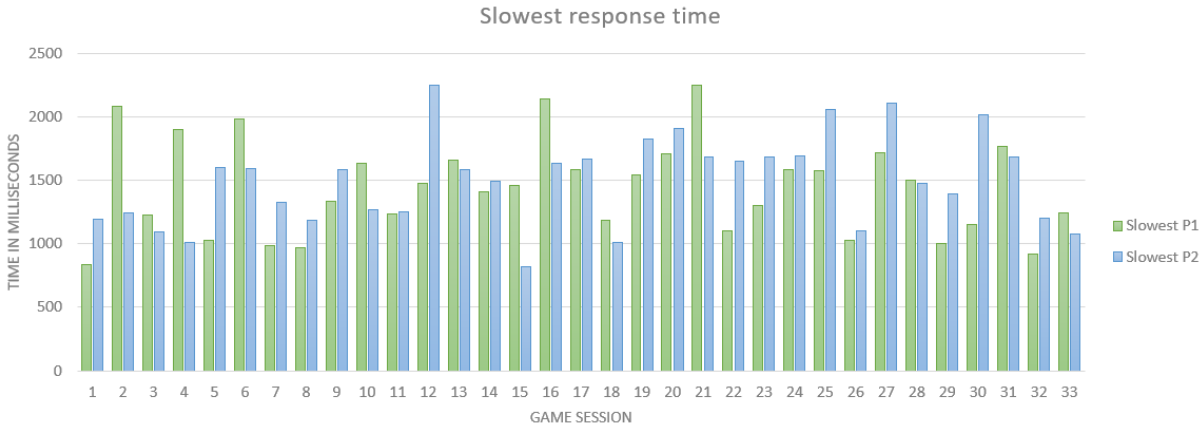


Figure 25: Slowest response that still awarded points in multiplayer

Lastly, we have the average reaction across all inputs done by the player, with singleplayer featured in figure 26 and figure 27. What is remarkably here is not only the immediate visible gulf between the two modes in average reaction, but that players in multiplayer is remarkably good at picking opponents with very close reaction speed. Some of the matches had as low as 1 ms difference between the two, and that is across a game lasting almost two minutes. Question then become to what degree in such competitions one need to accommodate people across the physical spectrum, as players seem to already manage to innately judge who to play against for a fair experience. This could of course be a result of people joining

with friends that are similar to them, and it also leave the possibility open that faster players deliberately hold back to give a more fair fight. In the table 3 you can see the average of these averages across the different modes.

Singleplayer	1150 ms
Player 1 multi	796 ms
Player 2 multi	775 ms

Table 3: Average of fastest reaction for each mode

Once more we can see the singleplayers falling behind, being almost 1/3 slower than multiplayer results on their average reaction time. This significant gulf can again most readily be explained by the competitive aspect driving players to outperform, as players gave it as their most prominent motivating factor in the questionnaire. The results seem to indicate that to drive the players to top performance, one need competition and one in real time against opponents. The latter is because all that played singleplayer indicated that competition was a motivating factor, and were noted under observation to announce their desire to not only beat their own best, but also their friends best. Response in interview of what the ones that tried multiplayer thought of the singleplayer was that it would be less motivating as they were not directly engaging with their opponent. Notably during multiplayer of this installation, the players would be facing each other and thus be aware of where the opponent was and what they were doing at most times. Regardless of reasoning, 354 ms difference between Player 1 and singleplayer and 375 ms between player 2 and singleplayer shows a much higher intensity of play when against another player.

