Mikko Ojares

DIFFERENCES IN HEURISTIC GAME EVALUATIONS BETWEEN DEVELOPERS AND PLAYERS

Master's Thesis in Technology
Faculty of Information Technology and Communication Sciences
Examiner: PhD Jari Varsaluoma
Examiner: PhD Kirsikka Kaipainen
February 2023

ABSTRACT

Mikko Ojares: Differences in heuristics game evaluations between developers and players
Master's Thesis in Technology
Tampere University
Information Technology
February 2023

This study aims to analyze the differences between game developers and players when evaluating mobile games using mobile game heuristics that focus on the user interface (UI) elements of the games. The focus is on the number and severity of problems found, as well as differences in behavior during evaluations. Evaluator effect is calculated using any-two agreement for game developers and players. Additionally, the study proposes a list of new mobile game heuristics for evaluating mobile games and analyzes their usability and effectiveness when evaluating games in development compared to fully developed games.

Eight participants, including four game developers and four players, participated in remote evaluation sessions where they played and evaluated two different mobile games.

The results show that players spent more time on evaluations and found more problems than game developers, although game developers identified more catastrophic problems. The study concludes that the differences observed may be attributed to the diverse backgrounds of the participants and especially, if the participants have experience with the type of game evaluated either from playing those kinds of games or developing them. Evaluator effects calculated showed that players had more agreements on the problems of the games compared to game developers. Additionally, the new game heuristics were considered usable in mobile game evaluations after some modifications. The heuristics were also more efficient with a game that is still in development.

Keywords: Usability evaluation, Skill level, Evaluator effect, Mobile games, Game evaluation, Game developers, Players, Heuristics, Game heuristics, User interface

TIIVISTELMÄ

Mikko Ojares: Differences in heuristics game evaluations between developers and players
Diplomityö
Tampereen yliopisto
Tietotekniikka
Helmikuu 2023

Tämän tutkimuksen tarkoituksena on analysoida pelinkehittäjien ja pelaajien välisiä eroja arvioitaessa mobiilipelejä käyttämällä mobiilipeli heuristiikkoja, jotka keskittyvät pelien käyttöliittymäelementteihin. Painopisteenä on havaittujen ongelmien määrä ja vakavuus sekä erilaiset käyttäytymiserot osallistujien välillä arvioinnin aikana. Arvioijan vaikutus (Evaluator effect) lasketaan pelin kehittäjille ja pelaajille. Lisäksi tutkimuksessa luodaan lista uusista mobiilipelien heuristiikoista, joita voitaisiin käyttää mobiilipelien arvioinnissa ja analysoidaan niiden käytettävyyttä ja tehokkuutta, mukaan lukien, kun niillä arvioidaan kehitteillä olevaa peliä verrattuna täysin kehitettyyn peliin.

Tutkimuksessa kahdeksan osallistujaa, mukaan lukien neljä pelinkehittäjää ja neljä pelaajaa, osallistuivat arviointi-istuntoihin, joissa he pelasivat ja arvioivat kahta erilaista mobiilipeliä.

Tulokset osoittavat, että pelaajat käyttivät enemmän aikaa arvioihin ja löysivät enemmän ongelmia kuin pelien kehittäjät, vaikka pelinkehittäjät havaitsivatkin suuremman vakavuuden ongelmia. Tutkimuksessa todetaan, että havaitut erot voivat johtua osallistujien erilaisista taustoista ja erityisesti siitä, onko osallistujilla kokemusta arvioitavasta pelityypistä joko pelaamisesta tai kehityksestä. Tulosten mukaan arvioijan vaikutus on suurempi pelaajilla verrattuna pelien kehittäjiin. Lisäksi uudet peliheuristiikat olivat käytettävät mobiilipelien arvioinneissa tiettyjen muutoksien jälkeen. Peliheuristiikat olivat myös tehokkaammat, kun niillä arvioidaan kehitteillä olevaa peliä.

Avainsanat: Usability evaluation, Skill level, Evaluator effect, Mobile games, Game evaluation, Game developers, Players, Heuristics, Game heuristics, User interface

PREFACE

I would like to thank my thesis supervisor Jari Varsaluoma for valuable feedback and guidance with my master’s thesis. I would also like to thank all the volunteers who participated in this study. I would also like to thank my family and friends for their support and especially my father who has been an important guide for me during my studies and all my life.

Tampere, 15.2.2023

Mikko Ojares

TABLE OF CONTENTS

1. INTRODUCTION

In usability evaluations, the skill level of the evaluator when it comes to the topic that is being evaluated affects the results and people with different skill levels, can have very different answers. Current studies haven't examined these differences when it comes to game evaluation. In this study I aim to analyse the differences and the reasons behind them between different skill level evaluators when it comes to evaluating mobile games. In this study the skill level is differentiated by having people who play games as participants and people who develop games. In current studies people who have the same skill level also have very different results when doing usability evaluations. This phenomenon was defined as the evaluator effect. [1] In this study I also analyse this evaluator effect with players and game developers. In this study I create new mobile game heuristics that focus on user interface usability issues, and I analyse how usable these heuristics are when evaluating mobile games and how effective they are with different kinds of games. The study is done by having participants evaluate two different mobile games with the new heuristics and analysing the answers along with conducting interviews with the participants. For this study the following research questions are made:

The first research question focuses on all the differences that may occur between game developers and players such as the number of problems found in the games and their severity, or the time spent doing the evaluations.

> **R**esearch **Q**uestion **1. What are the differences in results when using heuristics to evaluate mobile games between game developers and players?**

The second research question focuses on all the differences inside the group of game developers and the group of players. These differences are measured by calculating the evaluator effect.

> **R**esearch **Q**uestion **2. How large is the evaluator effect with game developers and players when evaluation a mobile game?**

The third research question focuses on the heuristics that are created in this study and how effective and usable they are considered by game developers and players when using them in evaluating mobile games. The usability of the heuristics is analysed from the subjective opinions and comments of the participants about the heuristics.

Research **Q**uestion **3. How usable are the new game heuristics when evaluating a mobile game?**

The fourth research question focuses on how well the heuristics work when they are used in evaluating mobile games that are in different phase of development. This is analysed from the results of the evaluations and the subjective comments of the participants from the interviews.

Research **Q**uestion **4. How effectively do the heuristics work when used in evaluating a game that is in development compared to a fully developed game?**

This thesis is structured in the following way. In chapter 2 I review the existing literature when it comes to general usability heuristics and game heuristics. I also examine the examples of differences in skill levels in existing literature and I introduce the evaluator effect in more detail. In chapter 3 I introduce the new mobile game heuristics that are created for this study. In chapter 4 I explain the research method of this study in more detail. This includes describing the mobile games used in the evaluations and the criteria for choosing them. I also describe the participants of the study and the detailed research process of the study. Finally in chapter 4 I describe the analyse methods used in the study in detail. In chapter 5 I present the results of the study. In chapter 6 I aim to analyze the different results and the reasons behind them. I will also answer to the research questions introduced above. In chapter 7 I will present the conclusions from this study.

2. RELATED WORK

In this section I will briefly go through the literature and discuss the existing studies that are relevant for this study. I will first go through more general heuristics found in literature. Second, I will discuss heuristics specifically made for games. Finally, I will discuss examples of differences in skill levels and the evaluator effect.

2.1 General heuristics

Jokela stated that when it comes to customer satisfaction there are must-have elements and then other factors that might not be that important. These must-have elements need to be met to satisfy the user. Also, if a product has usability issues, the user might still be satisfied because their must-have factors were met. [2]

Before discussing heuristics for mobile games, it is essential to discuss general usability heuristics and heuristics that are made for mobile applications but not necessarily to games. By examining these more general heuristics it could be possible to identify the must-have elements when it comes to mobile interfaces before even considering games.

Heuristics can be defined as broad usability principles [3,4] or as broad rules of thumb that are not specific usability guidelines [5].

In 1990 one of the earliest heuristics were made by Nielsen & Molich when they investigated the abilities of data processing professionals in recognizing interface problems. Based on the study’s findings Nielsen & Molich created a checklist of usability considerations. [6] Later in 1994 Nielsen refined these heuristics and created one of the most known heuristics for evaluating user interfaces [3]. These heuristics can be seen as a good base for general heuristics when it comes to evaluating user interfaces. However, it is important to consider heuristics that focus purely on mobile interfaces. Based on literature there have been several different heuristics made for mobile user interfaces but with different goals.

In 2006 Gu Ji et al. developed a usability checklist for mobile phone user interfaces in a study where the authors’ goal was to create a “must-have usability” checklist for mobile phones. This list of heuristics could then be used by mobile user interface developers to quickly evaluate the usability of mobile interfaces. [7] In 2014 Al-Razgan et al. introduced heuristics for elderly people that interact with mobile phones [8]. Even though this study is not focused on elderly people, this list made by Al-Razgan et al. contained some heuristics I considered relevant anyways.

Some of the newer heuristics were made by Johnston & Pickrell in 2016. In their study they created heuristics for designers creating mobile applications [9]. In 2019 Kumar & Goundar extended the general heuristics made by Nielsen [3] to create heuristics that support mobile learning applications [10]. Also, in 2019 Costa & Canedo created heuristics that focus on mobile applications on smartphones [11].

2.2 Game heuristics

Heuristics that focus on generic user interfaces are not detailed enough to be used in games, so heuristics related to games (henceforth game heuristics) are needed. In this section I will go through the different game heuristics found in literature that the author of this study considered relevant.

One of the earliest game heuristics were made by Malone in 1982. Malone created his heuristics by creating 8 different versions of the same instructional game by removing features that he considered motivational to the player one at a time. Then Malone conducted a study where 80 fifth grade students were given the choice to play one of the 8 versions of the game assigned randomly or a completely different game that was the same for all participants. Malone measured how long each student played their version of the game with 8 different version compared to the game which was same for all. The heuristics made by Malone were divided into three main categories: Challenge, fantasy, curiosity. Malone also mentions in his study that the main purpose of the framework he created is to work as a general checklist of heuristics that can be used for designing enjoyable user interfaces. [12]

In 1998 Clanton also created a set of game design principles. Clanton divided these guidelines into three different classes: game interface, game mechanics, and game play. He also demonstrated examples of different genre games. When comparing these guidelines to the heuristics Malone made. Clanton's guidelines focused more on how to engage the user. [13]

One of the most referred game heuristics were made by Melissa Federoff in 2002. Federoff created her heuristics by first reviewing literature and then spending time with a game development company and observing different people. Federoff divided her heuristics to the same classes used by Clanton: game interface, game mechanics and game play. Federoff also mentions that the study she made mainly focused on games that have the main goal of entertaining the user unlike Malone who focused on instructional games. [14]

In 2004 Desurvire et al introduced the heuristic evaluation of playability (HEP) which was a list of heuristics that were based on literature of productivity and playtesting heuristics. The heuristics made by Desurvire et al. were divided in to four different categories: game play, game story, game mechanics, and game usability. After the heuristics were made Desurvire et al. conducted user studies with four users that took part in two-hour playability sessions. Results of the study showed that HEP is especially helpful in early game design. Results also showed that HEP could find issues even before actual user interaction is possible. [15]

One of the most referred and popular heuristics made for mobile games were made by Koivisto & Korhonen in 2006. The heuristics made by Koivisto & Korhonen create a core model that has three different modules: game usability, mobility, and gameplay. According to Koivisto & Korhonen this core model can be used in any mobile game evaluation. The mobility modules heuristics describe issues unique to the mobile context which weren't considered in earlier heuristics. When developing the mobility heuristics Koivisto & Korhonen first analysed mobile phones and their context of use. The authors indicated that the mobile context is very different. Mobile phones can be used outdoors where lighting and noise can change. Users need to be more aware of their surroundings and other people need to be taken into consideration. Mobile phones are often also used in killing time during short breaks. The authors also mention that mobile phones' main purpose is different forms of communication which can cause interruptions for example when getting a call. Small screen sizes, battery limitations, insufficient audio capabilities and limited processing power also need to be considered. Based on these findings Koivisto & Korhonen define Mobility as "how easily the game allows a player to enter to the game world and how it behaves in diverse and unexpected environments". [16]

In 2007 Koivisto & Korhonen made another study that complemented the heuristics made by their earlier study discussed above. In this study Koivisto & Korhonen created heuristics for mobile multi-player games. The results of the study showed that these heuristics that focus on multi-player games can also be applied to non-mobile games. [17]

Schaffer created his own set of heuristics also in 2007. Unlike others who had created game heuristics such as Malone, Clanton or Federoff, Schaffer used examples of every heuristic such as screenshots and explanations. Schaffer indicates that: "The heuristics are concrete and specific, so it should be clear to game designers how to implement the heuristics".[18]

In 2008 Pinelle et al. created the first heuristics that exclusively focused on game usability, and which are also based on structured analysis of problems in large number of games thus covering several genres. Pinelle et al. identified game usability problems from game reviews and categorized them. After creating the heuristics, the authors conducted evaluations where five different evaluators used them to evaluate a game. [19]

In 2009 Pinelle et al. created heuristics called Networked Game Heuristics (NGH) which focused on networked multiplayer games. These heuristics were created by identifying problems from game reviews and categorizing them. These heuristics were tested by having evaluators use these heuristics and an existing heuristic list to evaluate two games. The authors indicate that their research was “the first to present networked game heuristics that are derived from real problem reports, and the first to evaluate the heuristics’ effectiveness in a realistic usability test”. [20]

In 2009 Desurvire & Wiberg created new heuristics of playability (PLAY). These heuristics were created in a follow-up study of HEP. These heuristics were created to help developers during the entire design process and especially at the starting phases. PLAY- heuristics were developed for only three genres: Real-Time Strategy, Action adventure and first-person shooter games. The PLAY- heuristics were divided into seven different categories: game play, skill development, tutorial, strategy & challenge, game/story immersion, coolness, usability/game mechanics and controller/keyboard. [21]

Also, in 2009 Papaloukas et al. introduced heuristics that can be used on usability studies that focus on new genre video games which Papaloukas et al. defined as “videogames that use specific and unique equipment or are part of a general software category such as platforms of social networking”. [22]

In 2010 Zaibon & Shiratuddin created heuristics that focus on mobile game-based learning. These heuristics were created by taking the heuristics made by Korhonen and Koivisto and adding the new component “learning content” to these heuristics.[23]

Year 2010 had also multiple other heuristics made. Koeffel et al. created heuristics that can be used in evaluating not only video games but also advanced interaction games (tabletop games)[24]. Omar & Jaafar created heuristics for educational games[25]. Paavilainen created his own heuristics for social games [26]. Tan et al created heuristics for instructional games[27]

In 2012 Sweetser et al. created heuristics focusing on real-time strategy games[28]. This list also had some relevant heuristics for this study even though the study made by Sweetser et al. focused on certain type of games.

In addition to all the examples mentioned above, over the years there have been multiple heuristics created for a large number of different other domains other than games [29].

2.3 Differences in heuristic evaluation skills

When evaluating games in general or when using these heuristics listed above, the skill level of the person using them can make a significant difference. This simple fact has been addressed in literature many times.

In a study by Nielsen & Molich made in 1990, they conducted experiments on heuristic evaluation. Nielsen & Molich believed that to improve usability in most industrial situations you need to study usability methods. In the study by Nielsen & Molich practical applicability of heuristic evaluation was tested. The study was conducted with four different experiments where people analysed a user interface heuristically. In the study these people were not usability experts. At very early stages of the study there was a discovery relating to different skill levels. Nielsen & Molich mention in their study that they had to modify their initial list of usability problems after an initial pass through the reports. This was because their evaluators in each experiment discovered problems that they had not discovered themselves. Nielsen & Molich make an interesting comment about this in their study. Nielsen & Molich mention: “This show that even usability experts are not perfect in doing heuristic evaluations”. [30] In experiment 1 evaluators tested a user interface of a video-tex system. The evaluators in this study were computer science students. In experiment 2 Nielsen & Molich used a design specifically made for the test. The design was a small information system which a telephone company would make available to their customers. The evaluators in this experiment were readers of the Danish Computerworld magazine. In experiments 3 and 4 Nielsen & Molich used “live” systems instead of specification-only designs that they used in experiments 1 and 2. According to the study both systems were “voice response” systems. Both experiment 3 and 4 were done with the same group. This group consisted of computer science experiments. Nielsen & Molich also highlight that this group of evaluators had no overlap with the group used in experiment 1. [30]

Nielsen & molich point out that based on the results from the four experiments, heuristic evaluation is difficult. Based on the results, in the experiment with the best results, the average proportion of usability problems found was 51% so only half of the problems were found. According to the study's results some evaluators did better than others. Nielsen & Molich point out that based on the results of their study even poor evaluators can find hard problems and good evaluators may overlook easy problems. This shows that the skill level of the evaluator can make a significant difference when making heuristic evaluations. The results of the study show that if you have several people conduct the evaluations and they are done independently of each other, the results of a heuristic evaluation will be better. Nielsen & Molich also recommend that heuristic evaluation should be done with between three and five evaluators.[30]

On another study made in 1992 Nielsen investigated the effects when the evaluators had varying levels and kinds of expertise. This study was carried out in the way that a same interface was used in heuristic evaluations and these heuristic evaluations were done by three different groups of evaluators. The system that was evaluated was a voice response system accessed through a touch tone telephone. In the study Nielsen mentions that because of the variety of evaluators employed in the study, printed dialogue was evaluated instead of an actual running system. From this point on in the study Nielsen referred to the system as the “BankingSystem”. In the study that Nielsen conducted, the three groups that performed the heuristic evaluations on the BankingSystem were different from each other based on the skill level of the evaluators. The first group consisted of computer science students with no formal knowledge of user interface design principles. Nielsen referred to this group as the novice evaluators. Nielsen also emphasized that these people were only novices regarding usability but not on the general use of computers. The second group in Nielsen's study consisted of “regular” usability specialists. Nielsen refers to them as “people with experience in user interface design and evaluation but no special expertise in voice response systems”. In his study Nielsen also mentions that the definition of a usability specialist on that study can be considered as a person with graduate degrees and/or several years of job experience in the usability area. The third group in Nielsen's study consisted of “double specialists”. Nielsen describes these people as “had expertise in user interface issues as well as voice response systems and therefore expected to indicate the best level of heuristic evaluation performance one might hope for”. The results of the study show that usability specialists are better at finding usability problems than people without usability training and having expertise with the type of user interface that is being evaluated helps. Based on the results of the study Nielsen conducted none of the groups did exceptionally well but “Double specialists” were able to find over half of the problems. Nielsen emphasizes that based on the studies result “double specialists” found more problems because they had specific experience for the type of user interface evaluated instead of just being better usability specialists. [4]

According to literature at least to the authors knowledge there are quite few actual studies made that mainly address different skill levels. However, when going through literature it was evident that there are multiple examples where the issue of different skill levels was discussed.

In 2010 Folstad et al. compared the performances of work-domain experts and usability experts. In this study Folstad et al. indicated that work-domain experts were characterized by high computer experience and low system experience. They conducted their study by doing a group-based expert walkthrough which was a method developed for supporting non-usability experts as evaluators. The study was done by 15 work-domain experts and 12 usability experts. The results of the study showed that work-domain experts generated equally valid usability inspection results when compared to the ones of usability experts, but they were less thorough. Folstad et al. indicated that this result showed that work-domain experts may be used as evaluators in usability inspections without compromising validity. [31] This again shows that people with “less” experience in making evaluations can be used in usability evaluations even though when compared to “experts” they behave differently.

One very interesting study related to skill levels of evaluators was done by Salian et al. in 2013 where they analysed the effectiveness of heuristic evaluation when the evaluators were children. The goal of their study was to find out whether children can perform heuristic evaluations. In their study 14 children evaluated a music making game on laptop. The results of the study showed that children could find problems, but they had difficulties understanding severity ratings and heuristics. Through observations the authors also noticed that the children didn’t understand the general purpose of heuristics and need more explanation. Some of the children also had troubles writing the heuristic number onto the evaluation comment sheet they were given. Based on the study’s results only 10 out of 27 problems identified had severity ratings attached to them so the authors indicated there must have been some problem with this part of the evaluation process. The children also only used 5 heuristics from a list of 12 and not a single child was able to relate their problems to other remaining heuristics. [32] Children can be considered in this context as novices when it comes to doing heuristic evaluations. This study shows that when doing heuristic evaluations with low level evaluators, it has great effects on the results but in addition to that, there can be problems with the actual evaluation process. The study made by Salian et al. gives direct questions that need to be researched in this study with game designers and players. Of course, children are its own special group and players, or game designers can’t be compared to them, but it will be interesting to study do players or even game designers with no earlier experience with heuristic evaluations have problems with the actual process like the children had.

There are also some smaller examples of the need to investigate effects of different skill levels in game related studies. In a study made by Fitchat & Jordaan in 2016 participants who were evaluating games mentioned that players differ in gender, age, playing style and skill level and these differences greatly affect the gaming experience. [33] In another study made by Phan et al. in 2016 the authors indicate after developing the Game User Experience Satisfaction Scale (GUESS) that it is currently administered on to players at least 18 years old with some high school education. According to the authors, future researchers could be interested in evaluating GUESS with younger and less educated people. [34]

There have also been other examples of different skill levels in literature. In 2019 Santos et al. conducted a study to investigate discrepancies between game reviews made by video game press and casual gamers. In this study Santos et al. referred to video game press as “experts” and to casual gamers as “amateurs”. The results of the study showed that reviews made by amateurs are highly polarized whereas expert reviews are more balanced. Results showed that amateur reviews often use emotionally charged vocabulary in the reviews. Santos et al indicated this fact shows that amateurs exhibit stronger sentiment compared to experts. [35] The results from the study made by Santos et al. show that even in other cases other than heuristic evaluation, the skill level of the person has a great effect on the results and overall opinions and behavior.

In 2019 Thewes et al. examined the usefulness of a set of heuristics in the evaluations of sociotechnical systems. In this study two groups which both consisted of fourth year students from a master- course were compared to each other. Both groups first evaluated a sociotechnical system and then after that a similar evaluation was performed but now with the use of a heuristic set. This was done to learn the base capability of participants’ base capabilities of evaluating sociotechnical systems and to investigate the improvements resulting from the use of heuristics. The results of the study showed that increase in productivity after using heuristics was not confirmed against the authors’ assumptions. Based on the results the number of observations that were documented decreased after the use of heuristics. The authors mention that further research needs to be done to investigate the effects and possible causes for this. [36] In the discussion part of the research made by Thewes et al. the authors mention several good points relevant for this study. Thewes et al. indicate that a minimal set of personal information was surveyed from the participants in respect to privacy concerns and because of this, the performances of individual participants cannot be matched to their respective prior knowledge. Individual work experiences and knowledge in evaluating work systems was not surveyed and the authors mention that these should have been controlled to help analyse the results. The authors indicate that additional research is needed to learn about the effects of novice evaluators and evaluators who are experts in related subjects. The authors also mention that prior to evaluating sociotechnical systems, in addition to thorough assessment of participants knowledge and experience, determining the domain of expertise is crucial. [36]

In 2021 Li et al. proposed an approach to analyse the playability of video games. This was done by mining a large amount of players' opinions from reviews they had made. After Li et al. had proposed their playability evaluation method, they conducted a study to verify the effectiveness of the method. This was done by conducting an experiment on a video game. After conducting this experiment Li et al. verified the correctness of the detected merits gathered from analysing player made reviews. This was done by comparing the results they had to the expert opinions about the same game extracted from critic reviews. [37] I will be focusing on the results of Li et al. that refer to the different skill levels. Based on the comparison made by Li et al. both parties pointed out similar topics in their reviews such as crashing or performance issues. Both parties also pointed out same positive and negative aspects related to gameplay but there were also some differences. Li et al. mention that critic reviews mentioned the sense of relaxing while this was not mentioned by players. [37]

All these examples show that there is need to investigate the effects of different skill levels in different fields of research and especially in the interest of this study, studies about game designers, players, and the skill level differences between them have not been studied at least to the authors knowledge.

2.4 Evaluator effect

The examples mentioned above show that in addition to people with totally different skill levels, people with the same skill level have also very different results when using the same system or when doing usability evaluation. This phenomenon was called “the evaluator effect” by Hertzum & Jacobsen. In 2003 Hertzum & Jacobsen conducted a study about three usability evaluation methods (UEMs) which were cognitive walkthrough (CW), heuristic evaluation (HE) and thinking-aloud (TA). The authors studied that, does evaluators who evaluate the same system with the same usability evaluation method find similar problems? In their study Hertzum & Jacobsen used the results of previous studies where this evaluator effect was evident, but it wasn’t addressed properly. In their study the authors used *any-two agreement* to measure the evaluator effect. This method measures the extend pairs of evaluators agree on problems. The *any-two agreement* was defined as “the number of problems two evaluators have in common divided by the number of problems the collectively detect, averaged over all possible pairs of two evaluators”. [1]

$$Any - two\ agreement$$
$$= Average\ of\ \left|\frac{Pi\ \cap Pj}{Pi\ \cup Pj}\right|\ over\ all\ {}^{1}/_{2}\, n(n-1)\ pairs\ of\ evaluators$$

Formula for calculating the any-two agreement [1]

In this equation of *any-two agreement Pi* and *Pj* are the sets of problems detected by evaluator *i* and *j* and *n* is the number of evaluators.