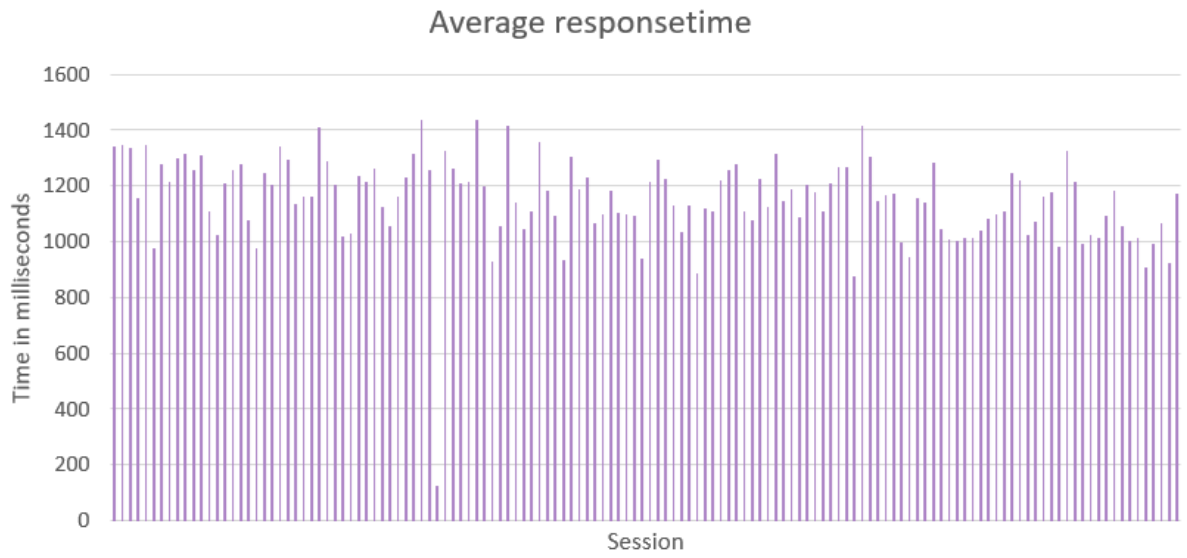


Figure 26: Average response in singleplayer

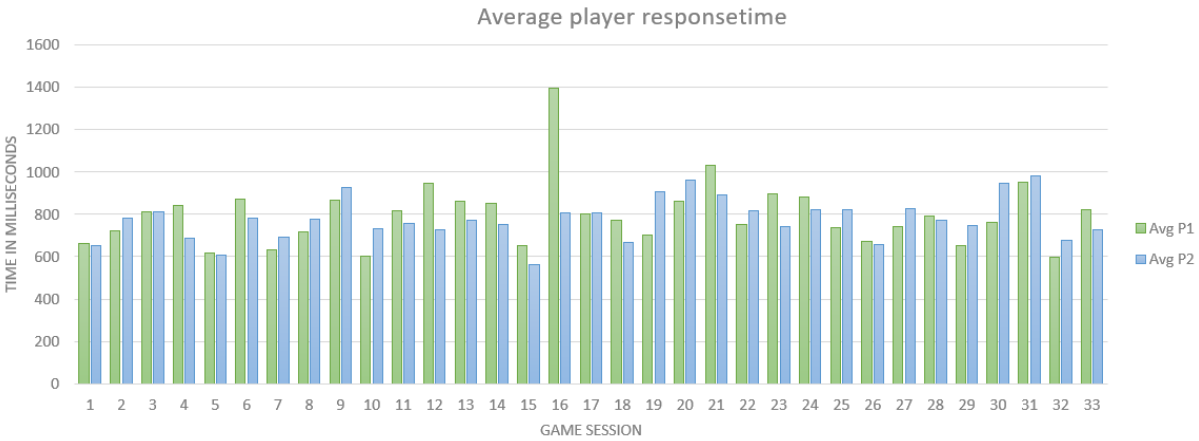


Figure 27: Average response in multiplayer

4.8 User understanding

As an obstacle a glowing button was introduced some way into each game that would reduce the score if pressed. This was to shake up the gameplay and make it more varied. Normal buttons that gave points glowed green and counted down, while red did not change count down and was unchanged for their duration. The red buttons were introduced past 5 points in singleplayer, and after first of three sets is won in multiplayer. During observation of both modes, some players expressed that they had red-green colourblindness which the design had not accounted for. In figure 28 one can view the amount of buttons pressed in multiplayer.

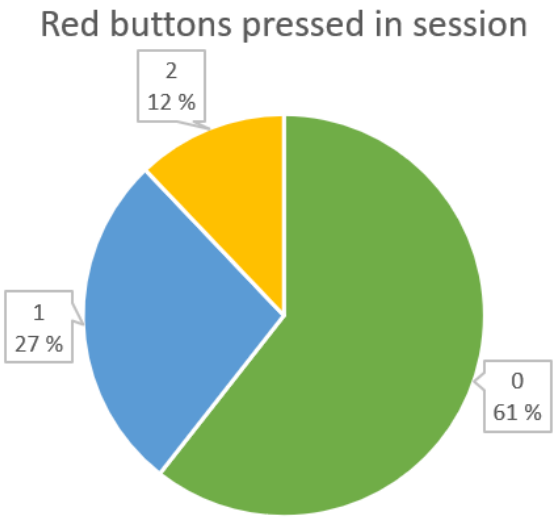


Figure 28: Amount of times a red button was pressed in multiplayer

As one can see, in 2/3 of the games neither player pressed the red button. As every game lasted between 24 and 45 points past set 1 based on the log, the chance of seeing no red button in any game less than 8% at the low end of score, and 0.9% chance at the high end. Thus in the games with no red buttons pressed, it is reasonable to claim these are due players avoiding the red buttons successfully. This is supported by the observations, as red buttons seemed to appear at the expected amount. Even so, players do end up failing to avoid the buttona and do end up losing score in the remaining third of the games.

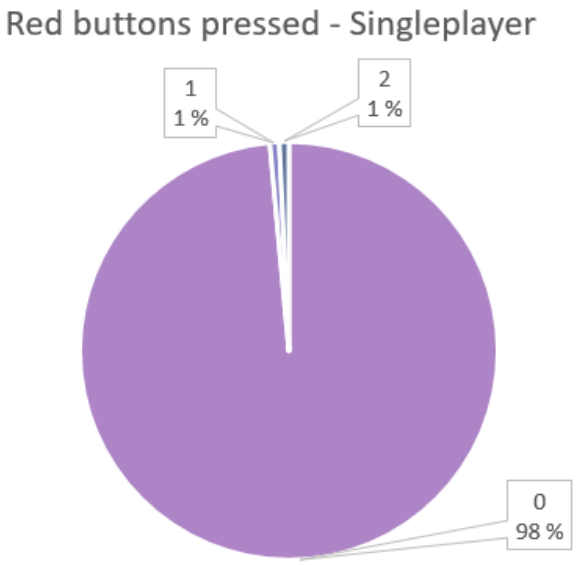


Figure 29: Amount of times a red button was pressed in singleplayer

Viewing the chart for red button presses in singleplayer shown in 29, a radically different results is gained. Of the 138 games, only 5 games had someone pressing any red buttons. Almost no player would have lost any points from red buttons, and the red button gained a radically different function. From observations you saw that as singleplayer gave players less time between sets of buttons activating, so these moments functioned as welcome breaks from the intensity, which was a functionality brainstormed during creation of the installation with another button functionality that was eventually scrapped. The randomness also made it so you never knew when it arrived, making any moments of respite a surprise. This result might explain the slower average responses in singleplayer discussed from 3, as players allowed themselves more time to think and judge what appeared, rather than having to beat both the game and another player on top.

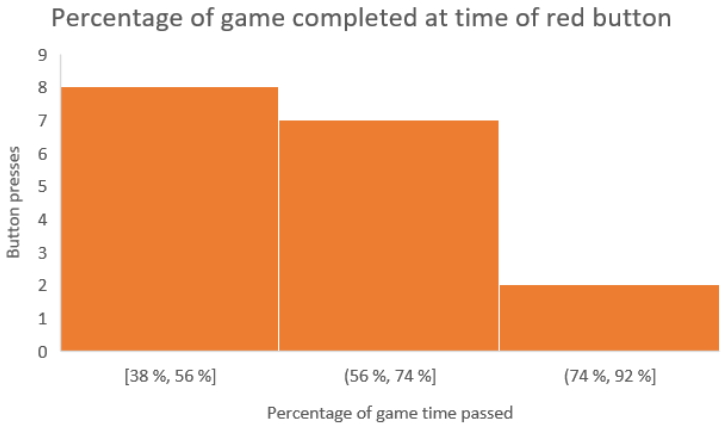


Figure 30: How far the game had progressed at the time of the red button for multiplayer

What can be seen in figure 30 is that in multiplayer where pressing the button happened in a sizeable amount of games, there was also a very present learning curve. Players predominantly made the mistake early on after the introduction of the button, with almost no incidents in the last part of the game. This shows, combined with the results from singleplayer, that red-green colourblindness was far less of an issue than initial observations had made it seem. Since the buttons were differentiated not only by colour but also by how they lit up, they were easy to differentiate regardless of colourblindness.

4.9 User relations

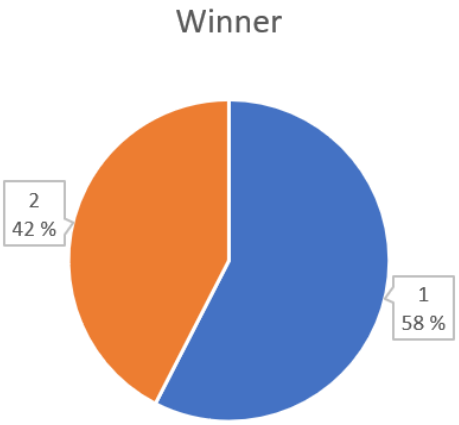


Figure 31: Which side of the installation won

When it came to the winner, both sides had decent chance of winning as can be seen in figure 31. What this does not show however is how sizable each of these victories were. In figure 32 we can see the score of player 1, and it shows that player 1 in 15 out of the 33 games won every set, as they got 9 points in each of the 3 sets. The worst performing player 1 had only 7 points by the end. Meanwhile, as seen in figure 33 player 2 was far less likely to dominate set victories. The worst player 2 had in addition a worse result than the worst player 1. Overall both player 1 and player 2 saw the player still getting above 2/3 of the points available to them in the majority of the cases. While players reported it to be a pretty fair fight, the domination of player 1 means that potential environmental factors like light condition could give player 1 some advantage. There is also the possibility that the randomness function of button selection favored the right hand side of player 1 (and thus the left hand side of player 2), which as shown in earlier graphs would give player 1 an advantage due to most players being right handed. This did not seem to be the case from observations and review of the code, but still remains as a possibility.

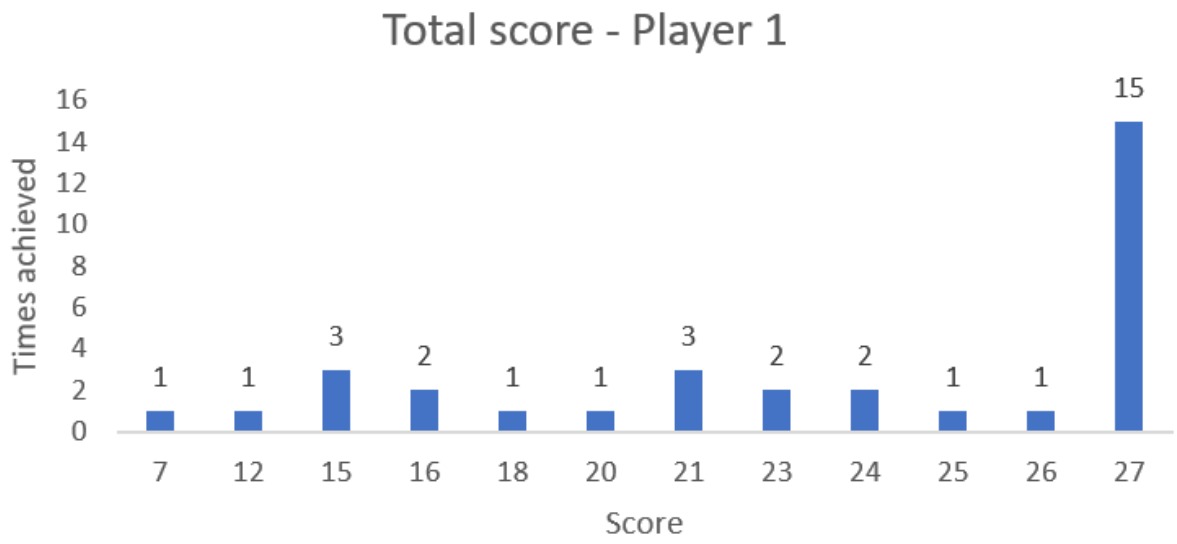


Figure 32: Total score of player 1 at the time of game completion

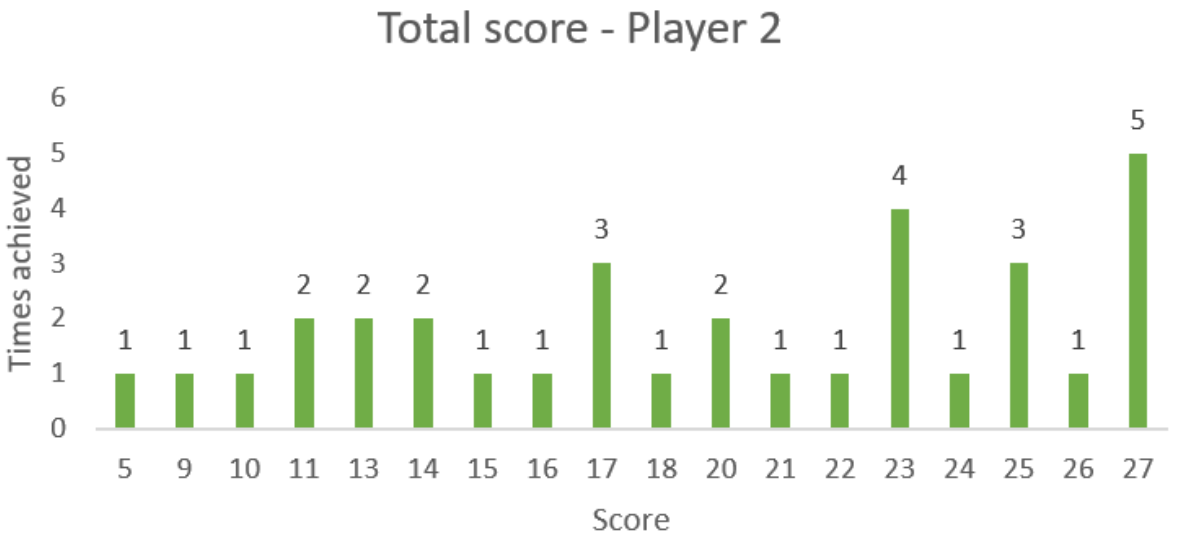


Figure 33: Total score of player 2 at the time of game completion

There was several metrics that could be compared to figure out who the winner was likely to be. The easiest assumption was that the winner would be the player that had the average fastest time of the two. But figure 34 dispel this notion, as the winner were more often than not the player with the worst average reaction on all their presses.