The results of the study made by Hertzum & Jacobsen showed that the evaluator effect is not restricted to novice evaluators or evaluators knowledgeable of usability, but it is relevant with all levels of skill. The results also showed that evaluator effect was found for evaluators with experience in the specific usability evaluation method they had been using. The results showed that the evaluator effect was also present in multiple kinds of systems and problems. Hertzum & Jacobsen mention that “The evaluator effect has been documented for different UEMs, for both simple and complex system, for both paper prototypes and running systems, for both novice and experienced evaluators, for both cosmetic and severe problems, and for both problem detection and severity judgment”. These results show that the evaluator effect is a real issue, and it needs to be considered.

[1] The authors make and interesting notice about the evaluator effect in their study. They mention that “The question is not whether the evaluation effect exist but why it exists and how it can be handled”. Hertzum & Jacobsen indicate that they believe the main cause for evaluator effect to be the fact that usability evaluation is a cognitive activity where the evaluator exercises judgment. The authors mention three contributors in usability evaluation methods that contribute to the evaluator effect: vague goal analysis, vague evaluation procedures, and vague problem criteria. [1]

There have also been other studies regarding the evaluator effect. Vermeeren et al. examined how characteristics of data analysis process may influence the evaluator effect. Three studies were conducted, and, in each study, two evaluators independently analysed same video recorded user test data. The results of the studies made by Vermeeren et al. again showed that evaluator effect can be found in different kinds of systems. Vermeeren et al. indicate that causes of the evaluator effect lay in the different kinds of interpretations of verbal comments and non-verbal behavior, guessing user intentions, decisions on how problematic interaction inefficiencies are, and differentiating usability problems from problems of the test set-up itself. The authors suggest that in order to manage the evaluator effect, researchers should conduct detailed data analysis with automated data logging, evaluators should discuss problems they are unsure of with other evaluators, and the analysis should be done by multiple evaluators. [38]

According to these examples mentioned above, the evaluator effect is a phenomenon that most definitely exists but the extend of it might be unclear. Also, when it comes to evaluating games with game heuristics, there hasn’t really been studies related to the evaluator effect at least to the best of the authors knowledge thus studying the evaluator effect and discussing the possible reasons for it in game evaluations was considered important by the author and considered in the second research question.

2.5 Conclusion of the relevant heuristics

In this section I will briefly conclude all the most relevant heuristics found from the literature that were used in creating the new list of mobile game heuristics and why these heuristics were chosen. The most relevant heuristics for this study and their sources are shown in table 1.

Table 1: Relevant heuristics for this study and their sources

Heuristics	Source
Malone	[12]
Federoff	[14]
HEP	[15]
Koivisto & Korhonen	[16]
Koivisto & Korhonen	[17]
Schaffer	[18]
Pinelle et al.	[19]
Pinelle et al.	[20]
PLAY	[21]
Papaloukas et al.	[22]
Omar & Jaafar	[25]
Tan et al.	[27]
Sweetser et al.	[28]
Gu Ji et al.	[7]
Al-Razgan	[8]

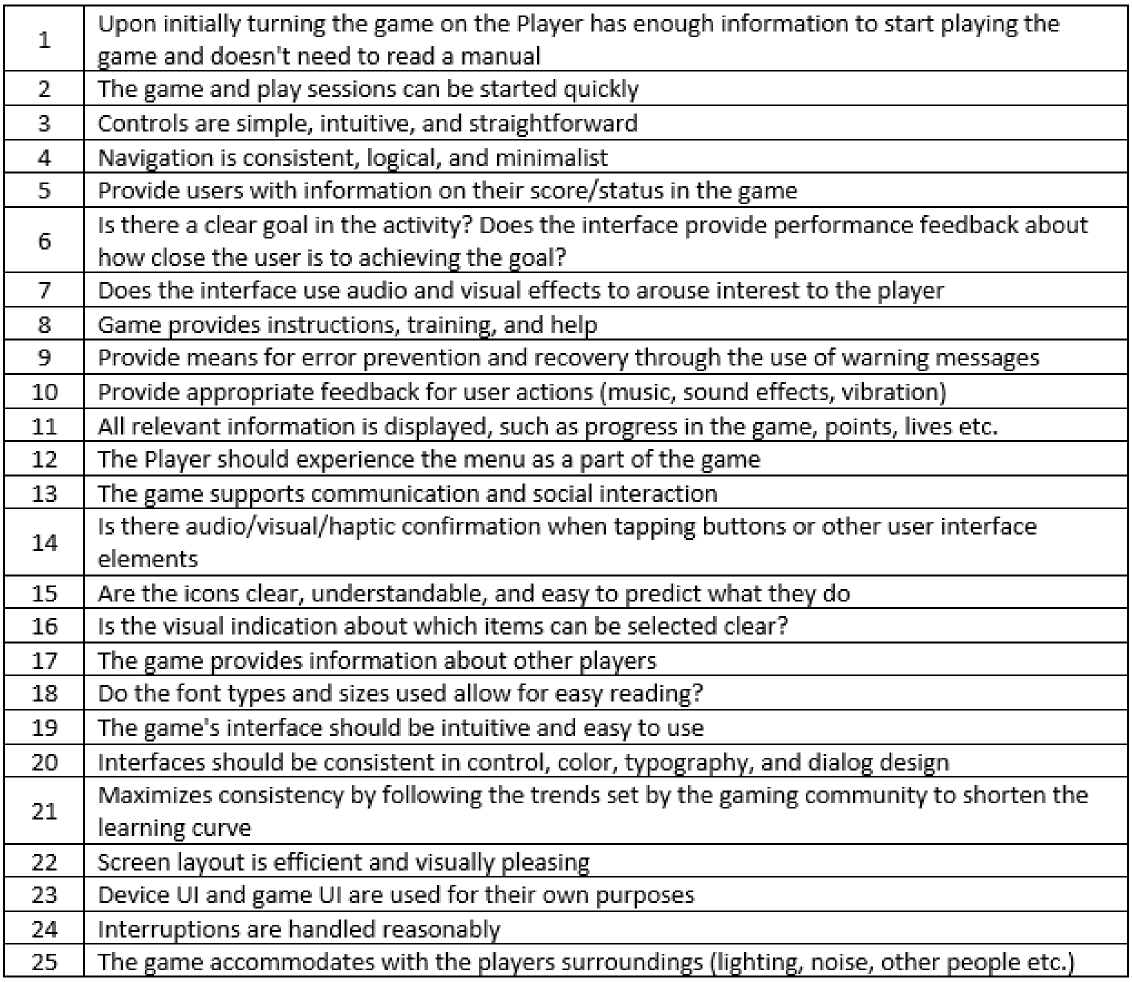
All of the heuristics except Gu Ji et al. [7] and Al-Razgan [8] had several good heuristics related to game user interface or general game heuristics that could be connected to the user interface and its usability, so these heuristics were chosen to be relevant. Koivisto & Korhonen [16] also had heuristics specifically related to mobility of the game and Koivisto & Korhonen [17] had heuristics related to multiplayer elements which are very common in mobile games, so I considered these heuristics to be very relevant. Gu Ji et al. [7] and Al-Razgan [8] had general heuristics related to mobile user interface usability which I considered to be also relevant when it comes to mobile games, so these heuristics were relevant.

3. FINALIZED HEURISTICS

To study these differences in skill levels and the evaluator effect, heuristics needed to be made for the test evaluations. Based on the heuristics presented in literature I created new heuristics to be used in this study. I also used the heuristics I created in my bachelor’s thesis as a base for these new heuristics [39]. I felt that these heuristics created in my bachelor’s thesis where not good enough for this study because I created them with the focus on general video games and only a few heuristics focused on mobile games. These old heuristics were also created based on the most referred heuristic lists in literature [12,14–17,21]. For this study I felt like these literature sources were not enough and the new heuristics should have references to other literature studies as well. I created these new heuristics by looking at heuristics lists from literature and choosing heuristics I felt were relevant in my own opinion when focusing on mobile games and their user interfaces. The heuristics were chosen in a way that they only focus on the user interface elements of the games, and they don’t focus on playability, or any aspects related to content of the game. Based on the literature many heuristics had similar items or even duplicates so I combined all the similar heuristics to only one and chose the wording of the heuristics so it would be as simple and coherent as possible. I also chose the order of the heuristics in the list so it would be a little more logical for the evaluators to start using them when playing a game for example heuristic referring to starting the game is first.

The new heuristics can be seen in table 2.

Table 2: New heuristics



1	Upon initially turning the game on the Player has enough information to start playing the game and doesn't need to read a manual
2	The game and play sessions can be started quickly
3	Controls are simple, intuitive, and straightforward
4	Navigation is consistent, logical, and minimalist
5	Provide users with information on their score/status in the game
6	Is there a clear goal in the activity? Does the interface provide performance feedback about how close the user is to achieving the goal?
7	Does the interface use audio and visual effects to arouse interest to the player
8	Game provides instructions, training, and help
9	Provide means for error prevention and recovery through the use of warning messages
10	Provide appropriate feedback for user actions (music, sound effects, vibration)
11	All relevant information is displayed, such as progress in the game, points, lives etc.
12	The Player should experience the menu as a part of the game
13	The game supports communication and social interaction
14	Is there audio/visual/haptic confirmation when tapping buttons or other user interface elements
15	Are the icons clear, understandable, and easy to predict what they do
16	Is the visual indication about which items can be selected clear?
17	The game provides information about other players
18	Do the font types and sizes used allow for easy reading?
19	The game's interface should be intuitive and easy to use
20	Interfaces should be consistent in control, color, typography, and dialog design
21	Maximizes consistency by following the trends set by the gaming community to shorten the learning curve
22	Screen layout is efficient and visually pleasing
23	Device UI and game UI are used for their own purposes
24	Interruptions are handled reasonably
25	The game accommodates with the players surroundings (lighting, noise, other people etc.)

The list of heuristics where you can see literature references from where that heuristic was adapted from can be seen in the appendix A.

Also worth mentioning is heuristic number 25: The game accommodates with the players surroundings (lighting, noise, other people etc.). This heuristic was referred in literature as: “The game accommodates with the players surroundings”, so it didn’t have any explanation to what “surroundings” means. I felt like this could be confusing for the evaluator, so I added my own brief explanation.

4. RESEARCH METHOD

In this section I will go through the games that are used in the study. I will describe the research group that were part of this study and the methods used in the study. I will also describe all the phases of the study in detail and all the data that is examined and how this data is analysed.

4.1 Games used in the research

In this subsection I will go through the criteria for selecting the games used in the evaluations, the process of selecting these games and briefly describe these games and their goals.

4.1.1 Criteria

When choosing the games that will be used in the heuristic evaluations, there were several criteria for these two games. These rules were:

1. *Both games are mobile games.*
2. *Both games need to available in both google play store and app store.* This criterion was chosen to make sure that participants who have different kinds of phones (Android/IOs) could participate in the study.
3. *Both games should be reasonably new*. This criterion was chosen because of the recommendation by Macey [40] so that the games would represent the most recent developments in their own genres.
4. *Both games should be released as close as possible to each other*. This criterion was chosen based on the recommendation made also by Macey [40] which was the chosen games should be published withing a certain timeframe from another, so they were comparable.
5. *One of the games needs to be in development phase and the other should be a finished game*. This criterion was chosen to answer to **RQ 4**.
6. *The genres of these two games can be different to each other*. The author of this study decided that the genres of the games are not that relevant for this study and that the heuristics could be considered usable with games from different genres.

7. *There must be multiplayer elements in one of the games and the other game cannot have any multiplayer elements*. The selection of this criterion was more unique compared to the others. This was chosen to add a small “trap” for the evaluators in the study. Heuristics 13 and 17 focus on elements that are only visible in multiplayer games. So, the other game should be totally single player with no multiplayer elements to see what the participants answer or think about these two heuristics.
8. *Both games should be as bug free as possible*. This criterion is kind of vague, but this was added so the games wouldn’t have any major issues or bugs that would affect the evaluations.

4.1.2 Selection process

The process of selecting these games started with discussing with the author’s friends if they know any good candidate games for this study with these criteria. These friends had quite a lot of experience from mobile games, so the author got some possible games as a list of candidates. After this initial list of possible games, the next step was browsing the Google Play Store and searching games from different categories. First, I browsed all the popular multiplayer games to find a game with multiplayer elements according to rule 7. This meant browsing the top 100 games in Google Play Store under the category multiplayer games. These games had several games that had millions of downloads and were really popular, so I considered these games also as finished games according to rule 5. From this list of multiplayer games, I was able to select a game that fits the criteria. For the other game I had to find a game that was more in a development phase and did not have any multiplayer elements. Going through the list of games my friend had made I was able to find a game fit for this criterion. After selecting these two games I made sure they were also available for app store. For both games I created game specific tasks that are done in the evaluation sessions. The idea of these tasks is to make sure the evaluator plays and explores enough of the game to have a proper idea of its features.

4.1.3 8 Ball Pool

The first game that was selected was 8 Ball Pool by Miniclip. This game was considered as a finished game since the game was released in 2013. [41] Overall pool games as an idea are quite easy to understand so I felt like this can be considered as a finished game. Even though the game was released in 2013 which is almost ten years ago, the idea of pool is timeless in the authors opinion so the game was acceptable for criteria 3 [41]. This game had several multiplayer elements in it. The idea of the game is that it is a pool game with traditional pool rules [42]. When you first start the game, the player is presented with a following screen seen in image 1.

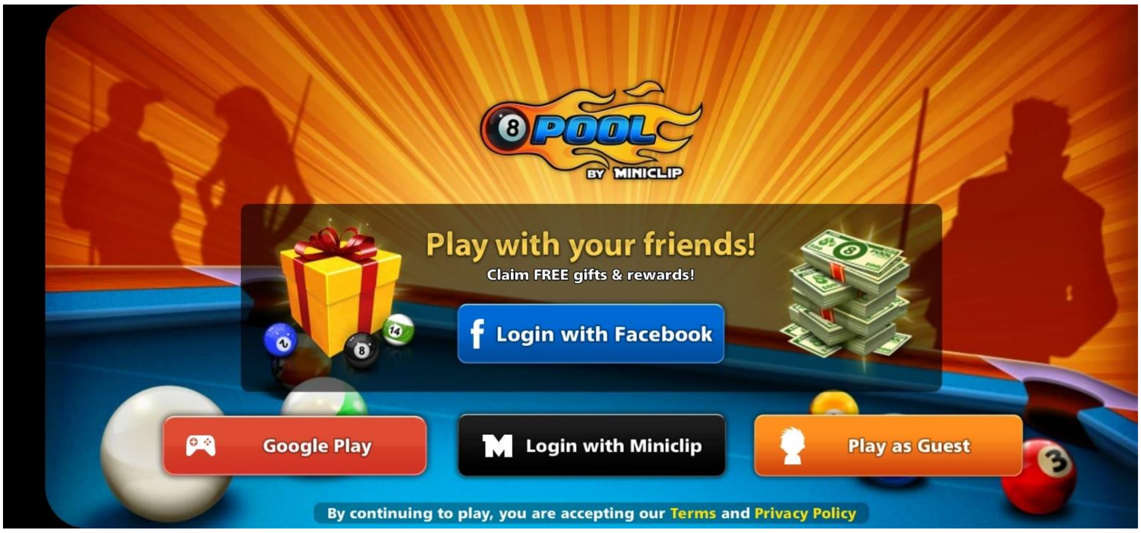
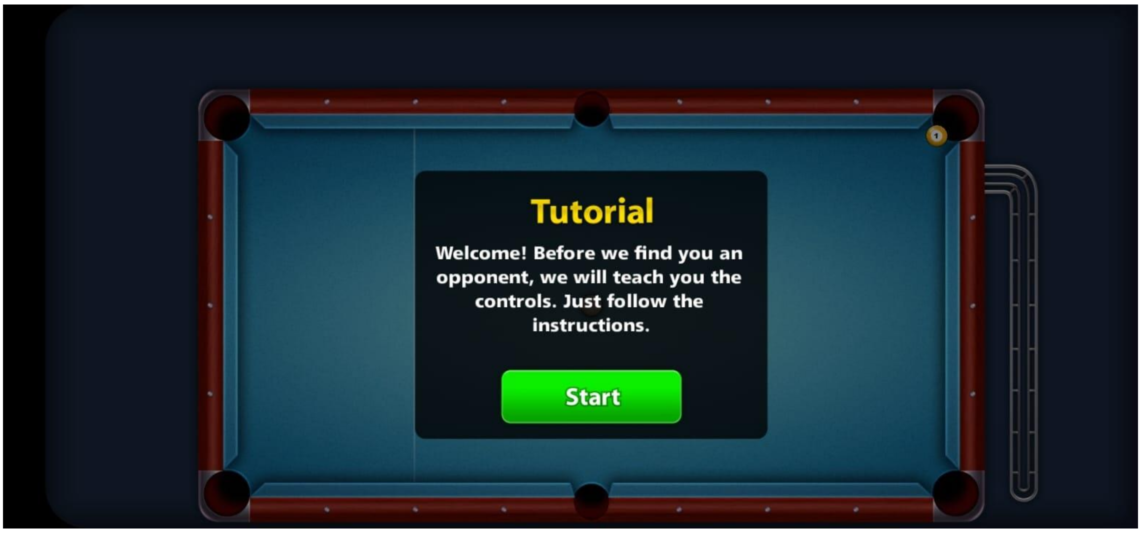


Image 1: Starting screen of 8 Ball Pool

In this screen the player can choose how they want to login to the game or do they want to play as a guest. The login type the player chooses is not relevant for this study.

After the player has chosen how to login, tutorial of the game starts. This view is seen in image 2.

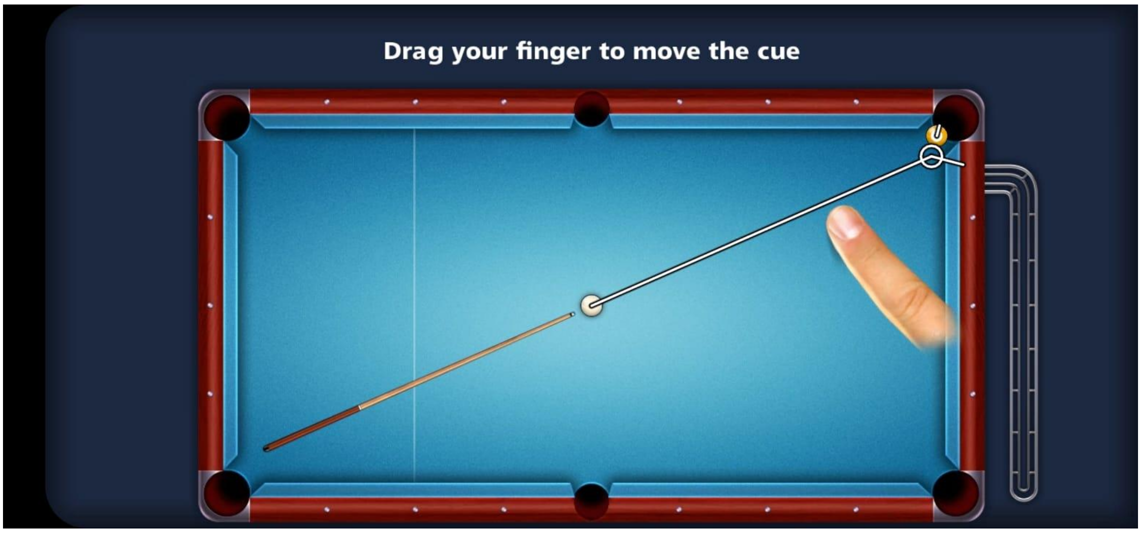


Tutorial

Welcome! Before we find you an opponent, we will teach you the controls. Just follow the instructions.

Image 2: First tutorial screen of 8 Ball Pool

After this first tutorial screen the tutorial starts which has the two main phases shown in images 3 and 4.



Drag your finger to move the cue

Image 3: First part of tutorial

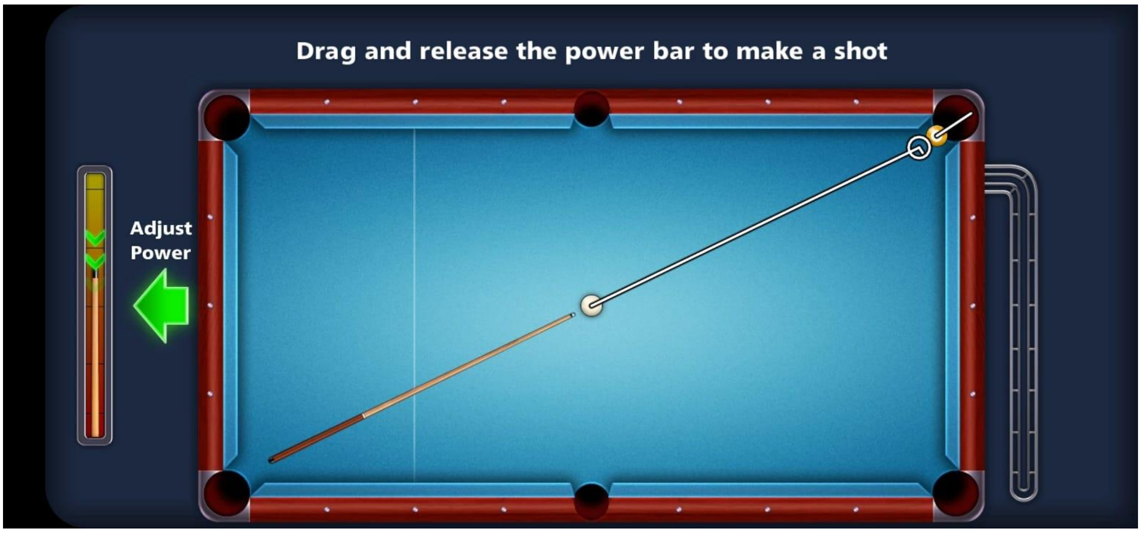


Image 4: Second part of tutorial

After the tutorial is completed, the player is taken to a normal match. This view of a normal match is seen in image 5.

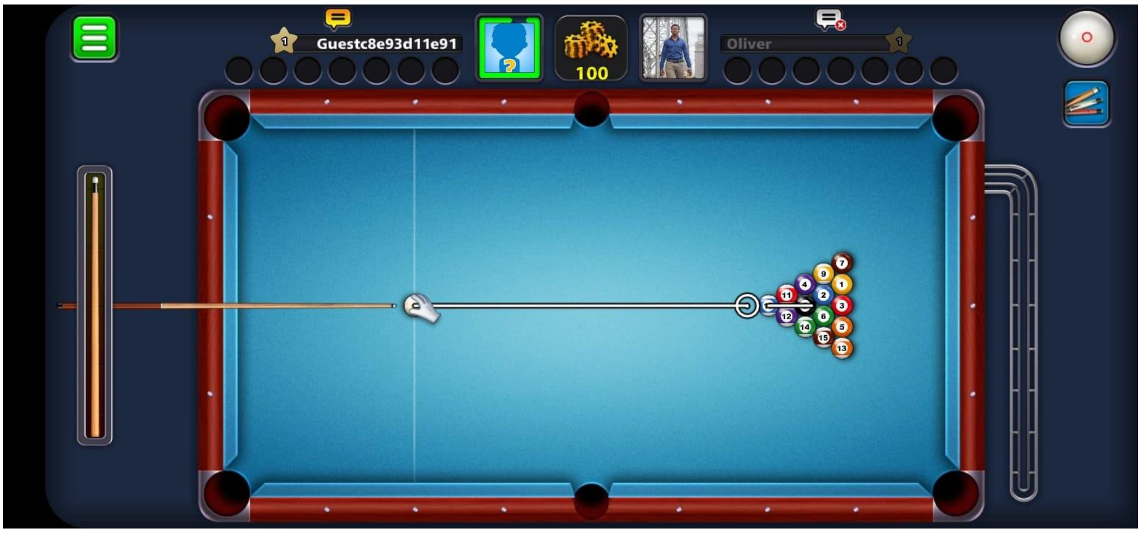


Image 5: Standard view of playing a normal match.

When the player completes this match or quits in the middle of it, they are taken to the main menu of the game. This is seen in image 6.

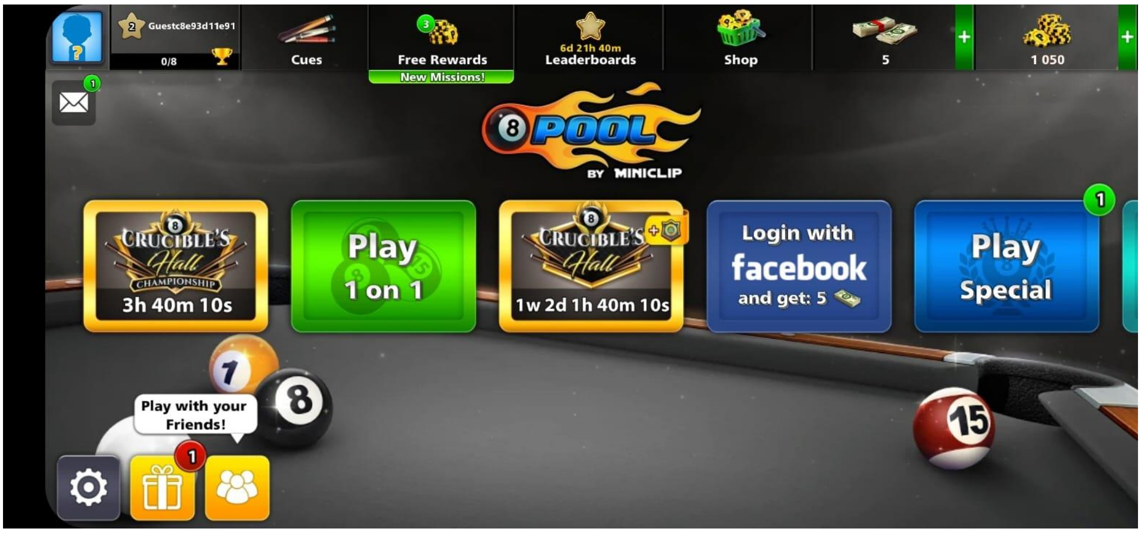


Image 6: Main menu of 8 Ball Pool

The game specific tasks assigned to 8 Ball Pool for the player were to play at least one match and to visit the shop and navigate through it. Playing one match makes sure the player knows what the game is all about. No more than one match is required to be played because it is still just a pool game so there shouldn't be anything that new to the player. The player is required to visit the shop so they would get a better understanding of the game's user interface and overall look.