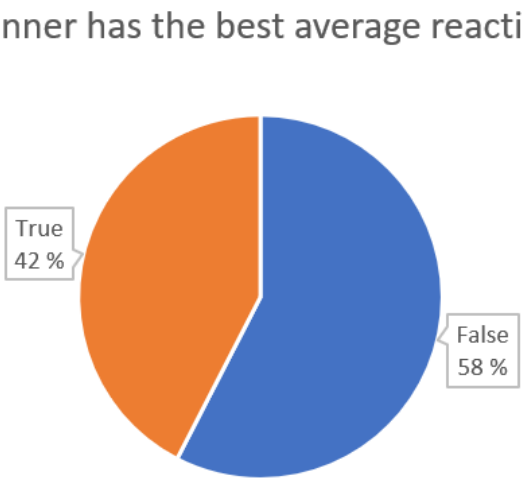


Figure 34: Whether the winner had the best average reaction

Furthermore, while close, the player who had the best reaction of the entire game as shown in figure 35 was usually also the loser overall of the game. While players often self selected players with relatively close average reaction, their fastest reaction were not nearly as close.

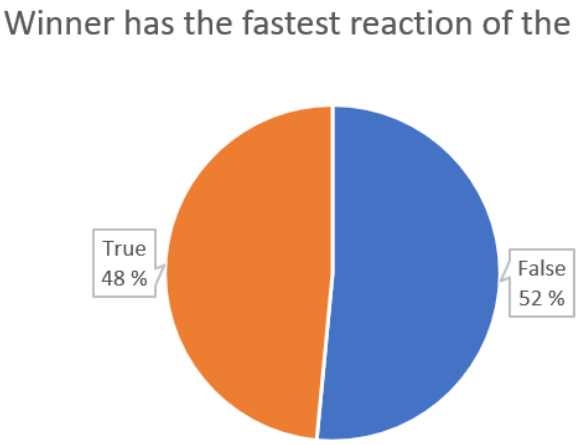


Figure 35: Whether the winner had the fastest reaction

Lastly, figure 36 shows that the winning player was twice as likely to have the slowest successful reaction in the entire game. Combined the results seem to indicate that the fastest player often ended up getting tunnel vision, losing points because the other player was better at taking in the whole playing field. And more than anything it shows that the game does not give undue advantage to the very fastest of participants. The later parts of the multiplayer game saw two buttons activate at once. Based on the numbers here, the fastest player usually won the first button in a close competition, while the slower was better at getting the second button at a higher rate than fastest player won the first.

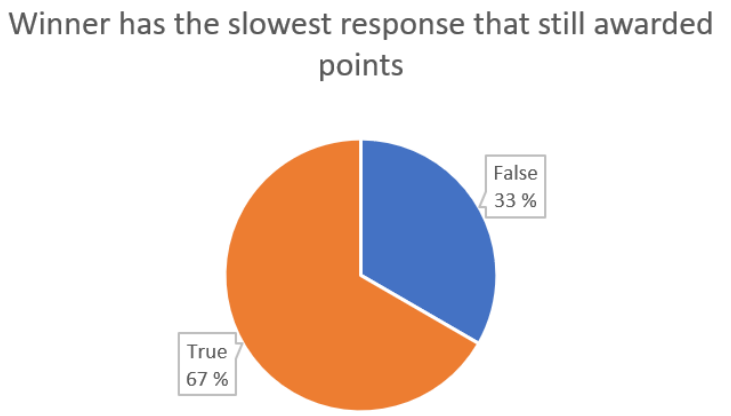


Figure 36: Whether the winner had the slowest time that still awarded points

5 Discussion

5.1 Design lessons

More than anything, what we have seen is that players push themselves to react significantly faster if there is another player involved, averaging around 500ms faster for their worst inputs and 350 ms on their average input. This also comes at a shorter time to comprehend what is happening. The player thus can be said to grow more reliant on instinct. One thing this indicates that for those wishing to design installations that challenges players reaction they would do very well if focus is given on including competitive aspects between players, as this did a far better job than solid single player challenge curve. Players pushed each other harder than the game could, even to the point that a player hurt themselves trying to beat the other in multiplayer as mentioned in observations, something that did not happen in singleplayer that featured more games played. Extra care should also be given to proper introduction to game mechanics in such competitive games.

Another part is the difficulty. Motivational factors were predominantly reported to be related in some form to the challenge it provides. The players were shown to adapt, as the slowest input were impossible to have happened any time but earliest part of the game in single player. Players first "life loss" came on average too late in the game for players to not have adapted to the intensity given, and with the life loss being spread around (average at 61%, 83% and 100% of game played for each of the three lives respectively) the game over was likely not due to incredibly sudden spike of difficulty the player could not cope with. A too big focus on being accessible can then very well end up being detrimental to the end goal of the installation, as player do not feel challenged. From what seen here one should have the willingness to push the player, and they will more often than not adapt and enjoy the experience more due to it. Enjoyment very much aligned with players responding positively to intensity and difficulty in the questionnaire seen in section 4.4 after all.

For multiplayer this can be tricky, as losing heavily is of course demotivating. But as seen players choose opponents that are close if not their equal in at least reaction. Though there is the possibility that some players "toned down" their performance to allow the other players a chance, this is less likely as both still performed noticeably better than even the same player in singleplayer. What can be done to motivate is to allow a catch-up mechanic, as was slightly unintentionally included in this installation. Here the randomness dictates that when two buttons activate, it has a higher chance that they are on two different columns. This is due to there being 3 buttons to select on the same column while four on the other column, and each of these have same weighting. Players can then see if their opponent is heading for one column and choose to aim for the other to keep up.

5.2 On the value of metrics

What working with the data showed, more than anything else, was that the metrics was incredibly good in tandem with other data gathering more than stand alone. The advantage was low effort to record, and low effort to decode in comparison to things like the interviews or even the questionnaire. This was true even though all these had lower amount of entries. Metrics are useful due to scaling incredibly well, being the most quantitative of the data types.

As seen with the results on the dimensions of the installation, it was really good at *identifying* problem areas, and the same could be said about the spread of players across difficulties. But whenever it gave an indication of "this might be an issue", it alone would have left open ended questions as to any confirmation if it truly was an issue, and if so why. It leaves more up to guesswork if all you have is numbers. The questionnaire and the observations were perfect then to answer the why given by the data. It showed that these potential issue areas more so than not worked not only as intended but to a good effect. Acting on metrics alone could have given an overall worse product. This is because if everything is working properly and well, the metrics become harder to interpret. What that means, is that metrics are stronger at showing deviance.

The inverse was also true, however. The qualitative approach often gave vague results that needed the numbers to explain how whatever was said was reflected in reality. A good example was the that players answered that competitive was a core motivation, even the most important one, but without the metrics what this meant for play was vague. What the metrics showed was that players put far more effort into reacting fast, but this came at a cost of placing the player immediately into action. Little room was given for onboarding as the chase of winning over the other player meant that players gave themselves far less time to process. Without the metrics how to deal with competition being motivating would be hard as players of both modes gave it as answer. Due to metrics, we can say "we need to look at how we can introduce the mechanics of play to multiplayer better" and "we need to look closer at how we can make players of singleplayer play at intensity closer to multiplayer".

Lastly, the metrics more so than the other forms did really well at challenging the preconceived notions to the different elements of the design. The vastly different results of the identically acting red buttons across the two modes was not readily noticed through any other data, but going back and looking at the filmed game sessions with the result of the metrics in mind suddenly made it clear that the different modes handled the same thing very differently. The idea that the fastest player was also the likely winner also seemed like common sense, once more debunked by data not seen in the more qualitative forms of data gathering.

5.3 What could have been done different

While working with the results, it became clear that there were gaps in the data gathered. The clearest knowledge that could have been bridged with more data was the lack of record of several inactions on the part of the player. All the data of the red buttons were only tracked if the player actually pressed it. Even though we know the mathematical odds for red buttons appearing, this introduces an element of uncertainty. If some issue caused them to appear at a lower rate or higher rate it would affect the way the data would be viewed. And though there was a vast gulf of difference between how singleplayer and multiplayer treated the button, this difference only measured how often they were pressed in each and not how often they were ignored.

This also goes for recording where green buttons appeared. We know how many were pressed from the score, and we know how quickly and which buttons were favoured, but the actual percentage split between how often certain buttons activated meant that we were working on an assumption that the system itself works as intended. This is an example of a "debugging" metric that would be useful to rule out possible factors on the result. The system was designed to have all buttons be equal chance of activating and seemed to follow this rule for both players based on observations. But if some issue caused the right hand column to be favoured for one player, that would give them an advantage. As it stands now we cant fully rule this out as why one player side won more on average.

This is perhaps the most important realisation, that keeping metrics used for debugging can be useful for maintaining that everything still works as designed. Public installations often gain wear and tear as they get more used, and thus removing these metrics once the design period is over (or not using them at all) is harming the possibility for accurate analysis.

5.4 On the designing of metrics for public installations

As said during the methodology, each installation and every installation needs to be judged on their own merits to see what would be needed. That however does not mean we can't try to generalise how this can be done. When designing the metrics, three questions to ask yourself in order, and after reviewing the results of testing for this installation they worked incredibly well for quickly finding solid metrics.

The questions are the same as in section 3.1, though here with some supplementary questions.

1. **What is possible to measure?**
   Find out what inputs you take in, and what internal logic (if any) that is worth measuring, both before creating output and as finished output. Is there some important data that cant be measured with your current installation, and is it worth to expand to be able to measure it? Is there something important you know that can happen, but don't currently know when?

2. **What is reasonable to measure?**
   Is there any privacy concerns with what you record, would you be able to connect a result to someones identity? Is there something you could measure that would overwhelm other results while giving little in return? Is there data you can gather that you don't have the means or knowledge to use?

3. **What is relevant to measure**
   Is the data you can record relevant to the goal of the installation? Is what recorded something that is necessary for the entire installation, regardless of result? Can the data at all help fill any gap in knowledge?

That section also mentioned the data groups relating to aspects that this installation used. Using these categories as a list to see if you captured the aspects relevant to installation could be useful, as finding them certainly helped here. The expanded list is shown below, based on lessons learned from section 5.3.

- **Dimensions of the installation**
  Relates to the physical size and accessibility of the installation. Could also relate to screenspace on software.

- **Engagement of installation**
  Metrics on usage, often length and recurrence.

- **Difficulty curve**
  Metrics on challenge and drop-off.

- **Understanding**
  Metrics measuring how and when mechanics gets used
- **User relations**
  How does users interact with others, and with who?

- **Environment of the installation**
  Metrics covering conditions of the environment around the installation, like light, wetness and time of day.
- **Debugging**
  Metrics intended to check if logic of installation works as designed.

5.5 Future work

There are plenty of avenues to continue on, as one can pick most of the aspects listed above in the design of metrics to look more in depth. These are less in the form of how to make metrics, but rather how we can utilise the metrics to their fullest.

One of those would be the engagement. This is a heavily researched topic, but generally focuses on trying to keep the user as long as possible. The installation here however kept players in short bursts, as the audience that tested generally did not have much time to be spent on it. A look at length of play vs return of physical activity would be a good topic. Does shorter bursts cause players to engage more often, and cause more people to engage? Single player for the thesis installation had significantly more players returning to play and more people trying across the same time period. This even when the mechanics were mostly identical but the game shorter compared to multiplayer.

Another aspect would be the environment. Do outside or inside installations give better success? Is there correlation between time of day and who ends up using it? How do you design an installation that stands out from the environment enough and looks inviting enough for people to try? These were aspects lost while keeping the installation in one spot. For multiplayer the light conditions were one player ended up with sun shining at their back, while the other would having it at their front could very well explain difference in player results.

Lastly, one could take a look at whether we can design metrics to the point where we can minimise the need for qualitative data to lessen manpower needed and have better scaling on data gathering. The conclusion here was that metrics provide a significant half of the puzzle, but more often than not need qualitative data to give a proper picture. Can we then design metrics that will do a far better job at providing the whole picture?

6 Conclusion

This thesis has explored the value of using metrics when it comes evaluating public installations. These are often heavily influenced by qualitative over quantitative data, but as we have seen the metrics ended up uncovering aspects that flew under the radar by the other forms of data gathering. There was little to no indication outside the metrics that multiplayer significantly caused players to up their reaction speed, much like they did not say much about the vastly different treatment players had for the red buttons. These were both parts that mechanically identical across each mode they were tested on, and comparatively low effort to track with metrics.