4.1.4 Gem Stack

The second game is Gem Stack by Rollic games. Gem Stack was chosen as the game that was still in development because the game was released in February 2022, so it is quite new. It is also updated quite often and at least in the author's opinion is quite narrow in its contents still, so it was considered as a game that is still heavily in development. [43] Gem Stack was also a game with no multiplayer elements in it. When the player first opens the game, they are greeted with the following view which is seen in image 7.

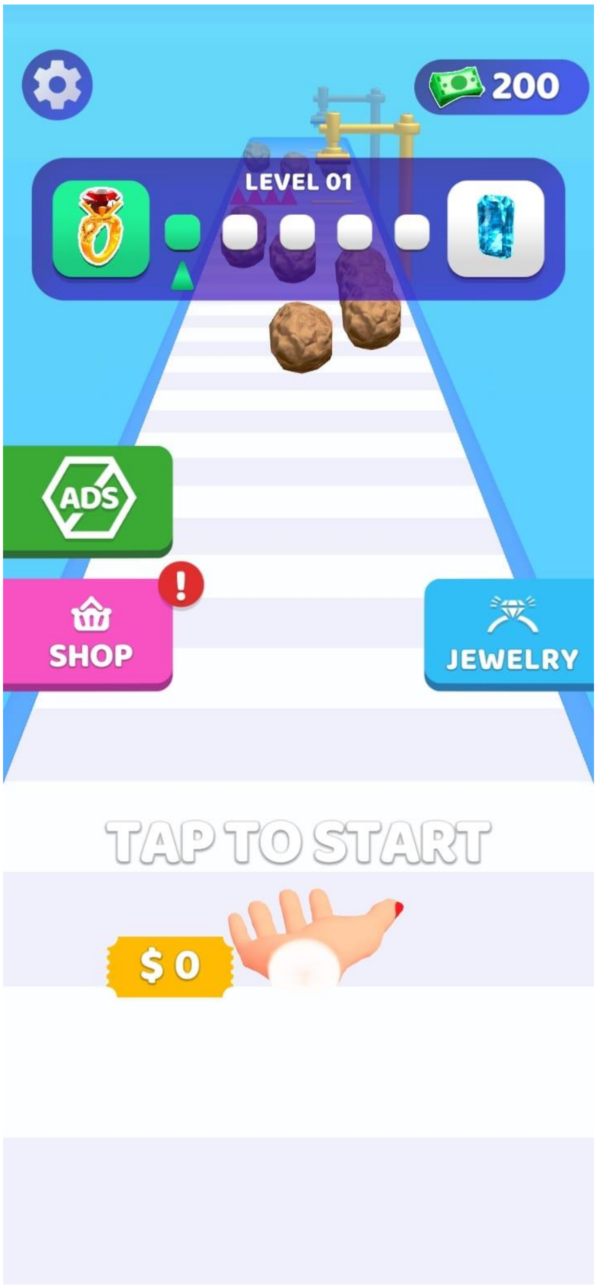


Image 7: Starting view of Gem Stack

This game didn’t have any starting tutorial like 8 Ball Pool had and the player is taken straight into the game. The goal of the game is to collect these rocks and go through different gates/mechanics. First these rocks are changed into rough looking gems. Then then the player can go through different gates to smooth out the gems and finally turn them into rings or other jewelry. In the end of a level the player can sell all the jewelry and the player’s score is set by the amount of money they get. This is the core loop of the game and after this the player is taken to next level. The core loop is presented in images 8, 9, 10, 11, 12.

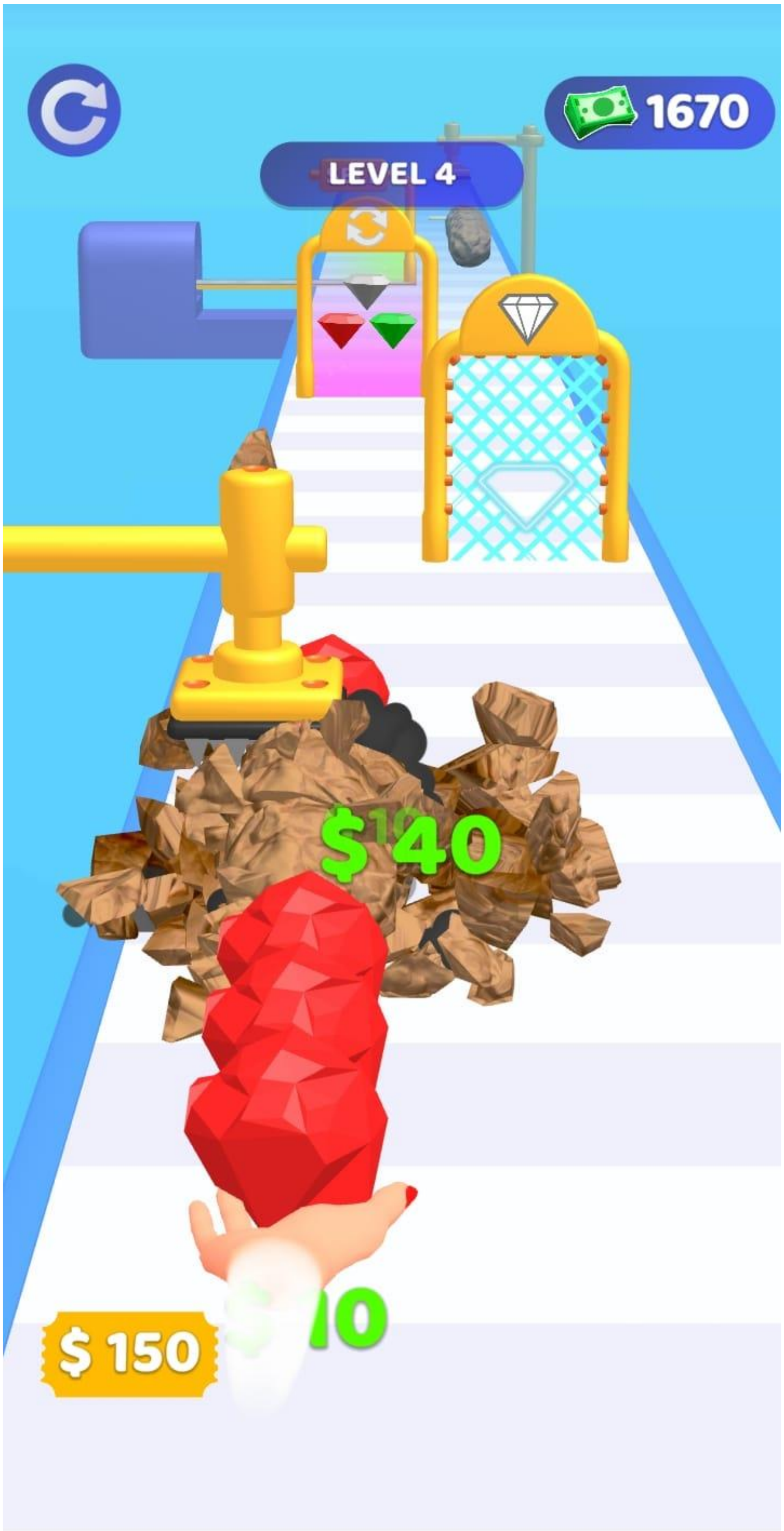


Image 8: After collecting the rocks, the player goes through a crusher to change them into rough gems.

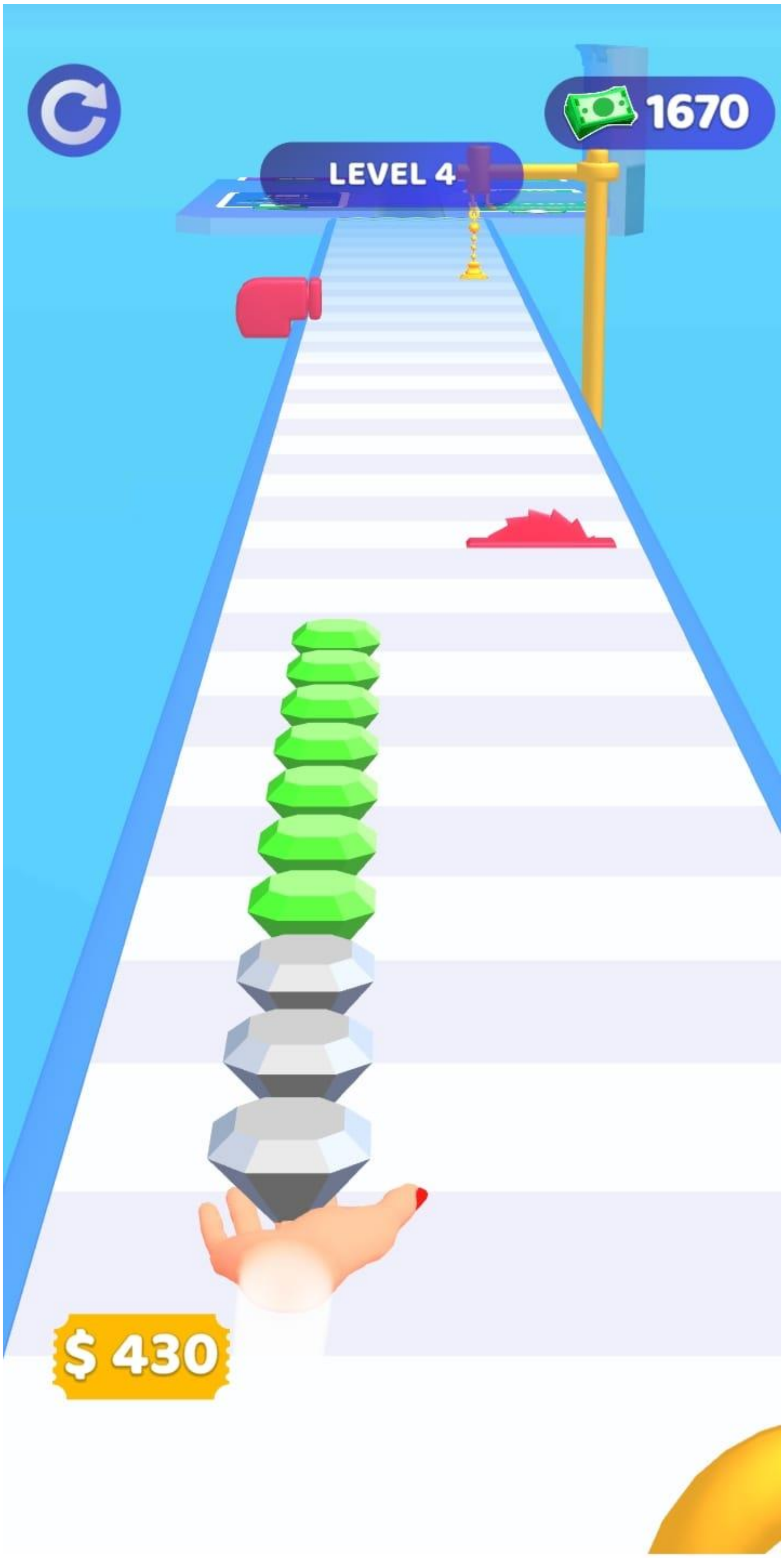


Image 9: After the player has gone through a gate, they have more refined gems.

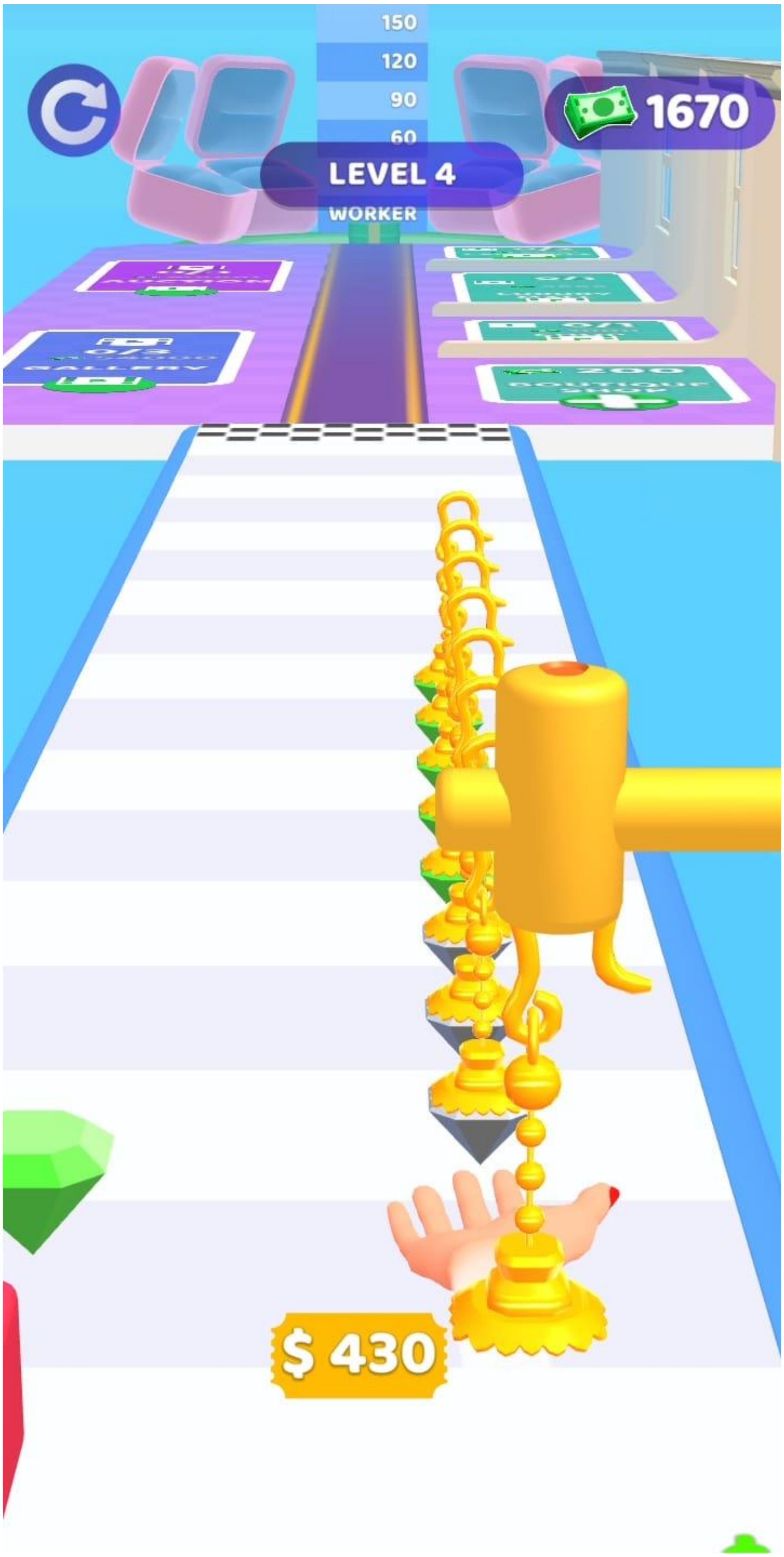


Image 10: Player goes through different objects to turn the gems into jewelry.

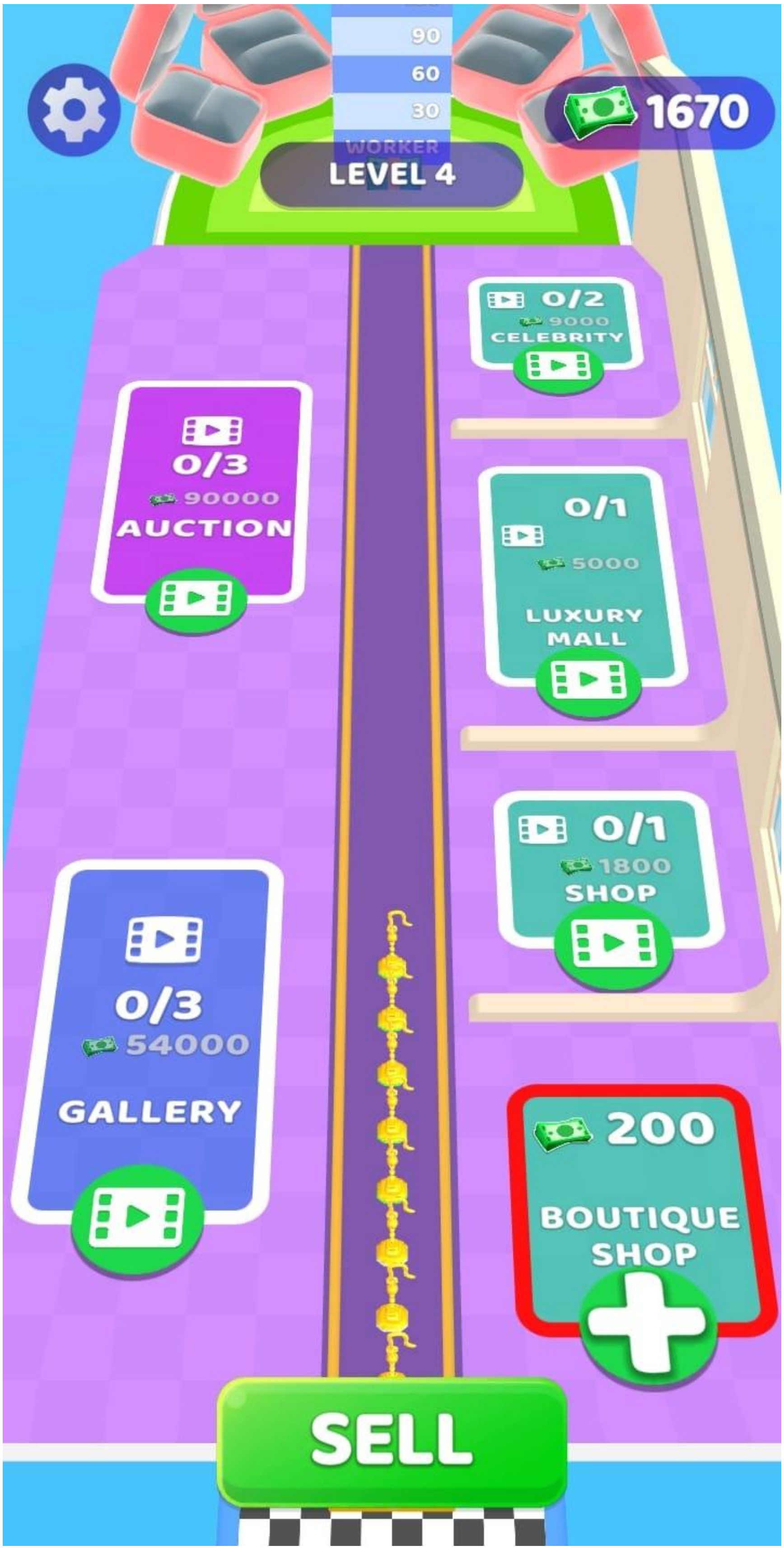


Image 11: End scene where the player sells the objects they gathered.

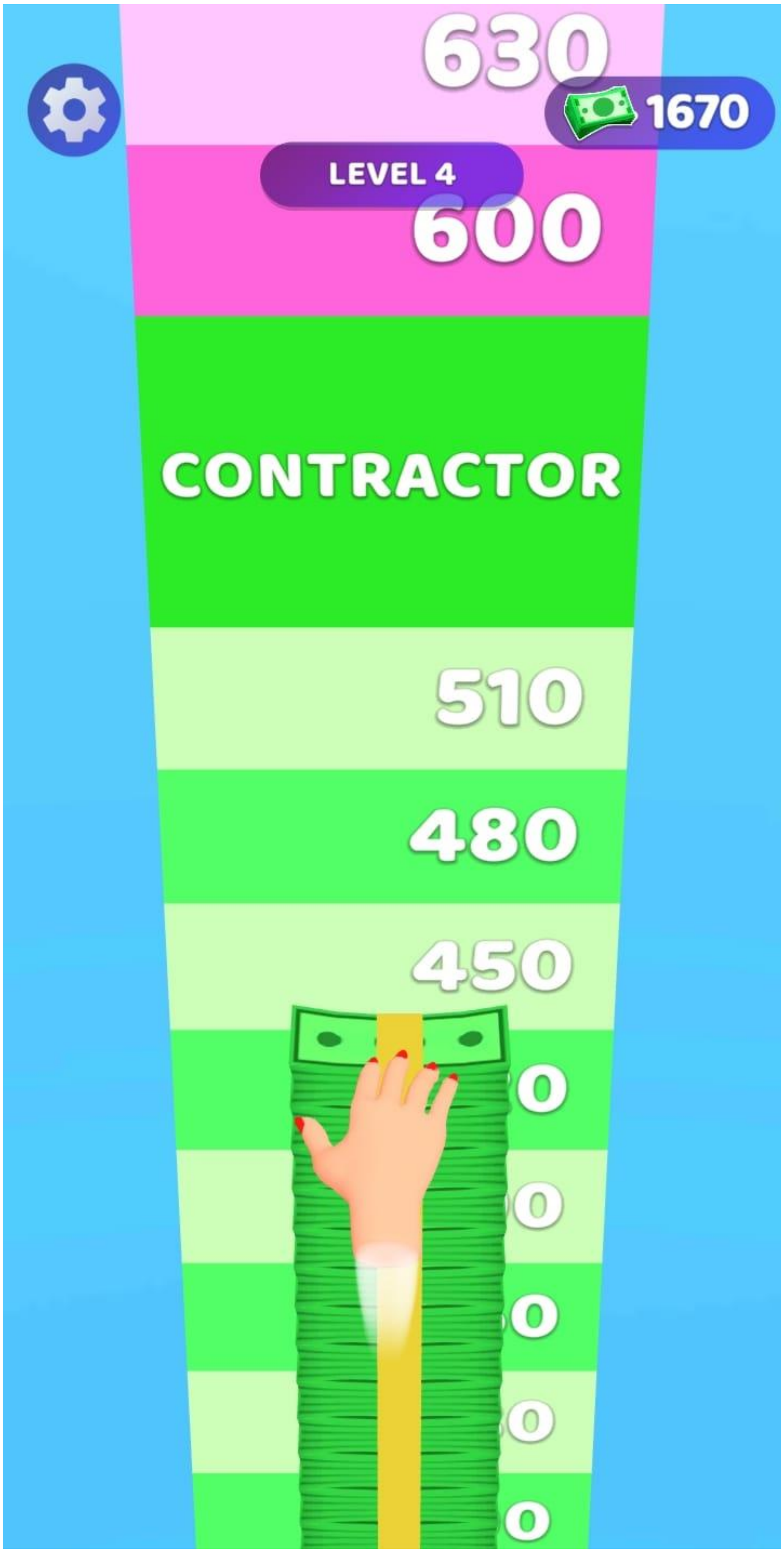


Image 12: Final animation where the height of the money stack the player got assigns the score.

The two goals assigned for this game were to play at least 5 levels and to visit the shop. The goal to play at least 5 levels was assigned because the game gradually introduces new gates and mechanics to the game through new levels so the player should at least play 5 levels to see the what the game is about. The second goal to visit the shop was assigned for the same reason as in 8 Ball Pool so that the player gets an understanding of the game's user interface and look.

The two games presented above were released in times that were very different from each other so criteria 4 was not that successful. However, both games have had updates that were close to each other, so I felt like they are acceptable and overall, the criteria 4 was not that important in the authors opinion compared other criteria such as 5 or 7. When it comes to criteria 8, these games had no critical or disturbing bugs so at least in the authors opinion so criteria 8 was met.

4.2 Participants

The participants for the research consisted of four game developers and four players. All the game developers were recruited from a Discord server with multiple game developers. The players were recruited from the authors' acquaintances. The participants and their roles in the order of recruiting can be seen in table 3.

Table 3: Participant roles

ID	Role
1	Game developer
2	Player
3	Player
4	Game developer
5	Game developer
6	Player
7	Player
8	Game developer

All the participants were male and were aged between 23-38.

Participant 1 was a game developer and their main roles in game development were game design, UI design, textures, 3D modeling and programming. Participant 1 had worked professionally with games for 1-2 years and had made games for mobile. He had worked in 3-4 game projects which were action/arcade games. Participant 1 also categorized himself as midcore/core video game player who plays 30-39 hours a week and usually 4-5 hours in one sitting. Participant 1 had experience with web design heuristics from a school project.

Participant 2 was a player that categorized himself as a casual video game player who plays video games 5-9 hours in a week and usually 3-4 hours in one sitting. Participant 2 played PC games and usually played shooting, sports, multiplayer, or role-playing games. Participant 2 had no experience with heuristics.

Participant 3 was a player that categorized himself as hardcore/expert video game player who plays video games 10-19 hours in a week and usually 4-5 hours in one sitting. Participant 3 played PC games and usually they were shooting, multiplayer or action role-playing games. Participant 3 had no experience with heuristics.

Participant 4 was a game developer, and their main roles were programming, game design and balancing. Participant 4 had worked professionally with games for 1-2 years and made games to mobile platform. He had worked on 3-4 game projects which were shooting, action or simulation games. Participant 4 categorized himself as midcore/core video game player who plays 20-29 hours in a week and usually 4-5 hours in one sitting. Participant 4 had experience from “general heuristics related to user interface” from university.

Participant 5 was a game developer, and their main roles were programming and balancing. He had worked professionally with games for 1-2 years and had made games for PC. He had worked on 3-4 game projects which were puzzle, adventure, shooting, strategy, or action games. Participant 5 categorized himself as midcore/core video game player who plays more than 40 hours in a week and usually 8 hours or more in one sitting. Participant 5 had no experience from heuristics.

Participant 6 was a player that categorized himself as hardcore/expert video game player who plays video games 20-29 hours in a week and usually 4-5 hours in one sitting. Participant 6 played PC and mobile games and usually they were adventure, shooting, strategy, action, driving, multiplayer, music/exercise/rhythm, or online role-playing games. Participant 6 had no experience with heuristics.

Participant 7 was a player that categorized himself as casual video game player who plays video games 5-9 hours in a week and usually 1-2 hours in one sitting. Participant 7 played console games and usually they were sports games. Participant 7 had no experience with heuristics.

Participant 8 was a game developer, and his main roles were lead developer and tech lead as in which they were responsible of used technologies, game engines, platforms and taking care of game logic and advanced coding. He had worked professionally with games for 5-10 years and had made games for mobile, PC and VR. They had worked on 7-10 game projects which were puzzle, adventure, shooting, simulation, multiplayer, role-playing or online role-playing games. Participant 8 categorized himself as midcore/core video game player who plays 5-9 hours in a week and usually 1-2 hours in one sitting. Participant 8 had experience with heuristic and had used them professionally/in their work. Participant 8 mentioned that they had used heuristics in game programming as temporary solutions to be later converted into more sophisticated solutions and sometimes left as they were.

All eight participants had approximately started to play video games at the age between the scale of 0-10.

4.3 Research process

The research started with a screening questionnaire to filter out possible participants that were applicable to join the study and those who weren’t. The questionnaire was sent to a Discord server with an explanation of the goal of the study and that it is part of the authors master’s degree. Within the explanation the author also informed the possible participants that basic English is needed in the study because the questions and heuristics were in English. Smart phone would also be needed to play the games in the evaluation sessions. The Discord server was a part of Finnish Game Incubator 2021 and there were several game developers and mentors. In addition to general explanations of the study, all the people were also instructed that between all the people that successfully complete the study, one person is picked to win a steam gift card worth 25€. All the people interested in taking part of the study were instructed to fill out the questionnaire. The goal of this questionnaire was to figure out are the people who answer it game developers or just people how play games. The definition the author made in this initial questionnaire for game developers was working with game development for at least 1 year. The decision to choose what is considered working with game development was left for the participants to decide but I also asked them to describe the roles and jobs they have had. By adding this additional question, I could filter out people who had said they consider themselves as game developers but for example only worked on music for the games which I didn't consider as proper game development. The questionnaire had a question regarding how much the participants play videogames in a week ranging from less than 1 hour to more than 40 hours. The questionnaire also had question about what kind of phone the participants had. This was relevant because the games chosen for the study were confirmed to be available for Android and iOS. If the participant's mobile device would have been something else, I could then exclude them from the study. The initial questionnaire also had a question regarding the participant's experience with heuristics. This question could be used in further filtering of the participants that move to the second stage of the study. The initial questionnaire can be seen in appendix C.

After this initial questionnaire I had a list of initial participants and in the next phase I filtered this list of participants to have the participants that best suited for the study.

When choosing the final participants, I used the following criteria:

1. *As a player you need to play at least 5-9 hours of video games in a week.* This criterion was chosen so that the participants are people who consistently to some degree play games and not just people who try some game and play it for an hour and then plays again after two weeks.
2. *As a developer your roles/jobs need to be related to game development.* The definition for jobs that relate to game development were made by the author. This meant that you need have some sort of programming or design compared to something like only making music for the game.

Information about these final participants can be seen in chapter 4.2.

These final participants were contacted to let them know they are part of the study, and they were asked to fill out a consent form for the study and a background questionnaire. In this background questionnaire I asked basic questions like age, gender, and education but in addition to these there were multiple questions related to what kind of player the participant is, what kind of platforms they play in and how long they have played. These were asked to gain insight on what kind of players they really are and to find out could these factors affect their results in the upcoming evaluations. In this questionnaire there was also a question about the model of the participant's phone. This was asked in case someone reported during the evaluation sessions that the games they are playing don't run smoothly or any similar issues. Remaining questions in the questionnaire were focused for the game developers. These questions focused on how many years they have worked with games and what kind of games they have been working on. These were asked to also gain information on factors that could affect the participant's answers in the evaluation sessions. These questions were asked to get information about the participants knowledge and experience in the relevant field that is being evaluated based on the recommendation made by Thewes et al. [36]. Consent form can be seen in appendix D. This background questionnaire can be seen in appendix E.

After the participants had filled out the consent form and background questionnaire successfully, they were informed about the basic steps of the evaluation session. A suitable time to complete the evaluation session was discussed with the participant in Discord chat. The participants were instructed to schedule at least 2.5 hours of time for the sessions. At least one day before the actual sessions were held, the author of this study sent the participants the heuristic evaluation form they would use in the sessions and instructed them to look at the heuristics and all the tasks they would need to do in the evaluation. The participants were instructed that this document would be browsed through together with the conductor of the study so if they would have any questions, they would have an opportunity to ask. The participants were also instructed that they could print the document if it was easier to them but in the end, they would have to send it back to the conductor of the study. The heuristic evaluation forms, one for each game were sent to the participants. The evaluation forms contained brief introductions to the evaluation sessions. They also contained the instructions for the two tasks the participants had to complete during the sessions. This evaluation form also had two written examples of how to answer to both tasks. These were added to try to teach the participant how to perform the evaluations similarly like in the studies made by Salian et al, Al-Razgan et al, Pinelle et al. [8,20,32]. This form also contained the full list of heuristics, severity scale with descriptions and examples of fulfillment related to the second task explained below. The heuristic evaluation form is presented in appendix F.

In the actual evaluation sessions, the participant and the author of the study were in a remote Discord call. The participants were welcomed and explained again the evaluation procedure. The participants would need to evaluate two games that were described in chapter 4.1 of this study. The order of these games was changed so that there was an equal amount of participants who started with game one and game two. The participants were instructed to spend their time playing the game and performing the actual evaluation as they wanted similarly as suggested in a study by Thewes et al [36]. The participants were instructed to use as much time as they wanted for the evaluations and to let the conductor of the study know they are done. After 60 minutes if the participants weren’t finished, they would be stopped, and the evaluation session would continue. In this evaluation session the participants had to complete two tasks while playing the game. In the first task they had to classify usability problems in the game using the heuristics created in chapter 3 of this study and a severity scale. This task was added to the study based on multiple similar studies in literature [8,20,24]. The severity scale was done according to these multiple examples from literature. The severity scale used in the study can be seen in table 4.

Table 4: Severity scale used in the evaluations [8]

Value	Description
0	I don’t agree that this is a usability problem at all
1	Cosmetic problem only – need not be fixed unless extra time is available on the project
2	Minor usability problem – fixing the problem should be given low priority
3	Major usability problem – important to fix, so should be given high priority
4	Usability catastrophe – imperative to fix this before product can be released

In the second task the participants had to assign a score from 1 to 5 (1 being worst, 5 being best) to every single heuristic based on how well the game fulfilled each of them. This task was added to the study based on the study made by Koeffel et al [24]. Examples of different fulfillments were added to the form that can be seen in table 5.

Table 5: Examples of different fulfillments

Value	Description
1	The game does not fulfill the heuristic at all
2	The game fulfills the heuristic poorly
3	The game moderately fulfills the heuristic
4	The game mostly fulfills the heuristic
5	The game completely fulfills the heuristic

Even though these examples were added, the participants were instructed to use their own judgement on what does different number of fulfillments mean to them because these example descriptions were created by the author of this study to possibly help the evaluators. Once the participants were done, they were instructed to inform the conductor of the study.

After the first game was evaluated, there was a short 15min break where the participant was given the chance to breathe a little or visit the bathroom.

After the break, the evaluation of the second game started with the same instructions as with the first game and maximum time they were allowed to spend evaluating the game was 60 minutes again.

After the game evaluations were done, the participants were asked to fill out the System Usability Scale (SUS) -questionnaire to evaluate the heuristics. SUS is a Likert scale that is used to get a global view of subjective assessments of usability. SUS is a 5-point scale (1-5) that indicates the degree of agreement and disagreement. [44] SUS-questionnaire questions can be seen in table 6.

Table 6: SUS-questionnaire questions [44]

I think that I would like to use this system frequently
I found the system unnecessarily complex
I thought the system was easy to use
I think that I would need the support of a technical person to be able to use this system
I found the various functions in this system were well integrated
I thought there was too much inconsistency in this system
I would imagine that most people would learn to use this system very quickly
I found the system very cumbersome to use
I felt very confident using the system
I needed to learn a lot of things before I could get going with this system

In this study I used the SUS-questionnaire to evaluate the list of heuristics, so I changed the word “system” to “list of heuristics” to make the questionnaire more understandable for the participants. The word system would most likely confuse them so they wouldn’t know what they are evaluating with the questionnaire. This word changing was also done by Gomez et al. when they used the SUS-questionnaire in their study where they applied the questionnaire for evaluation tools they created. They changed the word “system” to “tool”.[45]

After filling out the SUS-questionnaire the participants were asked to describe the experience of doing the evaluations and all its strengths and weaknesses and the strengths and weaknesses of the heuristics. This was done by having a semi-structured interview. The main questions asked in the interview can be seen in table 7.

Table 7: Interview questions

1. How did the game evaluations go in your opinion?
2. How did you do the evaluations?
3. What did you think about the games you played?
4. Were you able use the heuristics successfully when evaluating the games?
5. Are these heuristics easy to understand? (in general)
6. What are the strengths of these heuristics?
7. What are the limitations of these heuristics?
8. Which heuristics are the most useful?
9. Which heuristics are difficult to understand?
10. Does the heuristics cover all usability problems in these games you evaluated?
11. Do you think these heuristics would be useful in game development? Why? Why not?
12. (Game developers) How useful do you think these tasks would be if you could get an evaluation filled out by the players about your own game

Questions 5, 6, 7, 8, 9 and 10 were same as in the study by Al-Razgan et al [8]. These questions mainly focused on the heuristics the participants used in the evaluations. Through these questions I tried to gain knowledge on what the participants thought about the heuristics in general and were there any single heuristics that were confusing to them. In addition, there were questions the author of this study made (1, 2, 3, 4). There was a question about how the evaluations went in the participants opinion and questions about their overall feelings about the heuristics. There was also a general question about what the participants thought about the games they played. This was asked because if the participants would answer that they totally hated the game, this could subconsciously affect the results of evaluations the participants did. Questions 11 and 12 were also created by the author of this study. In question 11 I tried to gain even more insight of the usefulness of these new heuristics. Question 12 was asked from the game developers to get insight about the overall evaluation tasks and their usefulness in real life scenarios.

After the interviews were done and the participants had filled out the questionnaires the evaluation sessions were finished, and the participant was thanked for their participation.

4.4 Analyse methods

After completing all the evaluations with all the participants, the results were analysed with the following methods.

Answers to the SUS-questionnaires were scored based on the system introduced by Brooke [44]. First the sum of the scores were calculated. For items 1, 3, 5, 7 and 9 of the SUS-questionnaire the score contribution was the scale position minus 1. For items 2, 4, 6, 8 and 10 the score contribution was 5 minus scale position. The sum of scores was multiplied with 2.5 to get the overall SUS score [44]

The evaluator effect was calculated using the Any-two agreement which is shown in the formula below [1] .

$$Any - two\ agreement$$
$$= Average\ of\ \left|\frac{Pi\ \cap Pj}{Pi\ \cup Pj}\right|\ over\ all\ {}^{1}/_{2}\, n(n-1)\ pairs\ of\ evaluators$$

Formula of any-two agreement [1]

The evaluator effect was calculated for the group consisting of game developers and also for the group consisting of people who play games henceforth players. For calculating the evaluator effect, the problems used for making the calculations were defined as the two highest severities of the severity scale. This means all heuristics which had the severity of 3 or 4 were considered problems.

In addition to these, the number of critical errors (severity 4) with game developers and players were analysed and tabled using MS Excel. Also, the number of non-errors (severity 0) recorded by game developers and players were analysed and tabled using MS Excel.

The results of the evaluation sessions were analysed quantitatively by tabling the percentages of severity ranking scales between game developers and players similarly as in the study made by Al-Razgan et al. [8]

Relating to the second task of the evaluations the results of the heuristics’ fulfillment to the games used in the study were analysed by getting the sum of the ratings by single evaluators and the calculation of average ranking. This score was converted into a percentage scale indicating to which degree the game complied with the heuristics. This was done similarly as in the study made by Koeffel et al. [24]

Also, the times that the participants used in the study are analysed and tabled between game developers and players using MS Excel.

The interview answers were analysed by doing a thematic analysis using MS Excel. The answers were classified into different themes and answers to the questions were categorized based on the themes. The number of answers related to specific themes were summarized and the all the themes found for all the questions along with the amounts were tabled using MS Excel.

5. RESULTS

In this section I will go through all the results from the evaluations. First, I will present the overall times that the participants spend doing the evaluation along with the order of the games played. After this I will present all the severity rankings that the participants classified for the games. I will present all the any-two agreements that indicate the evaluator effect. I will also present all the fulfillments for the heuristics and the results from System Usability Scale (SUS). Finally, I will go through the interview answers and all the themes found from the answers.

Table 8 shows all the participants and the order of games they played in the evaluation sessions and the time they spend playing/evaluation specific game.

Table 8: Participants and the games they played along with the times they took.

Participant ID	Role	Game 1	Game 2	Time for Game 1	Time for Game 2
1	Developer	8 Ball Pool	Gem Stack	24:35:31	15:54:06
2	Player	8 Ball Pool	Gem Stack	49:29:46	35:03:60
3	Player	Gem Stack	8 Ball Pool	34:54:14	34:49:94
4	Developer	Gem Stack	8 Ball Pool	18:35:60	16:38:31
5	Developer	8 Ball Pool	Gem Stack	27:58:29	30:04:57
6	Player	Gem Stack	8 Ball Pool	38:47:95	18:01:69
7	Player	8 Ball Pool	Gem Stack	17:56:43	14:50:32
8	Developer	Gem Stack	8 Ball Pool	28:46:29	22:23:44

The results show that time it took to do the evaluations for each game varied a lot from 14 minutes to 49 minutes. Participants 4 and 7 did the game evaluation faster than other participants. Participant 2 took the longest time to do the evaluations.

5.1 Severity rankings

In this section I will go through all the severity ranks for all the games classified by players and game developers.

Table 9 shows the individual amounts of different severity rankings for all the heuristics for Players evaluating 8 Ball Pool

Table 9: Number of severity rankings for all the heuristics for players with 8 Ball Pool.

Severity ranking scale (8 Ball Pool)	PLAYER 1 (ID 2)	PLAYER 2 (ID 3)	PLAYER 3 (ID 6)	PLAYER 4 (ID 7)
No problem	20	18	20	11
Cosmetic	2	3	5	13
Minor	3	4	0	1
Major	0	0	0	0
Catastrophe	0	0	0	0

The results show that none of the players marked a severity value of “Major” or “Catastrophe”. Most of the players had ranked the heuristics with a rank of “No problem”. Player 4 had a majority of “Cosmetic” problems.

Table 10 shows the individual amount of different severity rankings for all the heuristics for players evaluating Gem Stack

Table 10: Severity rankings for players with Gem Stack.

Severity ranking scale	PLAYER 1 (ID 2)	PLAYER 2 (ID 3)	PLAYER 3 (ID 6)	PLAYER 4 (ID 7)
No problem	9	13	13	0
Cosmetic	4	5	4	1
Minor	6	4	5	4
Major	5	1	3	10
Catastrophe	1	2	0	10

The results show that all the players had very different rankings for the severities. Player 1 had all kinds of problems nearly even amount and had only one “Catastrophe”. Player 2 had mostly “No problem” rankings but also all the other ones with two “Catastrophe” rankings. Player 3 had also mostly “No problem” ranks but also the other ones excluding

“Catastrophe”. Player 4 had zero “No problem” ranks and mostly had “Major” or “Catastrophe” rankings.

Table 11 shows the individual amount of different scale errors for Game developers evaluating 8 Ball Pool

Table 11: Severity rankings for developers with 8 Ball Pool.

Severity ranking scale	DEVELOPER 1 (ID 1)	DEVELOPER 2 (ID 4)	DEVELOPER 3 (ID 5)	DEVELOPER 4 (ID 8)
No problem	15	19	22	21
Cosmetic	8	3	3	1
Minor	2	2	0	1
Major	0	1	0	0
Catastrophe	0	0	0	0

The results show that none of the developers had “Catastrophe” ranks. None of the developers except developer 2 had “Major” ranks. Developer 2 having one “Major” rank. All the developers had a majority of “No problem” ranks.

Table 12 shows the individual amount of different scale errors for Game developers evaluating Gem stack.

Table 12: Severity rankings for developers with Gem Stack.

Severity ranking scale	DEVELOPER 1 (ID 1)	DEVELOPER 2 (ID 4)	DEVELOPER 3 (ID 5)	DEVELOPER 4 (ID 8)
No problem	12	15	6	9
Cosmetic	5	1	6	6
Minor	3	3	5	1
Major	2	6	0	2
Catastrophe	3	0	8	6

The results show very different amounts for different ranks. Developer 1 had a majority of “No problem” with having all of the other rankings as well with three “Catastrophe” ranks. Developer 2 had a majority of “No problems” ranks and zero “Catastrophe” ranks along with some of the other severities. Developer 3 had a majority of “Catastrophe” ranks with no “Major” severities. Developer 3 also had six “No problem” severities which was the same amounts as “Cosmetic” and one more than “Minor” severities. Developer

4 had a majority of “No problem” severities with six “Catastrophe” severities along with severities from other scales.

Table 13 shows the percentages of different scale errors with players for both games.

Table 13: Percentages of severity rankings by players.

Severity ranking scale %	8 Ball Pool	Gem Stack
No problem	69%	35%
Cosmetic	23%	14%
Minor	8%	19%
Major	0%	19%
Catastrophe	0%	13%

The percentages indicate that with players, 8 Ball Pool mostly had no problems but some cosmetic ones along with few minor problems. With Gem Stack player also had the majority of “No problem” severities along with reasonably even amount of all the other severities but a bit more “Minor” and “Major” severities compared to “Cosmetic” and “Catastrophe”.

Table 14 shows the percentages of different scale errors with game developers for both games.

Table 14: Percentages of severity rankings by game developers.

Severity ranking scale %	8 ball pool	Gem stack
No problem	77%	42%
Cosmetic	15%	18%
Minor	5%	12%
Major	1%	10%
Catastrophe	0%	17%

The percentages indicate that with game developers, 8 Ball Pool mostly had no problems but some cosmetic ones along with few minor problems and only 1% of major ones. With gem stack game developers also had the majority of “No problem” severities along with all other severities having “Cosmetic” and “Catastrophic” severities a little more compared to “Minor” and “Major” severities.

5.2 Evaluator effects

In this section I will present all the Any-two agreements that indicate the evaluator effect along with the individual agreements between all the possible pairs. Any-two agreement was defined as the number of problems two evaluators have in common divided by the number of problems the collectively detect, averaged over all possible pairs of two evaluators [1]. Problems were defined as heuristics that had severity of 3 or 4 (Major or Catastrophe).

Table 15 shows the any-two agreement with players for both games along with agreements for all the pairs of players.

Table 15: Evaluator effect for players.

Pair / game	Player1(ID2) & Player2(ID3)	Player1(ID2) & Player3(ID6)	Player1(ID2) & Player4(ID7)	Player2(ID3) & Player3(ID6)	Player2(ID3) & Player4(ID7)	Player3(ID6) & Player4(ID7)	Any-2
8 Ball Pool	0%	0%	0%	0%	0%	0%	**0%**
Gem stack	33.3%	22.2%	23.1%	33.3%	13.1%	13.1%	**23.02%**

The results show that for 8 Ball Pool there was no agreements at all between any of the players so the final Any-two agreement for players and 8 Ball Pool was **0**%. For Gem stack all the players had some agreements with each other. Players 1 & 2 and players 2 & 3 had the biggest agreements with 33.3% and players 2 & 4 and 3 & 4 had the smallest agreements with 13.1%. The final Any-two agreement for Gem stack was **23.02**%

Table 16 shows the any-two agreement with game developers for both games along with agreements for all the pairs of developers.

Table 16: Evaluator effect for game developers.

Pairs / game	Dev 1(ID1) & Dev 2(ID4)	Dev 1(ID1) & Dev 3(ID5)	Dev 1(ID1) & Dev 4(ID8)	Dev 2(ID4) & Dev 3(ID5)	Dev 2(ID4) & Dev 4(ID8)	Dev 3(ID5) & Dev 4(ID8)	Any-2
8Ball Pool	0%	0%	0%	0%	0%	0%	**0**%
Gem Stack	18.2%	0%	15.4%	21.4%	21.4%	18.8%	**15.87**%

The results show that for 8 Ball Pool there was no agreements at all between any of the developers so the final Any-two agreement for 8 Ball Pool and developers was **0**%. For Gem stack all the developer pairs except developers 1 and 3 had some agreement. Developers 2 & 3 and 2 & 4 had the biggest agreements with 21.4%. Developers 1 & 4 had the smallest agreement with 15.4% excluding the developer pair 1 & 3 where there was no agreement at all. The final Any-two agreement for Gem stack and game developers was **15.87**%

5.3 Fulfillment of heuristics

In this section I will present all fulfillments for the heuristics from task 2 of the evaluation form. In task 2 the participants had to assign a score from 1 to 5 (1 being the worst, 5 being best) to every single heuristic based on how well the game fulfilled each of the. Maximum fulfillment being 125 where all the 25 heuristics have a score of 5.

Table 17 shows the total fulfillment values for players for both games.

Table 17: Player fulfillments.

	Player 1 (ID 2)	Player 2 (ID 3)	Player 3 (ID 6)	Player 4 (ID 7)
8 Ball Pool	101	88	108	105
Gem Stack	61	67	82	46

The results show that according to all of the players 8 Ball Pool complied with the heuristics more compared to Gem stack. Player 2 gave a fulfillment of 88/125 for 8 Ball Pool which was the lowest compared to other players and 8 Ball Pool. All the players gave a reasonably low fulfillments for Gem stack. Player 3 giving the highest of 82/125 and player 4 giving the lowest of 46/125.

Table 18 shows the individual fulfillment values for game developers for both of the games.

Table 18: Developer fulfillments.

	Developer 1 (ID1)	Developer 2 (ID4)	Developer 3 (ID5)	Developer 4 (ID8)
8 Ball Pool	109	105	113	116
Gem Stack	85	78	55	68

The results show that according to game developers 8 Ball Pool complied with the heuristics more compared to Gem Stack. All the developers gave reasonably high fulfillments the highest being Developer 4 with a fulfillment of 116/125. All the developers gave a reasonably low fulfillments for Gem stack. Developer 1 giving the highest of 85/125 and Developer 3 giving the lowest of 55/125.

Table 19 shows the average ranking of how both games complied with the heuristics according to players and game developers.

Table 19: Average ranking of heuristic compliance.

Game	Players	Game developers
8 Ball Pool	80.4%	88.6%
Gem Stack	51.2%	57.2%

The results show that according to both players and game developers 8 Ball Pool complied with the heuristics more than Gem Stack. Compared to players, game developers had higher rankings for both games.

5.4 System Usability Scale (SUS)

In this section I will present the results of the System Usability Scale (SUS)-questionnaire.

Table 20 shows the SUS-questionnaire scores for all the participants.

Table 20: SUS scores.

Participant ID	SUS (X out of 100)
1 (Developer)	55
2 (Player)	87.5
3 (Player)	30
4 (Developer)	77.5
5 (Developer)	97.5
6 (Player)	67.5
7 (Player)	42.5
8 (Developer)	95

The results show that participants 5 and 8 had very high scores. Participant 3 however had a reasonably low score. Based on the scores the average score for developers is 81.25 and the average score for players is 56.875.

5.5 Results from interviews

In this section I will go through all the interview questions and describe the themes found from the answers.

Q1: **How did the game evaluations go in your opinion**?

The themes found and their amounts are shown in table 21.

Table 21: Themes for Q1 and the number of finds (N=8).