From this, it feels safe to say that metrics is not only a valid but core part of being able to evaluate any such installation accurately. As introduction said, guidelines as to how one can approach making metrics for similar installations were made from the experiences of making and evaluating the ones from this installation. Using the questions to gradually shrink your potential metrics helps not in making sure you have covered the aspects you wish to cover, but also to highlight what to focus on in the more qualitative data gathering to give the whole image. There are a lot papers exploring metrics, but few beyond a financial view. Reviewing the way we use this data while also playing the part of the designer has been a fantastic learning experience both for the data understanding perspective but also from the perspective of designing what people interact with. And the technological progress means that without being able to use metrics to learn about what we create, it will be hard to compete against the multitude of sources ever more fighting for our attention from those that can use them.

References

[1] Max Roser and Hannah Ritchie. “Technological Progress”. In: *Our World in Data* (2013). https://ourworldindata.org/techno progress.

[2] Neville Owen et al. “Too much sitting: the population health science of sedentary behavior.” In: *Exercise and sport sciences reviews.* 38.3 (), pp. 105–113. ISSN: 0091-6331.

[3] Emmett B Keeler et al. “The external costs of a sedentary life-style.” In: *American Journal of Public Health* 79.8 (1989), pp. 975–981.

[4] Rick Kazman Len Bass Paul Clements. *Software Architecture in Practice, Third Edition*. Addision-Wesley, 2012. ISBN: 0321815734.

[5] Margaret Rouse. *DEFINITION: metric*. Apr. 2005. URL: https://whatis.techtarget.com/definition/metric. (Hentet: 02.05.2020).

[6] Amy Van Looy and Aygun Shafagatova. “Business process performance measurement: a structured literature review of indicators, measures and metrics”. In: *SpringerPlus* 5 (Dec. 2016). DOI: 10.1186/s40064-016-3498-1.

[7] Tom Tullis. *Measuring the user experience : collecting, an- alyzing, and presenting usability metrics*. eng. Amsterdam, 2013.

[8] Martin Klubeck. *Metrics*. eng. 2011th ed. Springer Verlag, 2011. ISBN: 1430237260.

[9] Laura Tyson. *How to choose the right metrics and set business goals*. Feb. 2017. URL: https://www.geckoboard.com/blog/how-to-choose-the-right-metrics-and-set-business-goals/. (Hentet: 29.04.2020).

[10] Theodore H. Poister. *Measuring Performance in Public and Nonprofit Organizations*. eng. John Wiley Sons, 2008. ISBN: 978-047036517.

[11] Laura Tyson. *Should I have this KPI on my dashboard? The IPA rule*. Jan. 2016. URL: https://www.geckoboard.com/blog/should-i-have-this-kpi-on-my-dashboard-ipa-rule/. (Hentet: 29.04.2020).

[12] Christina M Frederick-Recascino, Hana Schuster-Smith, et al. “Competition and intrinsic motivation in physical activity: A comparison of two groups”. In: *Journal of sport behaviour* 26.3 (2003), pp. 240–254.

[13] Magnus Johannesson, Robert Östling, and Eva Ranehill. “The Effect of Competition on Physical Activity: A Randomized Trial”. In: *The B.E. Journal of Economic Analysis Policy* 10.1 (2010). URL: https://www.degruyter.com/view/journals/bejeap/10/1/article-bejeap.2010.10.1.2555.xml.xml.

Appendices

Following are included

Appendice A – Questionnaire
Appendice B – Multiplayer observation schema
Appendice C – Singleplayer observation schema
Appendice D – Interview guide
Appendice E – Consent form (Without reference number)

Appendice A – Questionnaire

Dato:_____

Spørreundersøkelse om lekne installasjoner

Kjønn: Mann ☐
Kvinne ☐
Annet ☐

Alder:_____

Kryss av for hvorvidt du er helt uenig, litt enig, verken eller, litt enig eller helt enig i følgende påstander.

		Helt uenig	Litt uenig	Verken eller	Litt enig	Helt enig
1.	Jeg syns varigheten på spillet var for kort	☐	☐	☐	☐	☐
2.	Jeg syns intensiteten på spillet var for høy	☐	☐	☐	☐	☐
3.	Jeg syns vanskelighetsgraden på spillet var passelig	☐	☐	☐	☐	☐
4.	Jeg føler at jeg fikk brukt kroppen i spillet	☐	☐	☐	☐	☐
5.	Jeg ville kategorisert dette spillet som lek	☐	☐	☐	☐	☐
6.	Lydeffektene gjorde det lettere å forstå spillet	☐	☐	☐	☐	☐
7.	Lysene i spillet gjorde det lettere å forstå spillet	☐	☐	☐	☐	☐
8.	Jeg syns at spillet var gøy	☐	☐	☐	☐	☐
9.	Jeg ville anbefalt dette spillet til venner	☐	☐	☐	☐	☐

Kryss av for hvorvidt følgende aspekter bidro til din fornøyelse i spillet?

		Ingenting	Litt	En del	Mye	Svært mye
1.	Samarbeidet (hvis aktuelt)	☐	☐	☐	☐	☐
2.	Konkurransen (mot deg selv eller andre)	☐	☐	☐	☐	☐
3.	Lyden	☐	☐	☐	☐	☐
4.	Lyset	☐	☐	☐	☐	☐
5.	Bevegelsen av kroppen	☐	☐	☐	☐	☐
6.	Intensiteten	☐	☐	☐	☐	☐
7.	Kravet om konsentrasjon	☐	☐	☐	☐	☐
8.	At det er en aktivitet utenom det vanlige	☐	☐	☐	☐	☐

Appendice B – Multiplayer observation schema

Dato: _____

Test nr:	
Player 1 ID:	
Player 2 ID:	

Player	**Attitude/mood** Player 1	Uninterested	☐	Insecure	☐	Little energy	☐	Curious	☐	Excited	☐	**Other:**
	Attitude/mood Player 2	Uninterested	☐	Insecure	☐	Little energy	☐	Curious	☐	Excited	☐	**Other:**
	Communication between players	Yes	☐	No	☐	**Topic:**						
	Laughter	Yes	☐	No	☐	**Why?**						
Movement	**Whole body movement**	Yes	☐	No	☐	**If No, how?**						
	Green buttons intentionally not pressed?	Yes	☐	No	☐							
	Game pole placement Player 1	OK	☐	Too high	☐	Too low	☐					
	Game pole placement Player 2	OK	☐	Too high	☐	Too low	☐					
	Jumping	Yes	☐	Sometimes	☐	No	☐					
	Squats	Yes	☐	Sometimes	☐	No	☐					
	Effort to reach difficult button combinations	Yes	☐	Sometimes	☐	No	☐					
System usability	**Game-related questions?**	Yes	☐	No	☐	**If Yes, what?**						
	Confusion?	Yes	☐	No	☐	**If Yes, what?**						
	Asks for explanation?	Yes	☐	No	☐	**If Yes, what?**						