Theme	Number of finds
Problems with a game	3
No big problems but a few problematic heuristics	3
No problems at all	3
Problems with English	1

Problems with a game: three out of eight participants mentioned they had issues with a game, specifically Gem stack. The issues in this case meant that the participants hated the ads and had issues with the actual game and not the evaluation of the game.

Participant 5 (P5) (developer): *I am very annoyed about the last game. I am totally fed up with the ads.*

P8(developer): *First game very annoying. Felt like it was getting stuck because of the ads.*

No big problems but a few problematic heuristics: three out of eight participants mentioned that they didn't have any bigger issues but some of the heuristics were weird for them.

P1(developer): *No major problems, but there were a few heuristics that were difficult to understand.*

No problems at all: Three out of eight participants mentioned that they didn't really have any problems with evaluations.

P4(developer): *It went ok nothing special to say*.

Problems with English: One participant mentioned that biggest problem was the English language.

P7(player): *I did my best, I don't know, I felt like I could do it but my English is not so good that I understand everything 100% so I had to improvise a bit. Language was the biggest problem for me. I didn't necessary understand the questions, otherwise everything was simple*.

In conclusion some participants had issues with either the games they played, the heuristics or English language and some participants didn't have any issues.

Q2: **How did you do the evaluations?**

The themes found and their amounts are shown in table 22.

Table 22: Themes for Q2 and the number of finds (N=8).

Theme	**Number of finds**
First task 1 and after that task 2	8
Going through heuristics in order	6
Occasional playing and filling	6
Doing Task 2 from memory	4
Filling the document while ads are playing	4
Game specific task first	3
Looking at Task 1 answers when filling task 2	3
Not playing anymore after some point	2

First task 1 and after that task 2: All the participants mentioned that in the evaluations they first did task 1 and after that task 2.

P1(developer*): I didn't realize that you could do it in a different order. Task 1 first. Same tactics both games*

P2(player*): task 1 first until it ends, first I went through the game, played the game specific task and went to the shop and looked it through and started task 1 one by one in order and after that task 2.*

P5(developer*): first task 1 and then task 2,*

Going through heuristics in order: Six out of eight participants mentioned they went through the heuristics one by one in order while doing the evaluations.

P4(developer*): I did it in order, one task at a time, I played it, took 2-3 heuristics, and played the game and filled.*

P1(developer*): I took the first one and played a little. As soon as I got an idea, I took the next one. With this tactic from start to finish. In order 1 2 3 4*

Occasional playing and filling: Six out of eight participants mentioned that they were switching between playing and filling occasionally.

P5(developer*): task 1 After the first game, I pretty much completed all of the game specific tasks, there were a few points that I had to play a couple more games.*

P7(player*): I played for a while at the beginning and looked at the questions is there something I can answer. If there was something that I need more information, I stopped there and continued playing.*

Doing task 2 from memory: Four out of eight participants mentioned that they did the second task with more of a feeling instead of playing anymore because all the relevant stuff was already in their mind.

P3(player*): for task 2, I didn't play anymore, I mostly felt like I had just gone through the same things.*

P4(developer*): I did task 2 pretty quickly without playing anymore because I already had everything in my memory.*

Filling the document while ads are playing: Four out of eight participants mentioned that they answered/filled the form while ads were playing in the game.

P5(developer*): I filled up the document when I got to play, and I filled it up when watching ads,*

Game specific task first: Three out eight participants mentioned that they did the game specific task before anything else.

P6(player*): First I played the 5 levels, so I know what the game is about, after that I started to answer the questions*

Looking at Task 1 answers when filling task 2: Three out of eight participants mentioned that they looked at the answers they had made to task 1 when filling task 2.

P5(developer*): when I got task 1 ready, started task 2 right away, when playing the second game had to look my answers from task 1 so I could give a realistic answer*

Not playing anymore after some point: Two out of eight participants mentioned that at some point of the evaluations they didn't play the games anymore because of different reasons such as having everything in memory or previous playing experience.

P4(developer*): I did task 2 pretty quickly without playing anymore because I already had everything in my memory*

P6(player*): I have played 8 ball for about 40 hours, I knew the game pretty well so it made it easier to answer and I honestly didn't play any more after the first game, I went a couple of times to try out the sound it makes when I press something, etc. pretty much did it with this tactic but I knew I wouldn't have any big problems with it*

In conclusion all participants first did task and then task 2. Majority of the participants also went through the heuristics in order and occasionally played the game and filled the evaluation form. Some of the participants also filled the documents while there were ads, and some participants did task 2 out of memory.

Q3: What did you think about the games you played?

The themes found and their amounts are shown in table 23.

Table 23: Themes for Q3 and the number of finds (N=8).

Theme	Number of finds
Gem stack is a bad game	8
8 Ball Pool is a good game	5
Something to improve in 8 Ball Pool	5
Something good in Gem stack	2

Gem stack is a bad game: All the participants said that they didn't like the game "Gem stack".

P7(player*): there are no instructions, I don't know what to do, there are no tutorials, you get to the finish line, you don't know what to do, there are a lot of ads, I can't take it, my head hurts, I can't take it, no audio, no music, I went to the store, no no no no no, nothing good to say, amateur trash game, 1 out of 10*

P8(developer*): The first incredibly annoying, too simple controls, no clear goal, an absurd number of ads 1 out of 10*

8 Ball Pool is a good game: Five out of eight participants mentioned that they liked the game “8 Ball Pool.”

P8(developer*): The second game, clear and simple, also fun, clear instructions tell you all the time if you hit the wrong ball, it was quite clear to get to the store and the menus. If not 10 out of 10 then I would give 9 out of 10*

Something to improve in 8 Ball Pool: Five out of eight participants mentioned there are some things that should be improved in 8 Ball Pool such as new designs or gameplay.

P1(developer*): 8 ball: feels like a pretty old-fashioned game. Pretty basic billiards, but there are so many games on the market that bring a more pleasant experience. At this stage, it took some kind of crazy mode, so you wouldn't even have any boosters. I'm not a huge fan, its nice billiards, but it won't last long 5 out of 10.*

P6(player*): The game has a lot of crazy things. Its pay-to-win, there are cheaters, The gameplay should be a little different, it's stupid how easy it is to make a perfect shot, it would be good if you had pendulum going right to left so you could fail your shot, at the moment you can't fail your shots, but this design is great for casual players. 7 out of 10*

Something good in Gem stack: Even though all the participants didn’t like Gem stack, two out of eight participants said there is something good in the game.

P1(developer*): stylish but wouldn’t play for long.*

P4(developer*): Pretty trash game but I kind of like the idea of it*

In conclusion Gem Stack was considered to be a bad game and 8 Ball Pool a good game.

Q4: Were you able to use the heuristics successfully when evaluating the games?

The themes found and their amounts are shown in table 24.

Table 24: Themes for Q4 and the number of finds (N=8).

Theme	Number of finds
For the most parts	4
Few parts were confusing	4
No problems at all	2

For the most parts: Four out of eight participants mentioned that they could mostly use the heuristics successfully.

P2(player*): for the most part yes, in my opinion they were quite clear, and they have remained well used in these mobile games.*

P6(player*): Mostly yeah*

Few parts were confusing: Four out of eight participants mentioned that some heuristics were a bit confusing. These specific heuristics were covered in question 9.

P5(developer*): I had to adapt with some of the heuristics.*

P8(developer*): There were a couple I didn't understand.*

No problems at all: Two out of eight participants mentioned that they didn't have any problems.

P7(player*): Yeah, I could. Just put the numbers from 0-4 simple as that no problems*

In conclusion the participants could use the heuristics overall successfully with some parts of the heuristics being confusing

Q5: Are these heuristics easy to understand (in general)

The themes found and their amounts are shown in table 25.

Table 25: Themes for Q5 and the number of finds (N=8).

Theme	Number of finds
Most of them yes	5
Couple of confusing ones	3
Problems with the language	2
Bad heuristics	1
Heuristics as a word not familiar	1

Most of them yes: Five out of eight said that most of the heuristics are easy to understand.

P1(developer*): Yeah, most of them are.*

P4(developer*): Yeah*

P6(player*): Yeah mostly, couple of difficult ones*

Couple of confusing ones: Three out of eight participants mentioned that there were couple of confusing heuristics.

P2(player*): Yeah, most of them were easy but there were some, like three or four that I had to read a couple of times to understand them.*

P6(player*): Yeah mostly, couple of difficult ones*

Problems with the language: Two out of eight participants mentioned that they had some trouble with the English language.

P6(player*): The problems I had might have been because I am not really perfect with English language.*

P7(player*): The heuristics were not really that easy because I think there is a lot of difficult English which was the biggest problem for me.*

Bad heuristics: One of the participants said they think the heuristics are bad because there are too many questions, and the heuristics should be more detailed.

P3(player*): I did ok, I don't think the heuristics were very good, a bit strange and the heuristics have too many questions, it should be detailed and more separated. There was one heuristic about sounds, vibration, and visual effects, it might be better to be separated to individual ones. Some of the heuristics I had trouble saying anything.*

Heuristics as a word not familiar: One of the participants mentioned that the word heuristic can be a bit confusing.

P7(player*): Heuristics as a word is not that familiar and I think it could confuse some people.*

In conclusion most of the heuristics were considered easy to understand but some were confusing. Some participants also had trouble with language. One participant considered the heuristics to be bad.

Q6: What are the strengths of these heuristics?

The themes found and their amounts are shown in table 26.

Table 26: Themes for Q6 and the number of finds (N=8).

Theme	**Number of finds**
Good for basic and necessary things	6
Very broad list	2
Fast to use	2
Don’t know what to say	1
The ability to comment to problems important	1

Good for basic things: Six out of eight participants mentioned that the strength of the heuristic list is that it is good for the basic and most necessary things.

P1(developer*): A good list that can quickly cover the basic things. It's a good thing to see that no catastrophic things arise from there.*

P2(player*): Can clearly see what the game's biggest problems are.*

P3(player*): The list is not super narrow, there are different kinds of questions, if there's a big problem, you'll find it, there are lots of basic important questions*

P6(player*): It would make development easier if you could get even a couple of people to fill it, you'd know where to focus more, you could find easy fixes, I believe that if you make a game and get 10 people to fil it out and the developer goes through it, you can get a crappy game to be better*

P8(developer*): With the list you can keep the game clear and visually consistent and ensure smooth gameplay.*

Very broad list: Two participants mentioned that the strength of the heuristics list is that it is broad.

P3(player*): The list is not super narrow, there are different kinds of questions, if there's a big problem, you'll find it, there are lots of basic important questions*

P4(developer*): The list of heuristics is pretty broad in my opinion.*

Fast to use: Two participants mentioned that you can get a view of a game very quickly.

P1(developer*): A good list that can quickly cover the basic things. It's a good thing to see that no catastrophic things arise from there.*

P7(player*): you can find out quickly what does someone think of the game by filling out the form.*

Don’t know what to say: One of the participants couldn’t say any strengths about the heuristics.

P5(developer*): I don’t know what to think. Really don’t know what to say about this.*

The ability to comment to problems important: One of the participants mentioned that the strength of the heuristics list was the ability to explain what was wrong with something.

P3(player*): I think the additional field next to the slot where you put the value is really good. If it was not there it would be really bad. Now I can specify what I actually mean with the number I put there. For example, I could specify that the problem was with sound and not the visual when there was a heuristic that had both of them in the same one. Without the commenting field it would be unclear where is the problem*

In conclusion majority of the participants considered the main strength of the heuristics to be that they were good for basic and necessary things. Other strengths were that the list of heuristics was broad, and it was fast to use.

Q7: What are the limitations of these heuristics?

The themes found and their amounts are shown in table 27.

Table 27: Themes for Q7 and the number of finds (N=8).

Theme	Number of finds
Too general / broad	4
Too many things in one heuristic	2
Doesn’t take the user in to account	1
Unclear what to answer	1
Problems with English	1
Needs additions about interruptions	1

Too general / broad: Four out of eight participants mentioned that the list of heuristics is too general or broad.

P1(developer*): Can't check everything. Maybe some heuristics are better for a certain genre. You can't generally use these for every game either.*

P2(player*): there may be some things that are not necessarily taken into account, which may be related to the game in a surprisingly large area.*

P7(player*): I think 25 heuristics is too much. I think less would be better.*

Too many things in one heuristic: Two participants mentioned that they think there are too many specific points in one heuristic.

P3(player*): For example, heuristic 14. it could have been divided into multiple individual ones. Also, heuristic 20 I think there is too much in the same one.*

P7(player*): I think some of the heuristics have too many points in them.*

Doesn’t take the user in to account: One participant mentioned that the heuristics don’t consider the evaluator and the evaluators lack of ability to do something.

P5(developer*): not everyone has the opportunity to use vibration, in a few questions you had to turn on sounds or vibrations or something similar.*

Unclear what to answer: One participant mentioned that in some points it is unclear what to answer.

P6(player*): It is not clear what to answer if the game has not been implemented something, for example very few games react to what is happening in the environment, you can't really give the best value or the worst.*

Problems with English: One participant mentioned that the biggest issue is the English language.

P7(player*): I don’t really understand English that good. Don’t really understand all of the questions 100% so I had to adapt and improvise a bit.*

Needs additions about interruptions: One participant mentioned that the list of heuristics needs some additions. The participant felt like the should be a heuristic that focuses on interruptions.

P8(developer*): I haven't had much to do with heuristics, if I would add something, it would be about functionality, is there anything that interrupts the concentration of the game session, are there easy ways to get out of the interruption, it was not directly mentioned anything that interrupts the game session like long advertisements,*

In conclusion the main limitation was considered to be the fact that the heuristics were too general and broad. Other limitations were for example that there were too many things in one heuristic, the heuristics don’t take the user into account and it is unclear to answer to some points.

Q8: Which heuristics are the most useful?

The themes found and their amounts are shown in table 28.

Table 28: Themes for Q8 and the number of finds (N=8).

Theme	Number of finds
Heuristics related to navigation	5
Heuristics related to tutorials	5
Heuristics related to interface and easiness of use	4
Heuristics related to the most relevant information	4
Heuristics related to starting playing quickly	3
Kind of all of them	2
Heuristics related to controls	2
Heuristics related to sounds	1
Heuristics related to communication	1

Heuristics related to navigation: Five out of eight participants mentioned that the most useful heuristics were related to navigation.

P1(developer*): The most important are those related to navigation and that the icons are comprehensible, that the work between the UI side and the game scene during gameplay is smooth, everything related to that. So mostly those that focus on menus.*

P6(player*): Heuristic number 4. The usability in so many games is so shit nowadays. Navigation is also a really broad concept, so it is really good to get feedback about it.*

Heuristics related to tutorials: Five participants mentioned that heuristics related to tutorials are the most useful. The participants said it is important to know how to play the game and there should be some sort of help to return to if necessary.

P2(player*): Heuristic number 8. every game must have clear instructions and a tutorial to return to, there must be some sort of help so that the player can return to if he does not play the game for a while.*

P6(player*): Heuristic number 1. With this you know do you have to make a tutorial to the game. It would be good to not have to make it.*

Heuristics related to interface and easiness of use: Four out of eight participants said that heuristics that focus on the interface and how easy it is to use are the important.

P1(developer*): The most important are those related to navigation and that the icons are comprehensible, that the work between the UI side and the game scene during gameplay is smooth, everything related to that. So mostly those that focus on menus.*

P2(player*): Heuristic number 19. The interface has to be simple, so you can use it and you can understand everything, It can't be difficult so you start wondering where this and that is. It should be very practical, even if you have trouble understanding some text for example, you can still play the game with pictures, so you don't need language skills.*

Heuristics related to the most relevant information: Four participants mentioned that heuristics that focus on the most "relevant" things in their opinion are the most important.

P4(developer*): I think the most important things are the ones that handle the basic things. Like seeing your score or having an easy-to-understand interface. If the basic things are not there, the game is basically unplayable.*

P7(player*): Heuristics 1, 3, 8, 2, 4. They are all these basic things. I would like to know is the game easy to start it is important. Then some basic things like moving in the game and understanding what the goal is and the point of the game, what are the controls of the game, where do you find stuff. I think these are the most important.*

Heuristics related to starting playing quickly: Three participants mentioned that the heuristic related to starting to play quickly is important because otherwise you might stop playing.

P3(player*): Heuristic number 1. If you don't get to start playing withing 30 seconds especially in mobile games, you delete the game*

Kind of all of them: Two participants mentioned that they think all the heuristics are useful.

P5(developer*): I think kind of everything is a bit useful, everything needs to be answered, every one of them is a viable question, everyone helps in developing the game forward,*

P8(developer*): I think the list is very good and I don't think there are any really bad or useless ones in there.*

Heuristics related to controls: Two participants mentioned that heuristics related to controls are the most important because they are part of the basic things that need to be good and easy to use to play the game.

P7(player*): Heuristics 1, 3, 8, 2, 4. They are all these basic things. I would like to know is the game easy to start it is important. Then some basic things like moving in the game and understanding what the goal is and the point of the game, what are the controls of the game, where do you find stuff. I think these are the most important.*

Heuristics related to sounds: One participant mentioned that heuristic that focuses on sounds is one of the most useful ones because it is frustrating to play games where the sounds don't work.

P6(player*): Heuristic number 10. I have noticed a lot of problems in games when it comes to sounds. Usually, they are really bad or non-existent. I finally get to buy the final upgrade in some game and there is no sound effect when I get it. Its frustrating. I don't get the dopamine I should get.*

Heuristics related to communication: One participant mentioned that the heuristics related to communication are important because playing with friends is more fun.

P2(player*): Heuristic number 13. Playing is more fun when there are friends involved. You can play with them and also play against all the people in the world. It brings variety to playing. Also when you play against other people you are more motivated because you are competing against someone. There is the element of competition involved. In single player games you get more easily bored*

In conclusion the heuristics that were related to navigation and tutorials were considered the most useful. Heuristics related to interface and easiness of use and the most relevant information were also considered relevant. There were also other relevant heuristics such as the ones related to starting playing quickly and controls.

Q9: Which heuristics are difficult to understand?

The themes found and their amounts are shown in table 29.

Table 29: Themes for Q9 and the number of finds (N=8).

Theme	Number of finds
Heuristic #23	6
Heuristic #25	5
Heuristic #12	5
Heuristic #24	4
Heuristic #21	3
Heuristic #14	2
Heuristic #9	1
Heuristic #10	1
Heuristic #15	1

Heuristic #23: Six out of eight participants had difficulties with heuristic 23 because they didn't understand what it meant.

P2(player*): I didn't remember what UI meant first so I had some problems with the question.*

P3(player*): I didn't really know what is the problem with this. What does it meant if they are not used for their own purposes. Does it meant that you would use the side buttons on your phone in the game?*

P5(developer*): What does it mean. How does my device UI have something to do with the game UI.*

P7(player*): The question was hard to understand. Mostly the English was difficult for me. Like game ui and game ui are used for their own purposes?! I probably should have asked about this. What the fuck is UI. The question was just hard to understand.*

P8(developer*): I had no idea what this meant. Very difficult to understand.*

Heuristic #25: Five out of eight participants mentioned that heuristic number 25 was difficult because they didn't know what it meant or how to test it.

P1(developer*): Didn't really understand what it meant and how to test it.*

P2(player*): It was very unclear what it meant. Maybe it should be explained a bit better so it would have made more sense. Also as it was in English as a Finn you could understand it a bit differently.*

P3(player*): The game does not affect the environment in any way so pretty hard to answer anything to this.*

P4(developer*): I understood what it meant but how am I supposed to test it.*

P8(developer): On the first game it was easy to answer because it didn't have any kind of sounds. But with the second game I couldn't answer anything. I didn't know how it could affect the environment. It was a hard question. No idea what it meant.

Heuristic #12: Five out of eight participants mentioned that they thought heuristic number 12 is difficult because they didn't know what it meant or how to test it.

P1(developer*): I just didn't understand what the question was looking for and what it meant. How do I test it?*

P3(player*): I didn't really understand is it a good thing or not that the menu is part of a game. Is it a bad thing if its not. The question is weird I don't know what it meant.*

P5(developer*): I have no idea what it meant exactly. Does it mean that the menu is taken out of the examination or should it be taken out or not. I didn't understand. This could have been maybe be said in a different way like is the menu a part of the games theme or something*

P6(player*): The question sounded pretty rough. No idea what it meant. What even is the question, Its not in question form.*

P7(player*): Does it mean I have to experience the menu meaning I have to go there?. I don't understand about these worlds that good but why do you even have to ask this. It is just unclear.*

Heuristic #24: Four out of eight participants had difficulties with heuristic 24 because they didn't know what it meant or how to test it.

P1(developer*): I didn't understand what it meant and how should I test it.*

P3(player*): If you don't have any interruptions is pretty hard to test this.*

P4(developer*): Didn't really know what this means. How am I supposed to test this.*

P6(player*): Hard to say that interruptions mean. I think it mostly meant the ads in games. This should be more clear.*

Heuristic #21: Three participants said that they had difficulties with heuristic 21 because it is a vague and hard questions and there are so many different trends.

P2(player*): You can have many different opinions about this. What kind of trends? There are a huge amount of them in different kind of gaming communities. There are lots of different game formats and communities. They are in totally different words compared to each other. Their trends are different. Like for example mobile games should be simple but PC games have totally different important things.*

P4(developer*): This question is a bit vague in my opinion. I could say something to this but I imagine some people might have problems with this especially if you don't know about games or the trends that good.*

P6(player*): It’s a difficult question to answer to. I can’t really make it better anyway but its just a rough question to answer to. I had to think about much more compared to the other questions.*

Heuristic #14: Two participants said that heuristic number 14 is difficult because it should be separated into two and it was just unnecessary.

P3(player*): It has audio, visual and haptic in the same one. In the other game I played it had no sounds and if one of the things is wrong, I have to give it a bad score, so you think that everything is bad when it was only the sound. They should be separate.*

P7(player*): The question was in a way interesting and I kind of get it but why do you have to ask this. I think you could just delete this one.*

Heuristic #9: One participant mentioned that heuristic number nine was difficult to answer to because there were no errors.

P5(developer*): There was no error or anything at any point, so it was difficult to answer to this*

Heuristic #10: One participant mentioned that heuristic number 10 was difficult.

P6(player*): Number 10 is difficult because you don’t give feedback with numbers.*

Heuristic #15: One participant mentioned they had difficulties with heuristic 15 because the language was hard to understand.

P7(player*): I think the word icons is interesting. Could there be a bit easier word for it.*

In conclusion heuristics number 12, 23, 24 and 25 were considered difficult by the majority of the participants because these heuristics were confusing, or the participants didn’t know how to test them.

Q10: Does the heuristics cover all usability problems in these games you evaluated?

The themes found and their amounts are shown in table 30.

Table 30: Themes for Q10 and the number of finds (N=8).

Theme	Number of finds
Something should be added about ads	3
Heuristic for sounds	1
Heuristic about games idea	1
The game was lagging	1
Missing from gameplay	1
Bugs in the game	1
Everything is covered	1

Something should be added about ads: Three out of eight participants there should be something in the heuristics about ads. The participants felt like the heuristics didn't have anything about ads which were very disturbing.

P5(developer*): I would add something about ads. Like is the game fucking possessed with ads and made unplayable because of them?*

P8(developer*): There should be something for the most severe interruptions like the ads that totally break the game session and playing.*

Heuristic for sounds: One participant mentioned that there should be some sort of heuristic regarding sounds.

P1(developer*): There could be a question about the sounds of the games and are they in sync and in balance. Are the sounds appropriate for the theme of the game.*

Heuristic about games idea: One participant mentioned that there should be a heuristic about the idea of the game, so you know what the game is about.

P2(player*): In Gem stack not really. You should know the overall idea of the game. In 8 Ball Pool the idea is clear but in Gem stack they don't tell anything about the games' idea. What is your mission? Where do I try to end up?*

The game was lagging: One participant mentioned they had a problem which the heuristics didn't cover which was lagging.

P4(developer*): Not really because the other game was lagging on my phone.*

Missing from gameplay: One participant mentioned that the heuristics should have something for gameplay.

P1(developer*): No they don't. There are quite a lot missing when it comes to gameplay but it is pretty hard to make those kind of heuristics because there are so many different kind of games. Cannot say anything specific examples at the moment about gameplay but there are so many things when it comes to games and 25 heuristics are just not enough*

Bugs in the game: One participant mentioned that there were bugs in the game.

P6(player*): There were some visual bugs with Gem stack. I had this one bug where there was an ad, and it filled the whole screen and I could not click it away so had to restart the game. When I got the bug, I had no idea in what heuristic I should add it to.*

Everything is covered: One out of eight participants said that they think the heuristics cover everything.

P7(player*): Well, I thought that the interruptions meant the ads which were annoying so then I think everything is pretty much handled.*

In conclusion majority of the participants thought that there was something that the heuristics did not cover such as sounds, game idea or bugs. Many participants said that something about the ads should be added.

Q11: Do you think these heuristics would be useful in game development? Why? Why not?

The themes found and their amounts are shown in table 31.

Table 31: Themes for Q11 and the number of finds (N=8).

Theme	**Number of finds**
Takes care of the basic things	4
Depends on the person using	2
You are able to notice the development stage of the game	1
Should be used in the end phases of development	1
Awakes questions	1

Takes care of the basic things: Four out of eight participants said that they would be useful because it will handle all the basic things.

P1(developer*): Yes, I think they would be really good because there were some simple easy fixes that would have made the game better. For example, in Gem stack with sounds and vibrations. They could have been used when pressing a button or something. Would have made the game better.*

P4(developer*): I am sure they would be useful. The list is very common in a way that it looks at all the basic things in a game. You could immediately see if something basic is not really good.*

P6(player*): They would help in development if you could even a couple of people to fill them. You would now on what parts of the game to focus on. You would find easy fixes. I believe that if you make a game and 10 people fills the heuristics and the developers goes through them they could make a "bad" game much better*

P8(developer*): Yeah definitely. You will have kind of checklist that says are the basic and most important things in order.*

Depends on the person using: Two participants mentioned that the usefulness of the heuristics depends on the person.

P3(player*): If the heuristics are filled by a so called "normal person" then I think they wouldn't be so useful. They should be more simple and the questions should be more leading for the average joe. They should be more specific "is something good or bad". Now I have to think about problems too broadly. Am I able to find some problem about something. The questions should be strict about some certain issue or topic. If the heuristic are used by more of a gamer person then it would be a bit better.*

P7(player*): I am sure they would be useful to someone smarter who actually develops games but for average guy like me no use at all.*

You are able to notice the development stage of the game: One participant mentioned that with the heuristics you could see in what development stage it is.

P5(developer*): Yes, I think they would. The second game I played Gem stack felt like its in alpha state made in two weeks. The other game 8 Ball Pool felt like a ready-made game. From the evaluations I could see what are the stages of the games.*

Should be used in the end phases of development: One participant mentioned that it would be better to use towards the end of development because the early development is already slow and in the end, you have a version of the game that is more suitable for testing

P1(developer*): I don't think you need to start doing these kind of evaluations too early because making the game is slow as it is. Maybe these are more just to keep in mind in the design phase. Towards the end when you have some sort of MVP of the game might these be useful.*

Awakes questions: One participant mentioned that the heuristic brings out questions about problems and solutions.

P2(player*): Yeah, I think it's quite clear that they would be useful. With both of the games I could say what is good and what is bad. The heuristics awakes question and different kinds of solutions to problems, so they are a good development tool*

In conclusion the participants thought that the heuristics would be useful because they take care of the most basic things. Some participants also said that the usefulness depends on the person using the heuristics.

Q12: (Game developers) How useful do you think these tasks would be if you could get an evaluation filled out by the players about your own game?

The themes found and their amounts are shown in table 32.

Table 32: Themes for Q12 and the number of finds (N=8).

Theme	Number of finds
Depends on the answer	4
Really useful	2
Task 1 is more important compared to Task 2	2
Both tasks are good	1

Depends on the answer: All the developers mentioned that the usefulness of the tasks depends on who fills out the evaluation forms, how good, how many people and how good are the forms filled.

P1(developer*): If a singular player would do it, I would consider the evaluation be totally useless. There should be multiple players who fill them. Something like 100*

P4(developer*): It depends on the quality of the comments. If there is only a number but no comment on why something is good or bad I can't really get anything from it. It is also very relevant what kind of player fills the form and how many people do the evaluations. How good have they answered? If they have answered "good" then I think you could get something from them in the development phase.*

P5(developer*): If the data I would get as developer would be this comprehensive and people would actually spend time filling them probably, this would be the best kind of info you can get as a developer.*

P8(developer*): If the answer is just that something is good then it's not so interesting but if the answers actually say something is wrong or irritating and why they are like that then that information is very good.*

Really useful: Two out of four game developers said that the tasks would be really useful because all kind of feedback is important, and the tasks give quick information on what is wrong.

P5(developer*): Yeah, totally they would be useful if I had to choose I would take task 1 because you can comment on that but I would rather take both. Any kind of feedback is extremely important.*

P8(developer*): Yeah, they would be really useful. They would give me a quick info on what is wrong in my game and in what parts do the players focus on.*

Task 1 is more important compared to task 2: Two game developers said that the information of the first task would be better for them compared to the second task because you can comment on it, and it gives direct answers to what is bothering the player

P5(developer*): Yeah, totally they would be useful if I had to choose I would take task 1 because you can comment on that but I would rather take both. Any kind of feedback is extremely important.*

P8(developer*): Definitely Task 1 is more important than task 2. Task 2 is more of an grading scale. But if the questions from task 1 come back to me actually filled I would get a clear and direct answer to what is bothering the player.*

Both tasks are good: One game developer said both of the tasks would be good information.

P5(developer*): Yeah, totally they would be useful if I had to choose I would take task 1 because you can comment on that but I would rather take both. Any kind of feedback is extremely important. If I would get only task 2 and not task 1, I think it would still be very useful. I think both of the tasks are useful because they kind of asks the different thing, so they answer to different questions. In task 1 you kind of ask is the game even playable and in task 2 you are asked how does the game feel for you. So maybe in task 2 I answer more about my preferences and what do I feel compared to other games. Task 1 is just directly asking is something working.*

In conclusion the developers said that the usefulness depends on the answers. Some developers said that the tasks are really useful and some also said that they considered task 1 to be more important compared to task 2.

6. DISCUSSION

In this section I will analyse the differences between players and game developers. I will also analyse the differences between all players and the differences between all game developers. In the differences I will analyse all the results presented in chapter 5 along with other findings from the evaluation sessions. I will also discuss possible future work improvements and the limitations of this study.

6.1 Differences between players and game developers

In this subsection I will go through the differences between players and game developers. First, I will go through the different findings for all the participants about the evaluations. Then I will analyse the results from chapter 5 which includes severity rankings and fulfillments. With these analyses I aim to answer to answer **R**esearch **Q**uestion 1: **What are the differences in results when using heuristics to evaluate mobile games between game developers and players**?

6.1.1 Differences in evaluations

Overall, all the participants in the study could complete the sessions successfully and in the authors opinion without any bigger issues. When it comes to the times that the participants spend evaluating the games seen in table 8 in the chapter 5, there were big differences between all the participants. All the participants were able to complete the evaluation under the given time of 60 minutes. Participants 4 and 7 surprised me with how fast they were in their evaluations compared to the other participants. There were some factors that could explain these time differences which I will discuss later in this section.

When looking at all the evaluation sessions and how the participants differentiated in their answers and methods there were significant differences. When looking at task 1 of the evaluations and focusing on how the participants commented and answered on this task, the results were very different from each other. When going through the tasks and the evaluation form in the start of the evaluation sessions with the participants, the instructions given by the author in task 1 was that in the column of description of the problem / comments, it is advisable to write why they chose the severity value.

With this instruction in mind, the participants answered very differently. Participant 1 answered in English, and he wrote something to almost every one of the heuristics. The descriptions the participant wrote were very general in all the cases. Just very shortly explaining what was wrong. Participant 2 also answered in English but compared to participant 1 he answered to all the heuristics. Even if something was totally fine and the severity was 0 the participant answered for example “This is good”. So, he answered to all the good things also with some sort of little positive sentences/words. Participant 2 spend the most time doing the evaluations so he might have just wanted to be very careful and thorough with the evaluations which might explain why he commented to even the positive ones.

Participant 3 wrote to almost every heuristic with general comments like participant 1 but interestingly participant 3 wrote everything in Finnish. This happened also with participant 4. He also wrote general comments to almost every heuristic in Finnish. Participant 5 wrote no comments at all to any of the heuristics when it came to 8 Ball Pool. However, this could be explained with the fact that participant 5 didn’t rank any of the heuristics as a bigger problem than cosmetic one. With Gem Stack participant 5 answerer in English to about half of the heuristics and the comments were general. Participant 6 differentiated from the others the most in the authors opinion. Participant 6 wrote clear suggestions for the games that would fix the problems with the heuristics and would make the games better. With 8 Ball Pool participant 6 wrote to less than half of the heuristics in English. With Gem Stack he wrote to all the heuristics. Participant 7 wrote general comments in English, but he wrote to very few of the heuristics compared to other participants. In 8 Ball Pool he only wrote to two heuristics. In Gem Stack he wrote to six heuristics. Participant 8 was like participant 2. He answered in English and at least made some sort of comment to all the positive ones also. Participant 8 however was the only one who had heuristics where he didn’t answer anything. No severity ratings at all.

Why the participants differentiated in the language they used is difficult to analyze. In the interview results participants 6 and 7 mentioned they had some problems with English, but they still wrote in English, so it is very hard to say why participants 3 and 4 answered in Finnish. This might have been just because of the combination of evaluation session instruction. All the tasks, questions, heuristics, and instructions were in English but the discussion with the participants and the author in the sessions were all done in Finnish so it might have affected which language the participants used. All the participants were instructed that the language they use is up to them. There are no clear differences between players and game developers in the ways how the participants answered. The only significant differences were that one of the players gave improve recommendations and no single game developer did this. The other difference being that participant 8 who was a game developer had heuristics where he didn't answer anything. All players answered something to every heuristic. The fact that participant 8 had heuristics where he didn't answer anything when all the other participants answered to every single one agrees to the finding made by Nielsen & Molich to some degree. Nielsen & Molich mentioned that the skill level of the evaluator has a significant difference, but even good evaluators can overlook easy problems [30]. Participant 8 cannot be considered as a good evaluator in this case, but he can be considered as a "Double specialist" [4]. He had experience from heuristics, and he had developed mobile games. With this experience he still had heuristics that were empty when everyone else had something. It cannot be said that these issues are easy problems to be found but when there were participants with considerably less experience with games and heuristics who had found the problems, I believe there is some sort of similar finding to [30].

There is no clear reason why participant 8 also answered to the heuristics that were not an issue. It might have been because he was the only one with experience with heuristics form working life also, but this is total speculation. The fact that participant 6 wrote suggestions that would make the game better was very interesting for the author because participant 6 was a player and not a game developer. However, when looking at the background questionnaire that the participants filled, there are factors that might explain this. Participant 6 categorized himself as a hardcore/Expert video game player who played pc and mobile games. He also played lot more different game genres compared to any other participants. Participant 3 also categorized himself as a hardcore/expert player, but he only played pc games when participant 6 said he played pc and mobile games. This status of hardcore/expert player especially with mobile games might explain why he wrote suggestions. It is clear he knew a lot about mobile games and could say what would make them better. This finding seems to be aligned with the results of the study made by Nielsen where the people with expertise of the type of user interface that is being evaluated helps in the evaluation [4]. Why participant 7 wrote so little comments might be explained by the fact that he did the evaluations the fastest. Why he did them so fast is a bit harder to say. However, the background questionnaire answers give some insight to this. He categorized himself as a casual gamer who only plays console games and only sports games. Compared to all the other participants it was clear participant 7 was the one with least experience with games. This might explain why he did the evaluations so fast. Maybe his experience with games was not good enough to give him deeper thoughts or ideas about the heuristics and games he played so he just did very fast and basic evaluations.

In the evaluations forms regarding task 1 there was a free space for other issues/comments. The instruction given about this space was that if the participants find a problem that don't break any of the heuristics in the actual list of heuristics but in the participants' minds it is clearly a problem, they should write the problem there. The usage of this free space also differentiated between the participants. Participant 1 wrote nothing to this space. Participant 2 wrote something about both games. The comments for both games were:

- *8 Ball Pool: Having more training tips and tricks inside the game, so it's much easier to learn and play better.*
- *Gem Stack: The commercials and instructions are the biggest problem. Commercials make the game miserable, and the game does not even explain what is the point of playing the game.*

Participant 3 also wrote something about both games in Finnish.

- *8 Ball Pool: When hitting the ball, the player is shown a prediction line that shows where the ball you are hitting will go. The line is quite short, so you would have to use a piece of paper or something straight on top of the screen to see more accurately beyond the ball's trajectory (translated from Finnish).*
- *Gem Stack: Ads and opportunities to watch ads for extra benefits are pushed everywhere and it makes the game a bit confusing (translated from Finnish).*

Participant 4 only wrote something about Gem Stack in Finnish

- *Gem Stack: The game lags quite a lot (translated from Finnish)*

Participant 5 also wrote only about Gem Stack

- *Gem Stack: This game looks like it's dev has taken max 30 days. Cashgrab. Awful game*

Participant 6 wrote nothing to the free space. Participant 7 wrote about Gem Stack.

- *Gem Stack: Shit game*

Participant 8 wrote nothing to this space. Most of the comments seemed to be about Gem Stack and the problems with it. Participant 4 mentioned as the only person that the game is lagging on their phone. Perhaps this was just because he had a slightly older phone than the other participants based on the interviews. When comparing game developers and players, players overall used this empty space more. Based on the comments they are in line with the findings of the study made by Santos et al. where reviews made by amateurs often use emotionally charged vocabulary [35]. This is line with the comments above. You can see this from looking at the comment made by participant 6 but overall, the fact that more players wrote to this field compared to developers and even the comments made by participants 2 and 3 are more extensive compared to the comments made by participants 4 and 5 who were developers.

6.1.2 Number of questions asked

The number of questions the participants asked in beginning of the sessions before starting the evaluations and during the evaluations differentiated. The participants were given the instructions to ask anything that is in their mind before starting the evaluations. They could also ask during the evaluations. However, the participants were reminded that the conductor of the study might not be able to answer depending on the question to not affect the results of the evaluations. Some of the participants asked questions before starting the evaluations. The questions the participants asked in beginning of the sessions are shown below:

Participant 1 (P1) (developer):

- "*Heuristic 12 I don't really understand this.*"
- "*Having trouble understanding heuristic number 21.*"
- "*Heuristic 24 what is interruption.*"
- "*What does heuristic 25 mean*?"

P2 (player):

- "*What does navigation mean*?"
- "*What about heuristic 9 can you explain that*?"
- "*Heuristic 21 what does it mean? What trends*?"

P3 (player):

- "*What about heuristic number 9? What if I don't get a situation where I need them, or they won't show*"
- "*Heuristic number 24 what is an interruption*?"

P6 (player):

- "*I didn't really understand heuristic number 9.*"
- "*In heuristic 10 do I need to give feedback or does the game give feedback*?"
- "*Heuristic 12 does this refer also to the shop?*"
- "*Heuristic 23 a bit weird*"
- "*What does accommodates mean*?"

P7 (player):

- "*What is navigation*?"
- "*What does provide users mean*?"
- "*What is an interface*?"
- "*Do I need to keep my sounds on*?"
- "*What is number 9*?"
- "*What is heuristic number 10?*"
- "*What does icons mean*?"
- "*What does consistency mean*?"

Participants 4, 5 and 8 did not ask anything before starting the sessions. The number of questions that the participants asked came as a surprise. I was expecting some questions about the heuristics or the whole evaluation session but especially the number of questions about heuristics was surprising. I will go through all the difficult heuristics in more detail in later section of this chapter when analysing the interview answers. I will now only discuss what kind of questions the participants asked in the beginning of the sessions. Participant 1 only asked about heuristics that were difficult to him which I will discuss later. Participant 2 asked about different heuristics, but he also asked about navigation and what does it mean. It might be because of the English language. It might also be because participant 2 categorized himself as a casual player so all the terms related to games might not be that known to him. Participant 3 also asked about heuristics. Participant 6 asked about heuristics, but he also asked a question about word in English. Participant 7 also asked about heuristics and multiple questions about English words. He also asked about using sounds in the game evaluations. This was an interesting question for the author. I presumed that all the participants will keep sounds on during the evaluations or at least would turn them on because there were sound related heuristics. But because it had to be asked, I maybe should have added instructions about it. This also agrees with the finding made by Salian et al. to some degree. In their study if the evaluator has low level experience on the topic it affects results and there can even be problems with the evaluation process [32]. Based on the background information presented in chapter 4.2 of this study, participant 7 had the least experience with games from all the participants and he had no experience from heuristics.

There were also some questions asked during the sessions. These questions were asked by the participants at totally random times when they were evaluating the games, in the way that the participants opened their microphone and asked something from the conductor of the study. The questions the participants asked during the evaluation sessions are shown below:

P2 (player) (During evaluation of the second game):

- "*What does device UI and game UI mean? (Heuristic 23*)"

P3 (player):

- "*Question about the heuristics related to sounds. I only get sounds during the ads.*"

P5 (developer):

- "*What do I do when I don't have the possibility to use sounds or vibrations? What am I supposed to answer*?"

P8 (developer) (Asked after the evaluation of the first game):

- "*I didn't really understand device UI and game UI. What is device UI what is game UI? Does it mean the menus where you can change audio and stuff*?"

Participants 1, 4, 6 and 7 did not ask any questions during the evaluations. Participants 2 and 8 both asked about heuristics number 23, but they only asked about it after evaluating the first game. This brings out a question how the confusion about the heuristic affected the evaluation of the first game. It is also hard to say why they didn't ask it immediately before starting the evaluations. The question asked by participant 3 was one of the questions the author of the study couldn't really give a proper answer to not affect the results. I had to answer very vaguely to the participant that he should answer what he feels like. Participant 5 asked a question that came as a surprise to me. I had presumed that everyone is able to use sounds or vibrations. The specific reason why the participant was not able to use sounds was not figured out. The participant did the evaluation with his phone so maybe the phones speakers were broken or possibly he was doing the evaluation in a place where he could not use sounds for example in a public place. When comparing the number of questions asked by the participants players asked a lot more questions compared to game developers.

6.1.3 Game specific tasks

When it comes to the participants and the game specific tasks that were given as instructions for both games all participants except participant 1 completed the game specific tasks successfully. With participant 1, in the evaluation session after the first game which was 8 Ball Pool, participant 1 informed the conductor of the study that he is done with the evaluation but in the very next sentence the participant said he will "visit the store really quickly". Based on this the participant evaluated 8 Ball Pool without going to the store which was part of the task. In a way he did complete the task because he visited the store in the final seconds before ending the evaluation of the first game, but the participant did not complete in the sense the author of this study intended. In the authors opinion all the features of the store and its user interface and usability might not have been properly examined and this could have affected the results of the evaluation to some degree. This finding verifies the comment made by Nielsen & Molich. In their study they mention that even usability experts are not perfect doing heuristic evaluations. Participant was not an usability expert in any way but considering this study he could be considered an expert of some degree because he was a game developer who had experience from mobile games and he had also had experience from heuristics. He could be considered as a "double specialist" that was defined by Nielsen [4]. Even though the participant could be considered as more of an expert he couldn't complete the evaluation in a proper way.

6.1.4 Amount of irregularities

When looking at the evaluations with the participants there were some clear findings that the author considers mistakes the participants made. The tasks 1 and 2 where pretty similar to some degree and the answers to these tasks should have some sort of connection. In task 1 the evaluator had to classify usability problems in the game using the heuristics found in chapter 3 of this study and a severity scale found in section 4.3. In task 2 the evaluator had to assign a score from 1 to 5 (1 being worst, 5 being best) to every single heuristic based on how well the game fulfilled each of them. For example, if in task 1 the participant marked the severity to some heuristic as 4: *Usability catastrophe-imperative to fix this before product can be released.* The participants should give a reasonably low number as a fulfillment (1-5) for the same heuristics in task 2 and the same game in which they earlier said it's a catastrophe. With some of the participants there was a clear error between the answers according to the author. For example, participant 2 gave the severity ranking 1 for heuristic number 17 in Gem Stack which meant it was only a cosmetic problem but the fulfillment for the same heuristic in task 2 was only 1 which meant the lowest fulfillment. In the authors opinion if the fulfillment was that bad in the participants opinion it also should have been a bigger problem other than being just a cosmetic one. There were similar examples with participants 1, 3, 7 and 8. One of the clearest problems was with participant 8 in the authors opinion. He had marked the severity ranking as 0 which meant that it is no problem at all but for the same heuristic the fulfillment was the lowest which was 1. In the authors opinion there is no logic between the answers and would have been fascinating to understand better why he answered like that but with the time requirements of the session it was not possible. However, it is worth mentioning that these kinds of examples were only a problem according to the author of this study. Someone else might consider these totally normal because in a sense these two tasks ask different things when compared to each other's. Like participant 5 mentioned in the interviews: *I think both of the tasks are useful because they kind of asks the different thing, so they answer to different questions. In task 1 you kind of ask is the game even playable and in task 2 you are asked how does the game feel for you. So maybe in task 2 I answer more about my preferences and what do I feel compared to other games. Task 1 is just directly asking is something working.* This comment is a good example that in task 1 the participants might not consider something to be a problem in a sense but in task 2 just didn't feel like it applies to the game they evaluated. Good example of this was found from the answers of participant 1. They marked the severity for heuristic number 25 with a rating of 0 and had comment saying: *Game does nothing to accommodate with the players surroundings.* In task 2 the fulfillment for the same heuristic was 1. This example shows that the participant might have thought that this is not working with the game they evaluated but they didn’t consider it to be a problem. When looking at the “mistakes” the participants made, players had more of the compared to game developers.

6.1.5 Answers to trap heuristics

In the chapter 4.1.1 of this thesis, I described the criteria for choosing the games. One of the criteria was: *There must be multiplayer elements in one of the games and the other game cannot have any multiplayer elements*. The criterion was chosen to add a trap to the evaluations with heuristics 13 and 17 which both focused on elements that exist in multiplayer games. The results showed that the participants thought very differently about these heuristics. The ratings are all only from the game Gem Stack because in the authors opinion the results from that game are the relevant ones for these “trap” heuristics. This was because 8 Ball Pool was the game with multiplayer elements, so these heuristics didn’t have any issues. Gem Stack however was a game that was totally single player, so the answers are more interesting for this game. The severity ratings from task 1 for heuristics 13 and 17 for all the participants are shown in table 33.

Table 33: Severity ratings for “trap” heuristics with Gem Stack.

Participant ID/Heuristic	ID 1	ID 2	ID 3	ID 4	ID 5	ID 6	ID 7	ID 8
Heuristic13	2	3	3	3	4	2	4	4
Heuristic17	0	1	1	3	4	2	4	4

The table shows that participants 1 and 6 didn’t really consider these heuristics to be big issues. Participants 2 and 3 thought that heuristic 13 was a problem but heuristic 17 was not. Other participants: 4, 5, 7 and 8 thought that both heuristics were problems.

There are some possible reasons why participants 1 and 6 didn’t consider these to be a problem. With participant 6 it might be because of his background as expert mobile game player which I mentioned before. With participant 1 it is much harder to say. He was a game developer who had focused on mobile games only but so was participant 4 but he thought they were a problem. Participant 1 however played more games compared to participant 4 and he also played mobile games where participant 4 didn't so this might explain why participant 1 didn't consider these heuristics to be that big of a problem. When it comes to participants 3 and 4 both mention in the comments for both heuristics that they are issues but only marked the severity rating of 1 to heuristic 17 and severity rating of 3 for heuristic 13. There is no apparent reason for this. The other participants who thought that both heuristics are a problem didn't have any specific common things that could explain the answers other than the fact that they didn't like the game Gem Stack. Participant 2 however also mentioned he didn't like the game but only ranked heuristic 17 as 1 so the disliking the game cannot be marked as the main reason. When comparing game developers and players based on these answers, more developers thought that heuristic 17 was a problem and only one player thought it's a problem.

6.1.6 Differences in severity rankings and fulfillments

Next, I will analyse the differences from the results between game developers and players presented in the chapter 5. When looking at severity rankings in chapter 5.1 regarding 8 Ball Pool between players and game developers the differences aren't that big. Both groups ranked the majority of heuristics as "No problem" and had the same pretty similar number of other severities. I think this is just because 8 Ball Pool is such a finalized and well-developed game so there just isn't that many errors the participant could even find.

When looking at Gem Stack and the different severity rankings between the two groups there were some differences. With game developers there were slightly more "No problem" heuristics than with players. This might be because game developers can "overlook" some things easier compared to players. Based on the interview answers all the participants did not like the game Gem Stack but still game developers had more "No problem" severities. This might be explained by the game developers having more experience in making games and being aware of trends and what works. However, compared to players, game developers had more "Catastrophe" heuristics. This could be explained with the same idea that game developers know better what works, and, in this case, what doesn't work so when they saw something they didn't like they considered it to be catastrophic. This can be seen also in table 19 with the average ranking of heuristic compliances. Both players and game developers had very similar rankings for both games. Both groups thought 8 Ball Pool complies with the heuristics and Gem Stack doesn't. However, game developers had slightly bigger rankings for both games. This could be explained by game developers being slightly more forgiving when it comes to games, as they have knowledge of both the business and development aspects of the games.

When looking at the fulfillments of heuristics, the results show that developers had bigger fulfillments overall with both games. I believe this is because developers are more forgiving than players and can overlook things as mentioned before.

6.1.7 Conclusion

To answer **R**esearch **Q**uestion 1: **What are the differences in results when using heuristics to evaluate mobile games between game developers and players**?

Players spend clearly more time doing the evaluations compared to game developers. Game developers had more "No problem" severities compared to players. Game developers also had more "Catastrophe" severities compared to players, but players had overall more problems meaning "Major" and "Catastrophe" severities combined. Game developers also had bigger overall fulfilments with both games compared to players. There were also other differences related to how they behaved during the evaluations. Players asked a lot more questions compared to game developers. Players also utilized the evaluation form better which could be seen for example as writing comments to the white spaces next to heuristics or writing something to the empty space for other issues. Players however had more irregularities in their answers compared to game developers. Interestingly these results were not aligned with the results of the study made by Nielsen where specialist where better at finding problems [4]. In this study the players found more problems with the games compared to game developers who could be considered more as the specialists. Overall, the backgrounds of the evaluators had great influence on the results as found in these studies [35] [33]

6.2 Interview answers to questions about evaluations

In this subsection I will analyse some of the interview answers that were related to the evaluation sessions and discuss possible reasons why the participants answered this way.