Sluttkommentar: _____

Appendice C – Singleplayer observation schema

Dato: _____

Test nr:	
Player ID:	

Player	**Attitude/mood**	Uninterested ☐	Insecure ☐	Little energy ☐	Curious ☐	Excited ☐	**Other:**			
	Laughter	Yes ☐	No ☐	**Why?**						
Movement	**Whole body movement**	Yes ☐	No ☐	**If No, how?**						
	Green buttons intentionally not pressed?	Yes ☐	No ☐							
	Game pole placement	OK ☐	Too high ☐	Too low ☐						
	Jumping	Yes ☐	Sometimes ☐	No ☐						
	Squats	Yes ☐	Sometimes ☐	No ☐						
	Effort to reach difficult button combinations	Yes ☐	Sometimes ☐	No ☐						
System usability	**Game-related questions?**	Yes ☐	No ☐	**If Yes, what?**						
	Confusion?	Yes ☐	No ☐	**If Yes, what?**						
	Asks for explanation?	Yes ☐	No ☐	**If Yes, what?**						

Sluttkommentar: _____

Appendice D – Interview guide

INTERVJU
Åpning: Dette er testobjekt # og har prøvd game mode single/competitive

1. **Inntrykk av spillet.** Kan du fortelle litt om førsteinntrykket ditt av spillet?
   Hva syns du om spillet etter å ha spilt det?
   Var det noen ting du likte spesielt godt ved spillet? Hva? Hvilken påvirkning?
   Var det noen ting du ikke likte så godt ved spillet? Hva? Hvilken påvirkning?

2. **Forståelse.** Hvordan syns du det gikk å forstå hvordan spillet fungerer?
   Var det noen spesielle elementer som bidro til at du forstod spillet?
   Var det noe som forvirret deg? Hvis forvirring - effekt på innsats/motivasjon?

3. **Erfaring.** Har du prøvd noe som ligner på dette før? Hva var det?
   I hvilken grad anser du deg selv som en leken person?
   Pleier du å spille bordtennis/foosball/biljard/shuffleboard/etc hvis muligheten byr seg?

4. **Følelser/Playfulness.** Da du kom hit i sta, før du spilte, hvordan følte du deg da? Hvordan var humøret?
   Hvordan følte du deg når du spilte?
   Kan du si noe om humøret ditt nå i etterkant?
   Bidro spillet til noen endring i humør (eller motivasjon for det du skal etterpå)? Hva kommer det av, tror du?

5. **Fysisk aktivitet.** Hva synes du om mengden bevegelse spillet ga deg?
   Kunne/burde det vært høyere krav om fysisk bevegelse?

6. **Motivasjon.** Da du kom hit i sta, fikk du lyst til å spille da du så installasjonen?
   Hva fikk deg til å ønske å spille?
   Når du kom i gang, hva var motiverende i spillet?
   Hva skal til for at du ville brukt denne i hverdagen?
   Evt hva kan gjøre at du ikke har lyst til å bruke den?

   (Hvis competitive)
7. **Samspill.** Hvilken verdi har det for deg å spille sammen med andre?
   Fulgte du mye med på hva den andre spilleren gjorde? Hadde det noen påvirkning på din innsats i spillet?
   Kjenner du personen du spilte mot? Hvis ja, hvordan tror du det ville vært å spille mot en fremmed?
   Tror du spillet kunne fungert som en type icebreaker for å gjøre det lettere å bli kjent med fremmede?

   (Hvis single)
   Score. Hva tenker du om score i denne typen spill? Var det et viktig element for deg? Hvorfor/hvorfor ikke?
   Tenkte du over scoren underveis i spillet?
   Hadde scoren noen påvirkning på motivasjonen din i spillet?

8. **Sosiale omgivelser.** Hvilke steder tenker du at denne typen installasjon er egnet for? I hvilke situasjoner ville det være naturlig for deg å oppsøke denne typen installasjon.
   Hvordan tror du kø/ventetid ville påvirket din motivasjon for å bruke spillet?
   Hvordan ville et eventuelt publikum påvirket hvordan du føler deg i spillsituasjonen?

Gi objektet en liten oppsummering av hva du har fått inntrykk av i løpet av intervjuet.

Appendice E – Consent form (Without reference number)

Forespørsel om deltakelse i forskningsprosjektet

«DigiPlay: Fysisk aktivitet gjennom lekne interaktive installasjoner»

Bakgrunn og formål
Formålet med prosjektet er å få tilbakemelding fra brukere (i hovedsak studenter) på en konkret interaktiv installasjon som har til hensikt å oppmuntre til spontan fysisk aktivitet.

Prosjektet er del av et mastergrads-prosjekt ved Institutt for datateknologi og informatikk og Institutt for interaksjonsdesign, NTNU. I studien vil vi ikke evaluere helsegevinst, men kun samle inn tilbakemeldinger fra friske brukere om deres bruksopplevelse av installasjonen, i tillegg til data om ulike måter/strategier å interagere med i installasjonen på.

Du er forespurt om å delta fordi du er i målgruppen.

Hva innebærer deltakelse i studien?
Hvis du velger å delta i prosjektet betyr det at du er med på en brukertest av installasjonen som innebærer både en utprøving og et etterfølgende intervju og spørreskjema.

Dine svar fra spørreskjemaet blir registrert på papir og senere elektronisk.

Vi ønsker å kunne gjøre videoopptak av utprøvingen av installasjonen og lydopptak av intervjuet.

Det er frivillig å delta
Det er frivillig å delta i prosjektet. Hvis du velger å delta, kan du når som helst trekke samtykke tilbake uten å oppgi noen grunn. Alle opplysninger om deg vil da bli anonymisert. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg.

Hva skjer med informasjonen om deg?
Vi vil bare bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket.

Det vil kun være forskere tilknyttet prosjektet som har tilgang til dataene, og ikke noen utover dette, f.eks. din arbeidsgiver.