Based on the interviews and the answers to question number 1 participants had different feelings about the evaluations in their mind. Three of the participants had problems with a game and in this case Gem Stack. Mostly these problems were because of the ads in the game. Two of these participants who complained about this were players who did not have experience from mobile games. The third participant who complained about this was a game developer, but he had only made games for PC. I think this is a very interesting and relevant point for this study about the different gaming worlds. Game developers who had experience from mobile games didn’t mention anything and player participants who played more mobile games didn’t mention anything. This shows that for people who are not into mobile games, the number of ads can come as a surprise. But the people who are familiar with mobile games know this and developers might even accept this fact because it is the modern way to make money in mobile games.

Based on the interview answers to question number 2, the participants had different ways of doing the evaluations. All participants said they first did task 1 and after that task 2. I am not sure why all the participants did the tasks in this order because they were given the instruction to do the tasks in any order they wanted. Perhaps this was just because it is easy and natural to go through things starting from up and going down. Six out of eight participants said they went through the heuristics in order. Perhaps this can also be explained with the fact that its more natural to go through things in order. Interesting found from the interview answers was that two out of eight participants did not play the games anymore at some point. This was mentioned by participants 4 and 6. Participant 4 said he had everything in his memory, so he didn’t need to play the game anymore. Participant 6 said he didn’t play 8 Ball Pool anymore because he had played it for 40 hours before. In this scenario I believe it was acceptable for the participant 6 to not play the game and still give answers that could be considered relevant. With that many hours into one mobile game I believe he could very easily answer what he thought about the game. These comments made by participants 4 and 6 show that it helps in the evaluations when you have experience on the topic similar as in the study made by Nielsen [4]. Participant 4 had experience from mobile games and heuristics and participant 6 was very familiar with the game being evaluated.

Based on the answers to question 3 all the participants said they thought Gem Stack is a bad game in some way but the reason for this opinion differentiated between the participants. Participant 1 said there were quite a lot of ads and there was no sound. Participant 2 said that in the game you just collect diamonds, and it is a single player game, so he didn’t like it. Participant 3 said he didn’t understand the idea at all and there were way too many ads. Participant 4 said that the game is a bit poor but the idea of it is ok. Participant 5 complained about the ads. Participant 6 complained about the missing sounds and audio in the game. Participant 7 thought that everything in the game is basically thrash. Participant 8 said that the controls are too simple, there is no clear goal in the game and way too many ads. The interesting find from these comments was that participants 1, 4, 6 and 8 had way more detailed and technical thoughts and feedback about the game compared to the other opinions that just generally said that the game is bad mostly because of too many ads. Participant 4 even complimented the games idea and participants 1 and 8 gave more technical examples why the game was bad like missing sounds or bad controls. This is probably explained by the fact that these participants were game developers who had experience with mobile games. Participant 6 who was a player also gave feedback on more technical issue which was also the missing sounds. This is probably explained by the fact that participant 6 was the only player participant who played mobile games. This is also in line with the results of Nielsen's study [4]

Based on the interviews only participant 1 had played both games before. This was unexpected. I initially thought that probably someone has played 8 Ball Pool because it is a very popular game and quite old already. Gem Stack however was a much more recent game and not that popular, so I didn't expect anyone to have played that game. Participants 2, 5 and 8 had never played either of the games. Participants 3 and 4 said that they had played 8 Ball Pool many years ago when they were younger, but they hadn't played Gem Stack. Participant 6 that he hadn't played Gem Stack, but he had played 8 Ball Pool approximately 40 hours. Participant 7 said that he had played some billiard game when he was younger but couldn't remember was it this one. He hadn't played Gem Stack

Based on the interviews participants had very different answers when asked did they familiarize themselves with the instructions which was recommended by the author. Participants 1 and 8 said that they read the instructions. Participant 2 said he very quickly looked at the first task. He didn't look at the heuristics or task 2. Participant 3 said he looked at everything except the heuristics. Participant 4 said that very quickly looked through the instructions but not with great detail. Participant 5 said he read through the instructions once quite quickly. Participant 6 said he didn't look at the instructions at all. Participant 7 said he looked at the instructions but didn't even see task 2 for some reason. Based on these answers' players (participants 2, 3, 6, 7) didn't look at the instructions as good as game developers did. This might have affected the results of the evaluations or at least the number of questions asked by the participants.

6.3 Differences between players

In this subsection I will examine the distinctions among players based on the findings presented in chapter 5. The players were participants 2, 3, 6 and 7. This examination will serve as the foundation for Section 6.5, where I will address **R**esearch **Q**uestion 2: **How large is the evaluator effect with game developers and players when evaluating a mobile game** and provide insight into the reasons for the evaluator effect.

When looking at the severity ratings between the players from chapter 5, there are some differences to be found. With 8 Ball Pool the differences are very small probably because 8 Ball Pool is such a finalized game. However, when looking at table 9 about the severity rankings for players with 8 Ball Pool participant 7 had much less “No problem” heuristics and more “Cosmetic” problem heuristic compared to other players. There is no clear explanation for this because the participant didn’t have any comments next to most of the severity ratings.

With Gem Stack the differences between the players were more significant according to table 10 in chapter 5.1. The biggest differences can be seen from the answers of participant 7. He thought with every heuristic that there is some level of problem, and the most notable thing is that he ranked 10 “Major” and 10 “Catastrophe” heuristics out of 25 heuristics. This was very different from the other players. This might be explained with the participants frustration of the game. Immediately after stopping the evaluation of Gem Stack and starting the interviews there were comments about how bad the game was. So it might be that participant 7 was just so fed up with the game he thought that every heuristic is either major or catastrophic. However, other participants also had comments about the game and how bad they thought it was but didn’t have such a large amount of major or critical issues. It is hard to analyse what was the reason that participant 7 had such strong thoughts about the game. It might be because participant 7 was a casual gamer who only played console sports games so game like Gem Stack was a type of game he had probably never played and probably would never play in his free time. So having the participant play a game he would probably never play or enjoy might have caused the strong feelings towards the game. However, the amount of problems found by participant 7 even with his little experience with games agrees with the findings of Nielsen & Molich [30] . Other differences were mainly between participants 2, 3, 6. Participant 6 had no “Catastrophe” errors. This could be explained with the fact that he was a hardcore gamer who played multiple kinds of mobile games, so he had probably played similar types of games and knew overall about mobile games, so he didn’t consider the problems to be that big.

When looking at player fulfillments for the games in table 17 in chapter 5.3 all the players were fulfilled with 8 Ball Pool except for participant 3 who had a bit lower fulfillment score compared to others. It is difficult to say why he had lower score than others because according to table 17 he had similar severity rankings compared to the other players. With Gem Stack all the players had a reasonably low score except participant 6 probably because he was a hardcore mobile gamer like mentioned earlier. Also, participant 7 had significantly lower score than the other players which might have because of the reasons already mentioned.

6.4 Differences between game developers

In this subsection I will analyse the differences between game developers based on the results presented in chapter 5. This examination will also help answering the **R**esearch **Q**uestion 2: **How large is the evaluator effect with game developers and players when evaluating a mobile game**.

When looking at the severity rankings by game developers for 8 Ball Pool in table 11 in chapter 5.1 the results are mostly the same between the game developers. However, there is one noticeable difference. Developer 2 (participant 4) had one heuristic with the ranking of “Major” problem. With 8 Ball Pool participant 4 was the only one with a heuristic that had a severity ranking of three or bigger from all the participants of the study. The heuristic that the participant had marked as “Major” problem was heuristic number 24: *Interruptions are handled reasonably.* The reason why the participant gave the ranking can be seen from the comment the participant had made to the comment part of the heuristics. The participant had written: *Hard to say. Not really.* This comment implicates that the participant had more problems with the heuristic than the actual game. It seems he didn’t know how he should answer this or test it, so he gave it a severity of 3. This is verified from the interview answers found in chapter 5 where the participant said this about heuristic number 24: *Didn’t really know what this means. How am I supposed to test this.* Developer 3 (participant 5) had only “No problem” or “Cosmetic” problem heuristics compared to other developers who had some “Minor” problems also.

With Gem Stack the differences were bigger which are seen in table 12 in chapter 5.1. All developers except developer 2 (participant 4) had “Catastrophe” problems. This could be explained by developer 2 having experience only working with mobile games. However, developer 1 (participant 1) also worked with only mobile games and he had 3 “Catastrophe” problems so this cannot be the only reason. Developers 3 and 4 (participants 5, 8) however had twice the amount of “Catastrophe” problems compared to developer 1 so it seems having the experience from making mobile games shows in the answers of developers 1 and 2. Developer 4 also had worked on mobile games, but he also worked on PC and VR games compared to developers 1 and 2 who only worked on mobile games. Developer 3 only worked with PC games, and I believe this shows in his answers. He had the least amount of “No problem” heuristics compared to other developers and the most “Catastrophe” heuristics. It seems that developer 3 was not familiar at all with these kinds of mobile games which is seen also from the interview answers. According to the answers in chapter 5 of this study developer 3 (participant 5) was very annoyed by the game and the ads.

Interestingly these results don’t match with the finding made by Nielsen where double specialist found more problems [4]. In this study participants 1, 4 and 8 could be considered as double specialist because they were game developers, but they also had experience from heuristics. For 8 Ball Pool there was only one problem (severity 3 or 4) and the reason for this was discussed above. But for Gem Stack these double specialists had less or equal number of problems compared to the one game developer who was not a double specialist. So, for these game evaluations it would seem that being a double specialist didn’t help in finding more problems.

Game developer fulfillments shown in table 18 in chapter 5.3 show that all the developers scored high fulfillments for 8 Ball Pool. Developers 3 and 4 had very high scores. There are no clear explanations why they gave that high scores. For Gem Stack the fulfillment was lowest with developer 3 most probably because his lack of knowledge from mobile game development like discussed earlier. Expectedly developer 4 had the second lowest score. Interestingly developer 1 had a higher score than developer 2 even though developer 1 had more “Catastrophe” heuristics.

6.5 Evaluator effects

In this subsection I will discuss the evaluator effect results shown in chapter 5.2 for players and game developers and compare these findings. Finally, I will answer to **R**esearch **Q**uestion 2: **How large is the evaluator effect with game developers and players when evaluating a mobile game?**

6.5.1 Players

When looking at the evaluator effects for players in table 15 in chapter 5.2, the evaluator effect for 8 Ball Pool was 0%. The reason for this was because there were no issues with the game so the developers cannot agree with non-existent issues.

With Gem Stack and players, the evaluator effect was 23.02%. With players all the pairs had some sort of agreement on the problems. The pairs with the biggest agreements were players 1 & 2 and players 2 & 3. I believe the agreement between players 1 and 2 might be explained with the fact that both players only played PC games. The agreement between players 2 and 3 might have been explained by both participants identifying as hardcore players. So even though player 3 played also mobile games, both participants could agree on some problems with their background in playing games. The lowest scores where with the pairs of players 2 & 4 and players 3 & 4. I believe the low agreement might be explained by both pairs having player 4, a casual player who only plays console games. However, with the pair of players 1 & 4 the agreement was not that low. This might be because also player 1 was a casual gamer. Player 4 also said that he only plays sports games and player 1 played sports games along with other kinds of games. This was the other common thing between the participants, but I believe the effect of both participants playing sports games on the evaluation results is very low.

6.5.2 Game developers

With game developers and 8 Ball pool the evaluator effect was also 0% as seen in table 16 in chapter 5.2. With game developers only one developer had one problem with the game and none of the others didn't so there could be no agreement

With Gem Stack and game developers the evaluator effect was 15.87%. With the game developers all pairs except the pair of developer 1 & 3 had some agreement. The non-agreements with developers 1 and 3 are hard to explain. I first thought it might be explained with the fact that developer 1 is a mobile game developer and developer 3 is a PC game developer but with the pair of developer 2 & 3 the agreement was the biggest among the pairs of developers and developer 2 was a mobile developer so it cannot be because of that. One difference that developer 1 and 3 have is that developer 1 had much more roles in game development such as game design, UI design, textures, 3D modeling and programming whereas developer 3 only had programming and balancing. These role differences might explain why they didn't agree on problems. With developer 2 his roles were programming, game design and balancing which were very similar roles to developer 3 which might explain the higher agreement between them. Developers 2 & 4 also had the highest agreement of 21.4% among the pairs of developers. There were no clear connections between these two developers that might explain the agreements. Surprisingly for the author the agreements between developer 1 and 2 were not bigger because these two participants had very similar backgrounds. Both were only mobile game developers who had worked with games approximately the same amount and worked on the same number of projects. They both also had some earlier experience from utilizing heuristics. The only differences between the participants were that developer 1 had more roles and played games a bit more. However, the pair of developer 1 and 2 had an agreement close to the biggest one among developers.

6.5.3 Comparing players and game developers

Overall game developers had lower agreements compared to players. This might be because the differences between the players were mainly how much they play or what kind of games they play. The most important thing however is that these participants play something. They are all players who know something about games no matter what kind of games so they can agree on some things probably a bit easier. With game developers however the biggest thing in the authors opinion is the fact that these developers work on different kinds of games and with different roles. In game development you must look at games from a different perspective. You need to take the business side into account and heavily focus on different things depending on your roles. I believe that if you work on a mobile game for example and another person works on a PC game, you have to focus on totally different things. This is because these are totally different gaming worlds with different trends and focuses. I think this is why when evaluating a game, two different kinds of developers don't agree that easily.

6.5.4 Discussion

In general, calculating the evaluator effect doesn't really work for more finished games as seen from the results with 8 Ball Pool in the authors opinion. With this kind of very well-made games the number of problems found at all are low. Calculating the evaluator effect this way as done in this study does not give relevant results. However, there must be some sort of agreements between the participants about "Cosmetic" or "Minor" problems which could be interesting to examine in future studies. With a game that is still more in development and not finished like Gem Stack, calculating the evaluator effect works in the authors opinion and can bring out interesting findings.

To answer the **R**esearch **Q**uestion 2: **How large is the evaluator effect with game developers and players when evaluating a mobile game?**

The evaluator effect is larger with players with **23.02%** for Gem Stack. With game developers and Gem Stack the evaluator effect is **15.87%**. With 8 Ball Pool the evaluator effect for game developers and players is **0%.** Based on the different studies and their results showed in the study made by Hertzum & Jacobsen, the evaluator effect calculated with any-two agreement has been from 5% to 65% [1]. It is important to remember however that the values depend on the specific studies and methods used in them. I believe the biggest reason for the evaluator effect and how large they were was the backgrounds of the evaluators. However, this is not the main reason because developers 1 and 2 had very similar background but their agreement was still reasonably small so there are more factors that affect the evaluator that should be studied more in the future.

6.6 Heuristics

When looking at the heuristics and how good and usable the participants considered them to be, the interview answers and the System Usability Scale (SUS)-questionnaire results give an overall idea. I will first go through and discuss the interview answers related to questions about heuristics from chapter 5.5 which were Q4-Q11 and then look at the SUS-questionnaire answers. These answers work as a foundation for answering **R**esearch **Q**uestion 3: **How usable are the new game heuristics when evaluating a mobile game**? and **R**esearch **Q**uestion 4: **How effective does the heuristics work when used in evaluating a game that is in development compared to a fully developed game**?

Q4: Were you able to use the heuristics successfully when evaluating the games?

Most of the participants were able to use the heuristics successfully when evaluating the games. Some participants had some parts that were confusing. The answers to this question didn’t surprise me. I was expecting some issues with the heuristics but overall, I am satisfied that all could use them without any bigger issues.

Q5: Are these heuristics easy to understand?

Most of the participant said that they could understand the heuristics but some of the heuristics were confusing. As mentioned before this did not come as a surprise. Participants 6 and 7 mentioned they had problems with language. This came as a little bit of surprise to me because I had explained to the participants in the initial info about the study that basic English is required. However, the term basic English can mean a different thing to different people so I don’t know what I could have done about this. Very surprising answer was that participant 3 said that he thinks the heuristics are bad.

- Participant 3(player): *I don't think the heuristics were very good, a bit strange and the heuristics have too many questions, it should be detailed and more separated. There was one heuristic about sounds, vibration, and visual effects, it might be better to be separated to individual ones. Some of the heuristics I had trouble saying anything.*

The comment that there are too many questions seemed weird to me. He also said that some of the heuristics should be separated to individual heuristics which would in fact make the heuristics even longer, so I personally think his criticism was a bit contradictory. But individually the comment about separating some of the heuristics made sense.

Participant 7 also mentioned that the word “heuristic” was not familiar. I agree that the word can be a bit confusing. However, the word was briefly described in the background questionnaire in which the author wrote:

- *Heuristics can be defined as broad usability principles or as broad rules of thumb that are not specific guidelines.*

It is unclear if the participant read the background questionnaire properly, did they just forget the meaning or is the definition unclear. Perhaps the definition should have been added to the evaluation form.

Q6: What are the strengths of these heuristics?

Majority of the participants said that the heuristic list is good for basic and necessary things. Participants also mentioned that the list is very broad, and its usage is fast. The heuristic list was created to be a general list for all kinds of mobile games, so the interview answers indicate that in that sense the list was a success. How fast the heuristic list is used is very much in connection with the person using it. Participant 3 also mentioned that one of the strengths is the ability to comment to problems. This comment was more about the evaluation form than the actual list of heuristics, so it seems that having a separate comment section is beneficial.

Q7: What are the limitations of these heuristics?

Half of the participants said that the list is too general / broad. As mentioned earlier the list was designed to be very general and broad. However, this issue should be considered in future studies. It was also seen from the answers to the previous question that majority of the participant thought that it’s a good thing that the list handles basic and necessary things. Some participants said in some heuristics there are too many things in one heuristic. I agree with these comments to some degree. Some of the heuristics could be separated to make the heuristics more simple. For example, heuristic 7: *Does the interface use audio and visual effects to arouse interest to the player* could be divided to heuristics “Does the interface use audio to arouse interest to the player” and “Does the interface use visual effects to arouse interest to the player”.

Participant 5 mentioned that the heuristics don’t take the evaluator in to account meaning that the heuristics don't consider the evaluator and their limitations. I agree with the participant that this was a clear weakness. However, I think that this issue should be handled with the evaluation’s instructions and not with the heuristics themselves. The heuristics that focus on sounds or vibrations are important in the author’s opinion because they bring good elements to the game and its playability and make the game more fun. This is why I think they shouldn’t be removed or modified. The instructions about the evaluations should more clearly indicate that you must turn on sounds and vibrations and you should do the evaluations in an environment where you can safely use them. It was unclear why the participant couldn’t use sounds or vibrations. Every phone should have working sounds and vibration unless the phone is somehow broken. I believe the case was that the participant was in an environment where he couldn’t use them.

Participant 6 mentioned that it is unclear what to answer when something has not been implemented. I believe this has more to do with the individual heuristics that were confusing which I will discuss later in this chapter. Participant 7 mentioned he had issues with English, but I have analysed this earlier. Participant 8 mentioned that he thinks additions are needed. This addition example he gave was mostly related to ads so I will analyse this with question 10 where there were similar comments.

Q8: Which heuristics are the most useful?

The answers to this question showed that the participants considered many kinds of heuristics in the list to be important and there were no single heuristics that stood out from the answers. The answers show that the heuristic list was successful in having all kinds of heuristics that focus on different things which some were more basic and general whereas some focused on more specific things. I believe the list of heuristics succeeded in being basic and general.

Q9: Which heuristics are difficult to understand?

Answers to this question showed the biggest issues with the heuristics because the participants explained all the heuristics that they thought were confusing individually and explained what was wrong with them. I will go through these in answers in more detail.

Six out of eight participants said that heuristic number 23: *Device UI and game UI are used for their own purposes* was difficult to understand. Participant 2 mentioned that he didn't remember what UI meant so he had problems with the questions. I agree UI can be confusing and some people might not know at all what it means so instead of UI it should be user interface. Participant 7 had challenges with English language and did not know the term UI. Participant 3 was confused what does it mean if they are not used for their own purposes, and does it mean they should use the side buttons of their phone to play the game. Participant 5 was confused how does their device UI have anything to do with the game UI. Even though some of these issues would have probably been fixed with having user interface instead of UI, I now see that this heuristic does not work at all. When Koivisto & Korhonen introduced the heuristic, they only wrote that this heuristic was part of a group of heuristics that are related to visual design and how information is presented [16]. There was no detailed explanation what the meaning of this heuristic was and why was it important to have. I believe this heuristic shouldn't be used at least without proper explanations in future studies.

Five out of eight participants said that heuristic number 25: *The game accommodates with the players surroundings (lighting, noise, other people etc.)* was difficult to understand. With this heuristic the problem with basically all the five participants was once again that it was difficult to understand or using it was very hard. With this heuristic I added the small explanation to the original heuristic like explained in chapter 3 of this study because I thought the heuristic was too confusing. Based on the results this clarification did not help at all and the heuristic was still confusing, so I think the best solution is to not use this heuristic at all

Five out of eight participants said that heuristic number 12: *The player should experience the menu as a part of the game* was difficult to understand. All the participants mentioned that the problem with this heuristic was the overall meaning of it. The participants didn't understand what it meant or how to test it. One participant mentioned that he didn't understand is it a good thing or a bad thing if the menu is part of the game. I agree that this heuristic might be confusing specially to people who are not that much into games and tend to play casually. However, it was surprising to the author that five out of eight participants said it was difficult to understand. This clearly shows there is something wrong with the heuristic. Perhaps it could have been explained a little bit more. For example, explaining that the style of the menu and the game world should be similar. At its current form I don't believe it to be usable in these kinds of evaluations.

Four out of eight participants said that heuristic number 24: *Interruptions are handled reasonably* was difficult to understand. Most of the participants mentioned that the issue with this heuristic was that they didn't know how to test it. Participant 6 said he didn't understand what interruptions meant and thought that it meant the ads in games, and this should be more clearer. When it comes to testing, I agree with the participants. Something like this is very hard to test because it basically requires that someone texts or calls the participants while they are playing the games. This is however very improbable to happen during the evaluations. This should have been tested in the way that the conductor of the study calls the participant while they are playing but this was not possible in the scope of this study. I agree with one of the participants that it can also mean the ads in the games because they are considered interruptions to the normal gameplay. However, in this scenario this probably should have been clarified that it means the ads because calling and interrupting the participants was not possible.

Three out of eight participants said that they had difficulties with heuristic 21: *Maximizes consistency by following the trends set by the gaming community to shorten the learning curve.* Participant 2 said that it was a question where you can have so many options because there are so many different trends with different kind of gaming communities. Participant 4 said the question was vague and hard to answer if you don't know about games that much. Participant 6 said that it was difficult to answer, and he had to think about it much more compared to other questions. I agree with the participants that this question can be a bit difficult to answer to. When it came to the first comment that there are different kinds of trends I initially thought when adding this heuristic to the list that the participants would understand to think about mobile game trends because they were playing a mobile game. I understand there are multiple trends and gaming communities but in the context of playing mobile games I thought it was clear what trends we are talking about. However, even if you focus on just mobile game trends, you have to know something about them. So, if you only play console games it is very difficult to say anything about the trends. The trends can also change very quickly and even game developers could not know what the most recent trends are. These facts make using these heuristic difficult.

Participants 3 and 7 said that heuristic number 14: *Is there audio/visual/haptic confirmation when tapping buttons or other user interface elements* was difficult. Participant 3 explained that the issue with the heuristic is that audio, visual and haptic are in the same heuristic. The participant pointed out that if the problem is only with audio, he has to give a bad score for the heuristic even though visual and haptic were totally fine and these should be separated. I agree with the participant that these should be separated. Participant 7 said that he did understand the heuristic but did not know why you have to ask this. This comment might have come from the participant because he was only a casual player who played only with console and only sports games so it might be he was not that educated about the different parts of games and what make them good. However, this is just speculation. I however consider the heuristic to be important and no other participant didn't have similar comments.

Participant 5 mentioned that heuristic number 9: *Provide means for error prevention and recovery through the use of warning messages* was difficult to understand. The participant commented that there were no errors at any point, so it was difficult to answer to. I agree with the participant that using this heuristic in the evaluations can be hard. If the game is well developed there are probably very good warnings and error preventations. However, if the game is very simple, it is very hard to even make a mistake. This is the case with Gem Stack. It is a very simple hyper casual game made in a way that you cannot really lose or make a mistake. In these kinds of cases the evaluator does not know what to answer. In any case I personally think that the heuristic itself is important and should be there. If you evaluate a game where you can make mistakes and the game does not prevent or warn about them, it is very valuable information.

Participant 6 mentioned that heuristic number 10: *Provide appropriate feedback for user actions (music, sound effects, vibration)* was difficult because you don't give feedback with numbers. When it comes to this heuristic, I strongly believe the participant was just confused how the evaluation and the heuristics work. The comment he made suggests that he thought he has to provide the feedback with the severity rating he was giving. The heuristic meant that the game provides the feedback. The confusion might have come because this participant also had issues with English so maybe he just understood it wrong. He was also the only one to complain about this heuristic, so I consider that this heuristic was not a real problem. I also think that the layout of the question is also fine because no one else had no issues with it.