Navnet og kontaktopplysningene dine vil vi erstatte med en kode som lagres på egen navneliste adskilt fra øvrige data. Datamaterialet (video og lydopptak) vil bli lagret på en forskningsserver på et innelåst rom.

Deltakere i prosjektet vil ikke kunne gjenkjennes i publikasjoner. Her publiseres kun anonymiserte data.

Hva skjer med opplysningene dine når vi avslutter forskningsprosjektet?
Prosjektet skal etter planen avsluttes 01.09.2021. Ved prosjektslutt vil datamaterialet bli anonymisert slik at du ikke kan gjenkjennes. Dette gjøres for etterprøvbarhet og eventuell senere forskning.

Studien er meldt til Personvernombudet for forskning, Norsk samfunnsvitenskapelig datatjeneste AS.

Dine rettigheter

Så lenge du kan identifiseres i datamaterialet, har du rett til:

- innsyn i hvilke personopplysninger som er registrert om deg,
- å få rettet personopplysninger om deg,
- få slettet personopplysninger om deg,
- få utlevert en kopi av dine personopplysninger (dataportabilitet), og
- å sende klage til personvernombudet eller Datatilsynet om behandlingen av dine personopplysninger.

Hva gir oss rett til å behandle personopplysninger om deg?
Vi behandler opplysninger om deg basert på ditt samtykke.

På oppdrag fra NTNU har NSD – Norsk senter for forskningsdata AS vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket. (Ref. Meldeskjema [FYLLES INN REFERANSENUMMER]).

Hvor kan jeg finne ut mer?
Hvis du har spørsmål til studien, eller ønsker å benytte deg av dine rettigheter til å trekke ditt samtykke, ta kontakt med:

- NTNU ved førsteamanuensis Yngve Dahl ved Institutt for datateknologi og informatikk (yngveda@ntnu.no, mob.: 905 27 892)
- Vårt personvernombud: Thomas Helgesen (thomas.helgesen@ntnu.no)
- NSD – Norsk senter for forskningsdata AS, på epost (personverntjenester@nsd.no) eller telefon: 55 58 21 17.

Med vennlig hilsen

Yngve Dahl (Prosjektansvarlig)

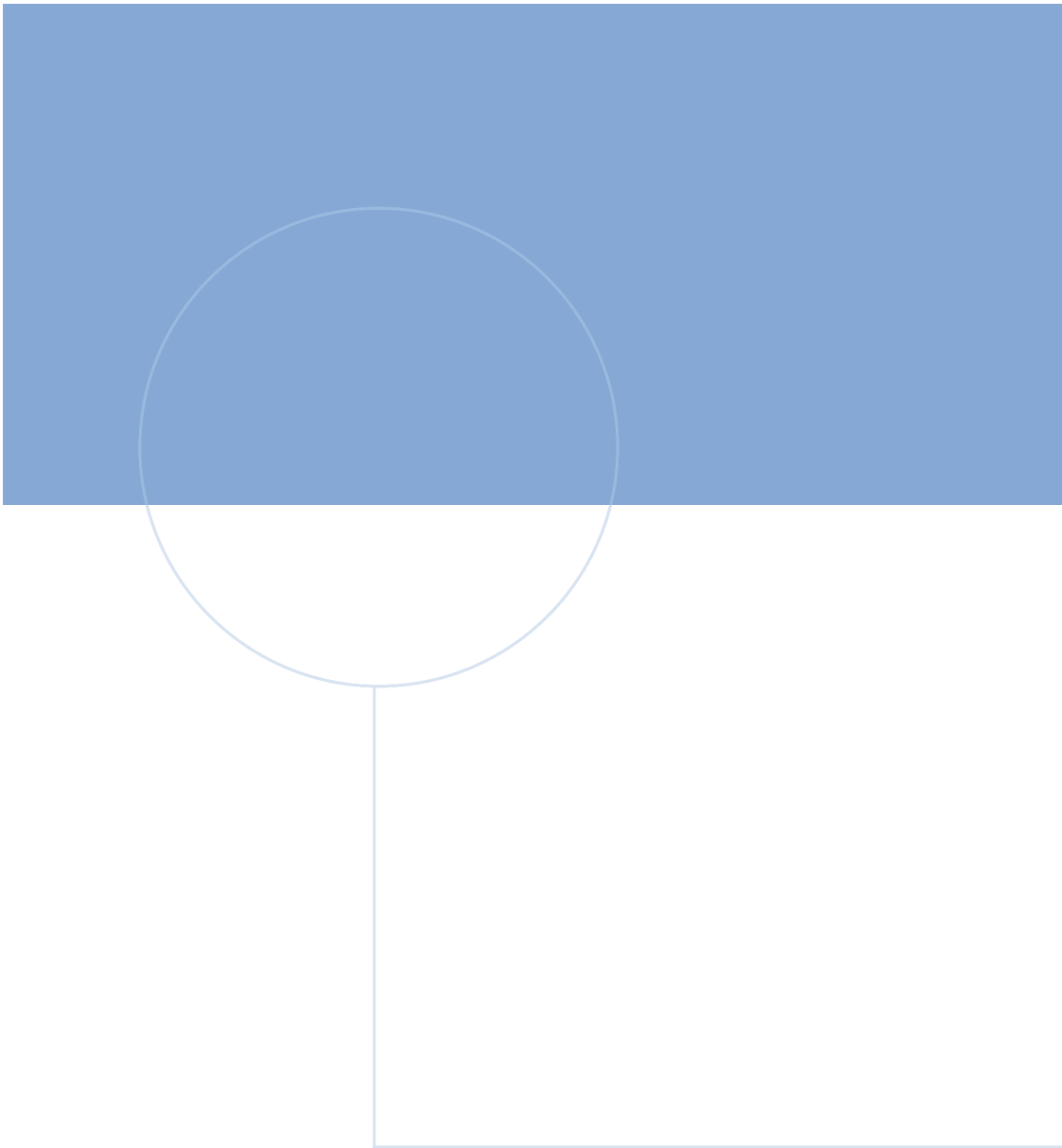
Samtykke til deltakelse i studien

Jeg har mottatt informasjon om studien, og er villig til å delta

---

(Signert av prosjektdeltaker, dato)

☐ ***Jeg samtykker til å delta i studiet.***
☐ ***Jeg samtykker til at personopplysninger kan publiseres/ lagres etter prosjektslutt.***



NTNU

Norwegian University of
Science and Technology