Participant 7 said he had difficulties with heuristic number 15: *Are the icons clear, understandable, and easy to predict what they do.* The participant said that the issue was with the word icons and not really the heuristic. This was probably because of the participants lack of English skill which he mentioned. I believe the word icons is generally understandable, so I think this was just an individual opinion.

.

Q10: Does the heuristics cover all usability problems in these games you evaluated?

Participant 1 said the heuristic list is missing heuristics about gameplay. This is true but the heuristic list was designed to be focusing on UI elements of the games and not the actual gameplay. Participant 1 also mentioned that there should be heuristic about the sounds of the game. I think this is a good idea but this kind of heuristic is more about the gameplay and playability and not related to the UI elements of the game so it cannot be added to these heuristics. Participant 2 said that the list needs something about the idea of the game. This is a good idea and could easily be added to the heuristics. The heuristic could be something like: “The idea of the game should be clear”. Participant 4 said that the game was lagging on his phone, and this was not covered in the heuristics. This was most probably because of the participants phone and not about the game itself because no other participant had issues with lagging so I don’t think there should be a heuristic about it, but it is an unfortunate issue which should be considered somehow.

Three participants said there should be something about ads in the heuristics. These comments were very interesting to the author. Nowadays mobile games are known to be full of ads because that is mostly their model of revenue. The games have a lot of ads and then the player can remove them by making a purchase in-game with real money. I understand that for a player who doesn’t want to buy the ads off they can be very annoying but for the developers of the game the ads are essential. So, it is very hard to say what kind of heuristic would be good about ads. Maybe it would be something like “The number of ads should be reasonable, and the ads shouldn’t interrupt the game sessions”. This kind of heuristic could work but having a heuristic that is totally against ads just wouldn't work I believe. Of course, there might be games that don't have ads but with mobile games it is very rare I believe. Also, the type of game needs to be considered. In 8 Ball Pool there wasn't really that many ads but with Gem Stack the game was full of them. This is just because the games are very different from each other. Gem Stack is a good example of hyper-casual game where there are usually a lot of ads as when 8 Ball Pool was not a hyper-casual game.

Overall, I agree there should be something about ads. One participant said that there were bugs in the game. I agree that there should be a heuristic about bugs. I tried to pick the games for the study to be as bug free as possible and I think I managed to mostly succeed but some can still exist.

Q11: Do you think these heuristics would be useful in game development? Why? Why not?

Participants 3 and 7 said that this depends on the person who is using the heuristics. I agree with these comments. At the current stage and form of the heuristic list there are questions that are too difficult for a "normal" person to answer. The person would need to be familiar with the gaming world to be able to answer all the questions. Participant 5 said that with the heuristics you can notice the stage of the game whether the game is still in alpha or if it is ready. Four participants said that the heuristic list would take care of the basic things. Participant 1 mentioned that the heuristics should be used in the end phases of development. Participant 2 said that the heuristics awakes questions. These comments suggest that the heuristics would be useful in game development at least in some degree. Only one participant mentioned the stage of development where it would be useful, but he also said that you should keep the heuristics in mind in earlier design. The overall answers would suggest that the participants thought the heuristics would be useful. I believe the heuristics would be useful but only after some changes had been made for them.

When it comes to System Usability Scale (SUS)-questionnaire results show in table 20 in chapter 5.4 there are no significant differences between game developers and players. With both groups there were some who liked the heuristics and some who did not. To analyse these differences, I need to look at the individual results of the participants within the groups.

With SUS-questionnaire the results with players differentiated a lot according to table 20 in chapter 5. Participant 2 gave the list of heuristics used in the evaluation very high score but participant 3 gave the lowest score of all the participants in the study.

Participant 6 gave the list an average score while as participant 7 gave the second lowest score. The score given by participant 3 was expected because of his comments that he didn’t like the heuristics.

The SUS-questionnaire results show that game developers gave overall bigger scores than players. Participants 8 and 5 gave very high scores. Participant 4 gave a reasonably high score. Participant 1 gave a little over the medium score which was the lowest among developers but has much higher than the lowest one of the players. Overall, the scores would suggest that the majority of the participants considered the heuristics to be usable.

To answer the **R**esearch **Q**uestion 3: **How usable are the new game heuristics when evaluating a mobile game**?

The game heuristics are generally usable. The game heuristics are good for evaluating the most basic and necessary things in games, but some modifications need to be made. Some of the heuristics should be divided to multiple separate heuristics. Heuristic 7: *Does the interface use audio and visual effects to arouse interest to the player* should be divided into separate heuristics for audio and visual effects. Heuristic 10: *Provide appropriate feedback for user actions (music, sound effects, vibration)* should be divided into three different heuristics (feedback from music, sound effects, vibration). Heuristic 14: *Is there audio/visual/haptic confirmation when tapping buttons or other user interface elements* should also be divided into three different heuristics based on the different confirmation types. Some heuristics should be removed. I believe if heuristics 23, 24 and 25 are removed, the heuristics would be considered much more usable. Heuristic 12: The player should experience the menu as a part of the game should include a brief explanation to make it clearer. New heuristic about the idea of the game should be added. Also, especially when evaluating mobile games, there should be a heuristic about ads. Based on the SUS-questionnaire results the heuristics also seem to be generally usable.

To answer the **R**esearch **Q**uestion 4: **How effective does the heuristics work when used in evaluating a game that is in development compared to a fully developed game**?

Based on the evaluation results the heuristics seem to be more effective with a game that is still in development compared to a fully developed game. This could be seen from the number of problems found in the evaluated games. With 8 Ball Pool which was considered a fully developed game, the participants simply could not find any issues from the game but with Gem Stack the participants could find plenty of issues with the heuristics. However, the heuristics can work with fully developed games because some minor issues were found from 8 Ball Pool but overall, the heuristics seem to be more effective with a game that is still being developed.

6.7 Future work and limitations

One thing that should be improved in future studies in my opinion are the games that are used in these kinds of evaluations. In this study I focused on mobile games which have some elements that are distinct to them for example all the ads in them. This kind of study where you examine the differences between developers and players should also be conducted using non-mobile games to see how the results differentiate from this study. Doing a study with different kinds of games also remove the elements that might have significant impact on the results which in this case are the ads in mobile games. Future studies should also focus on games of specific genres. In this study the games had random genres that were totally different from each other. Future studies should investigate the evaluator effect and other differences with only single genres for both games for example sports games. In this study I also used games where one of the games was a multiplayer game and the other one was a single player game. This kind of study should be done with similar type of games to understand how the game type effects the results, either both games being single player or both being multiplayer. One big thing that should be improved in future studies in my opinion is the factor that in this study I used one game that was in development phase and the other game was a ready-made game. In this study I think the usage of a ready-made game was not that good of a choice because all the evaluation results showed that the participants thought there is nothing wrong with the game which is understandable because it is a ready-made game but in this kind of study the results were not that interesting. The evaluator effect was 0 for both developers and players just because neither of these group thought that there is anything wrong with the game. For this kind of study these results were not ideal. For future studies I suggest that the games are not ready-made games to avoid the issues I had in this study.

One relevant limitation in this study was the low number of participants. With the given resources it was not possible to have a larger group of participants. The actual data and the results of the study represent a very small group so these results cannot be considered that valid. In future studies the number of participants should be much bigger to get valid data. Another limitation of this study was also the fact that the individual participants were too similar to each other. All the participants were male and approximately same age. For game developers only one of the developers had several years of experience from making games. Originally, I had hoped to get multiple game developer participants who had lots of experience but because of problems getting participants I couldn’t achieve this. In this study the other three game developers all only had a few years of experience from game development. Future studies should focus more on getting participants with longer backgrounds in game development to get data from experienced game developers. The game developers that participated in this study had different kinds of backgrounds when it came to development. Two of the game developers had worked only on mobile games. One developer had only worked on PC and the fourth and final game developer had worked on PC, mobile and VR games. In future studies all the game developers should have similar backgrounds when it comes to what type of games they have made. My proposition is that if future studies use for example mobile games in them, then all the game developers should be mobile game developers. The backgrounds of the players also differentiated heavily regarding the games they play. One player only played console games. Two other players only played PC games. Only one player played mobile games that were used in this study. This could be seen in the results of this study. I suggest that future studies should only have players that are focused on the types of games used in the study for example mobile games. Based on the results of this study, one player had played one of the games that was evaluated before, and he had played it for 40 hours. When selecting the participants for this study, the earlier experiences with the games that are being evaluated was not asked. The participant having this much experience with one of the games affected heavily to the evaluations. Future studies should find out early when selecting the participants if they have experience with the games or not. I believe future studies should have participants with similar amount of experience from the evaluated games.

One limitation of this study that should be improved in future studies was that testing some features in the evaluation sessions was not possible. During evaluations participant 5(game developer) asked what he is supposed to do when he didn’t have the possibility to use sounds or vibrations. This was a surprising issue for me. Future studies should take these kinds of issues into account by maybe doing the evaluations in a specific location with specific equipment instead of doing the evaluations remotely and letting the participants use their own equipment. I believe this would solve majority of the unexpected issues when it comes to technological issues or social situations where the evaluation could be disturbed. Some of the participants said that they didn’t look at the instructions of the evaluations properly. This might have affected the results. This could also be avoided if the evaluations were done in a more controlled environment.

There were also some points mentioned in chapter 6 which were hard to analyse. One point was that some of the participants asked questions after already starting the evaluations instead of asking them in the beginning of the evaluations when they were instructed to ask any questions. I believe this was because the participants hadn’t looked at the instructions or heuristics properly and only asked questions after looking at the heuristics during the evaluations. I believe this issue could be fixed by giving the participants enough time to read the instructions and ask questions but more importantly the conductors of the studies should make sure they read the instructions. This could be handled by having the evaluations in a more controlled environment like mentioned before. One interesting finding from the evaluations were that the participants differentiated in the language they used in the evaluations between English and Finnish. This was because they were instructed to answer in any language they wanted. Future studies should have a clear instruction on the language the participants should use.

In the last question of the interviews, I asked the game developer participants of this study what they thought about the tasks that they did in the evaluations. In task 1 the evaluator had to classify usability problems in the game using the heuristics found in chapter 3 of this study and a severity scale found in section 4.3. In task 2 the evaluator had to assign a score from 1 to 5 (1 being worst, 5 being best) to every single heuristic based on how well the game fulfilled each of them. All the game developers said that the usefulness of these tasks depends on the answers that are made by the people doing them. The game developers said that if the answer is good and high quality and the people doing these tasks spend time with them, the tasks would be useful. One game developer said that he wouldn’t think the tasks would be useful at all if they were only done by a few people. He said that there should be around 100 answers and only then he would consider the answers to be relevant. This statement complies with the previous notice that in the future the number of participants need to bigger. One participant said that if the heuristics only have the severity number but no comment at all why the evaluators think that way, the information wouldn’t be that useful. Perhaps in future studies it is mandatory for the participants to comment on all the heuristics. One game developer also said that if the answers tell that something is good then it’s not so interesting but if the answers actually say something is wrong or irritating then that information is very good. Perhaps this notice could be used in future studies as well. Based on the interview answers only one game developer thought that both of the tasks are good, and two developers said that task 1 is more important than task 2. In the interviews some participants also mentioned that they looked at task 1 answers when they were filling out task 2 which might mean that the tasks are very similar and task 1 is a bit more relevant. In future studies I recommend using both of the tasks but if the resources are limited, I would say removing task 2 is acceptable.

7. CONCLUSION

In this study I researched the differences between game developers and players when evaluating mobile games using mobile game heuristics that focused on the UI elements of the games. The focus was on the number of problems found and the severity of these problems. In addition to this I also examined other differences that might have been found between game developers and players such as how they behave in the evaluations and how many questions they asked. In this study I also focused on the differences between individual game developers and the differences between individual players. These differences were measured by calculating the evaluator effect. In this study I also created a list of new game heuristics from existing heuristics that could be used when evaluating mobile games and analyzed how usable are these new heuristics when evaluating mobile games. I also examined how effectively do these new heuristics work when they were used evaluating a mobile game that is in development compared to a game that is fully developed. The research questions of this study were:

> **R**esearch **Q**uestion **1. What are the differences in results when using heuristics to evaluate mobile games between game developers and players?**
>
> **R**esearch **Q**uestion **2. How large is the evaluator effect with game developers and players when evaluation a mobile game?**
>
> **R**esearch **Q**uestion **3. How usable are the new game heuristics when evaluating a mobile game?**
>
> **R**esearch **Q**uestion **4. How effectively do the heuristics work when used in evaluating a game that is in development compared to a fully developed game?**

In this study I conducted game evaluation sessions with eight different participants. Four of these participants were game developers and four were players. These evaluation sessions were conducted remotely and in the sessions each of the participants played two different games and evaluated them using the list of heuristics I created. One of these games played was a game still in development and the other one was fully developed. After the participants had evaluated both games, interviews were conducted.

The results of the study showed that players spend more time doing the evaluations compared to developers. Players found more problems in the games compared to game developers, but game developers had more of problems that were classified as catastrophic. The results also showed that game developers are more forgiving with issues of the games. These results could be explained by the game developers who participated in this study having very different backgrounds, which is reflected in the results. With players this same phenomenon was found. The evaluator effects calculated showed that players had more agreements on the problems of the games compared to game developers. Overall, the results showed that the background and the experience of the evaluators affect the results heavily. This finding was also present in these studies [35] [33]. Based on the results, people who had more experience with the types of games that were evaluated, succeeded in the evaluations better, gave lower severity ratings to the heuristics and had more agreements with participants who also had experience. The new heuristics created for this study were found to be usable and the list is good when it is used in evaluating the most basic and necessary things in games. However, the results showed that some modifications needed to be done to the heuristics to make them better and more usable in future studies. After making these modifications presented in chapter 6.6 the new heuristics can be seen in appendix B. Heuristics 7, 10 and 14 were divided into multiple different heuristics (heuristics 7&8; heuristics 11&12&13; heuristics 17&18&19). The original heuristics 23, 24, 25 were removed. Small explanation was added to heuristic 12 (now 15). New heuristics were also created (heuristics 28 & 29). The results also showed that these heuristics were more effective when used in evaluating a mobile game that is in development compared to a mobile game that is fully developed.

The results of this study show that game developers and players can have very different results when doing mobile game evaluations and their backgrounds affect heavily on these results. The results also show that the evaluator effect is present when it comes to mobile game evaluations and game developers and players. In the scope of this study the evaluator effect could be explained to some degree with the backgrounds of the participants. This means mainly the types of the games they had played or developed. However, future studies are needed to study more of the reasons for the evaluator effects that are not related to different backgrounds. From this study we now also have a solid foundation for a list of heuristics that can be used in evaluating mobile games. The study also showed an important result about the type of game used in mobile game evaluations. The games used should be still in development and not fully complete games.

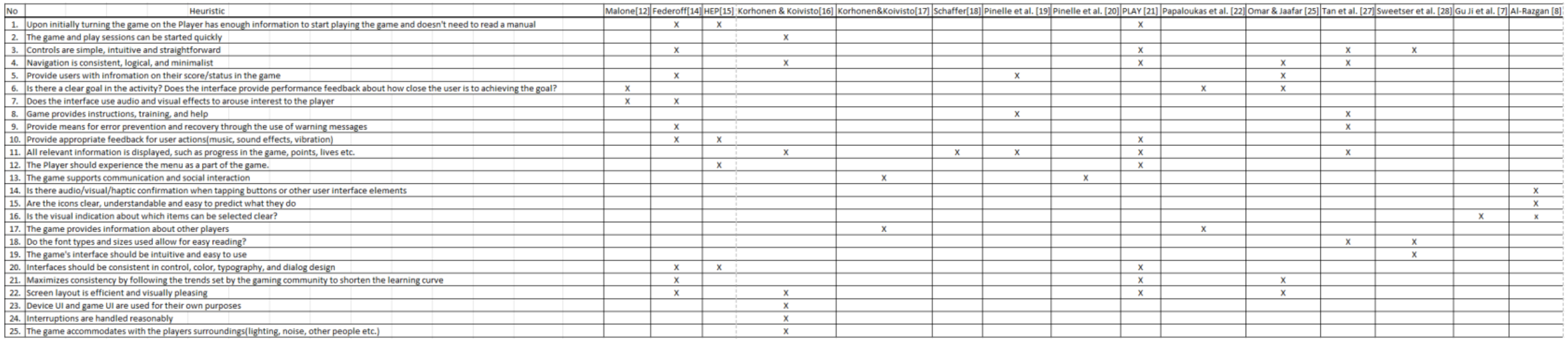
8. REFERENCES

1. Hertzum M, Jacobsen NE, The evaluator effect: A chilling fact about usability evaluation methods, International Journal of Human-Computer Interaction. Taylor and Francis Inc., 2003, p. 183–204.
2. Jokela T, When Good Things Happen to Bad Products: Where are the Benefits of Usability in the Consumer Appliance Market?, 2004
3. Nielsen J, Enhancing the Explanatory Power of Usability Heuristics, Human Factors in Computing Systems, April 1994, p. 152-158
4. Nielsen J, FINDING USABILITY PROBLEMS THROUGH HEURISTIC EVALUATION, May 1992, p. 373-380
5. https://www.nngroup.com/articles/ten-usability-heuristics/. Referenced 24.1.2023
6. Molich R, Nielsen J. Improving a Human-Computer Dialogue, COMPUTING PRACTICES volume 33 number 3, March 1990, p. 338-348
7. Gu Ji Y, Ho Park J, Lee C, Hwan Yun M, A Usability Checklist for the Usability Evaluation of Mobile Phone User Interface, INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION, 2006, p. 207-231
8. Al-Razgan MS, Al-Khalifa HS, Al-Shahrani MD, LNCS 8517 - Heuristics for Evaluating the Usability of Mobile Launchers for Elderly People, 2014, p. 415-424
9. Johnston A, Pickrell M, Designing for technicians working in the field: 8 usability heuristics for mobile application design, 2016, p. 494-498
10. Kumar BA, Goundar MS, Usability heuristics for mobile learning applications, Education and information Technologies, 2019, p. 1819-1833
11. Parente da Costa R, Dias Canedo E, A Set of Usability Heuristics for Mobile Applications, 2019, p. 180–193.
12. Malone TW, Heuristics for Designing Enjoyable User Interfaces: Lessons from Computer Games, ASSOCIATION FOR COMPUTING MACHINERY, 1981, p. 63-68
13. Clanton C, An interpreted demonstration of computer game design. April 1998, p. 1–2.
14. Federoff MA, Heuristics and usability guidelines for the creation and evalaution of fun in video games, 2002.
15. Desurvire H, Caplan M, Toth JA. Using Heuristics to Evaluate the Playability of Games, 2004, p. 1509-1512
16. Korhonen H, Koivisto EMI, Playability Heuristics for Mobile Games, September 2006, p. 9-16
17. Korhonen H, Koivisto EMI, Playability heuristics for mobile multi-player games, 2007, p. 28-35.
18. Schaffer N, Heuristics for Usability in Games, April 2007,
19. Pinelle D, Stach T, Wong N, Heuristic evaluation for games: Usability Principles for Video Game Design, April 2008, p. 1453-1462
20. Pinelle D, Gutwin C, Stach T, Wong N, Usability heuristics for networked multiplayer games, May 2009, p. 169-178
21. Desurvire H, Wiberg C, Game Usability Heuristics (PLAY) for Evaluating and Designing Better Games: The Next Iteration, 2009, p. 557-566
22. Papaloukas S, Patriarcheas K, Xenos M, Usability assessment heuristics in new genre videogames, 2009 - 13th Panhellenic Conference on Informatics. 2009, p. 202–206.
23. Zaibon SB, Shiratuddin N, Heuristics evaluation strategy for mobile game-based learning, 6th IEEE International Conference on Wireless, Mobile and Ubiquitous Technologies in Education, 2010, p. 127–131.
24. Koeffel C, Hochleitner W, Leitner J, Haller M, Geven A, Tscheligi M, Using Heuristics to Evaluate the Overall User Experience of Video Games and Advanced Interaction Games, Evaluating User Experience in Games, Chapter 13, 2010
25. Omar HM, Jaafar A, Heuristics evaluation in computer games, 2010. p. 188–193.
26. Paavilainen J, Critical review on video game evaluation heuristics: social games perspective, May 2010, p. 56-65
27. Tan JL, Goh D, Ang RP, Huan VS, Usability and Playability Heuristics for Evaluation of an Instructional Game, October 2010, p. 363-372

28. Sweetser P, Johnson D, Wyeth P, Ozdowska A, Gameflow heuristics for designing and evaluating real-time strategy games, July 2012.
29. Hermawati S, Lawson G, Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? Vol. 56, Applied Ergonomics, 2016, p. 34–51.
30. Nielsen J, Molich R, HEURISTIC EVALUATION OF USER INTERFACES. April 1990, p. 249-256
31. Følstad A, Anda BCD, Sjøberg DIK, The usability inspection performance of work-domain experts: An empirical study, Interacting with Computers, 2010, p. 75–87.
32. Salian K, Sim G, Read JC, Can children perform a heuristic evaluation? September 2013, p. 137–41.
33. Fitchat L, Jordaan DB, TEN HEURISTICS TO EVALUATE THE USER EXPERIENCE OF SERIOUS GAMES, INTERNATIONAL JOURNAL OF SOCIAL SCIENCES AND HUMANITY STUDIES Vol 8 No 2, 2016, p. 209-225
34. Phan MH, Keebler JR, Chaparro BS, The Development and Validation of the Game User Experience Satisfaction Scale (GUESS), December 2016, p. 1217–1247.
35. Santos T, Lemmerich F, Strohmaier M, Helic D, What’s in a review: Discrepancies between expert and amateur reviews of video games on Metacritic, November 2019
36. Thewes F, Herrmann T, Kluge A, Validating a heuristic evaluation method an application test, 2019, p. 593–597.
37. Li X, Zhang Z, Stefanidis K, A data-driven approach for video game playability analysis based on players’ reviews, 2021
38. Vermeeren APOS, Bekker T, van Kesteren IEH, Bekker MM, Managing the “Evaluator Effect” in User Testing, Human-Computer Interaction, 2003, p. 647-654
39. Ojares M, Peliheuristiikkojen hyödyntäminen pelikäyttöliittymien kehityksessä, April 2020
40. Macey J, Heuristics for Evaluating Video Games: A Two-Tier Set Incorporating Universal and Genre-Specific Elements. October 2016.
41. https://appmagic.rocks/google-play/8-ball-pool/com.miniclip.eightballpool/info. Referenced 24.1.2023
42. https://www.rulesofsport.com/sports/pool.html. Referenced 24.1.2023
43. https://appmagic.rocks/google-play/gem-stack/com.bytetyper.gemstack. Referenced 24.1.2023
44. Brooke J, SUS: A quick and dirty usability scale, November 1995
45. Yanez-Gomez R, Font JL, Cascado-Caballero D, Sevillano JL, Heuristic usability evaluation on games: a modular approach. Multimedia Tools and Applidations, February 2019, p. 4937–4964.

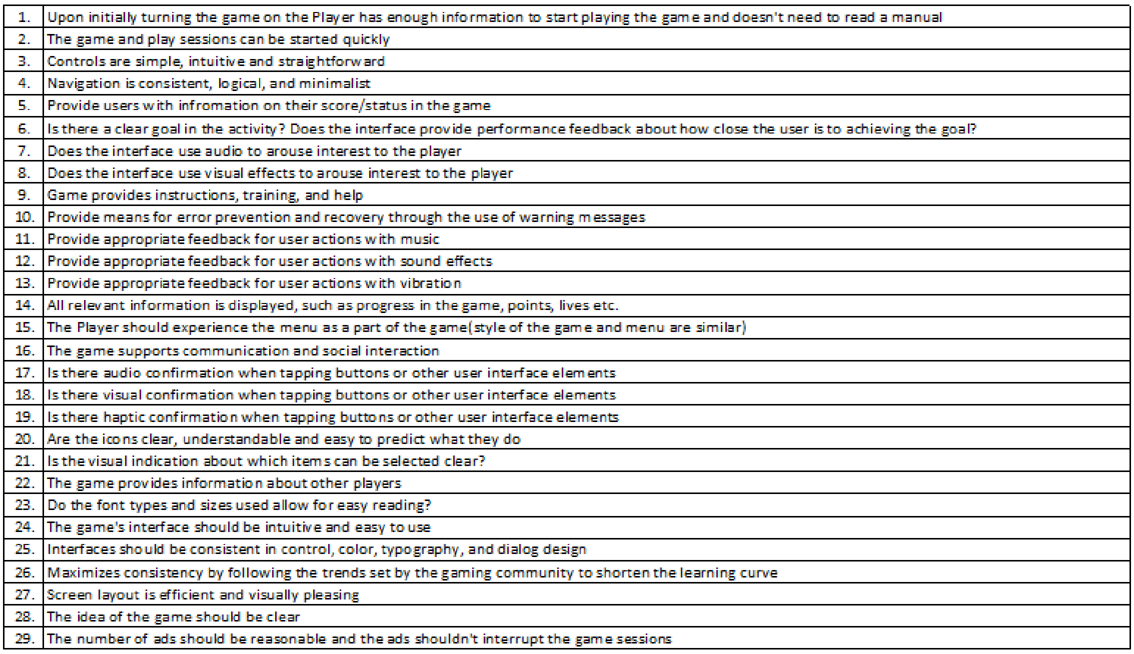
9. APPENDICES

Appendix A: Full list of heuristics and their sources.



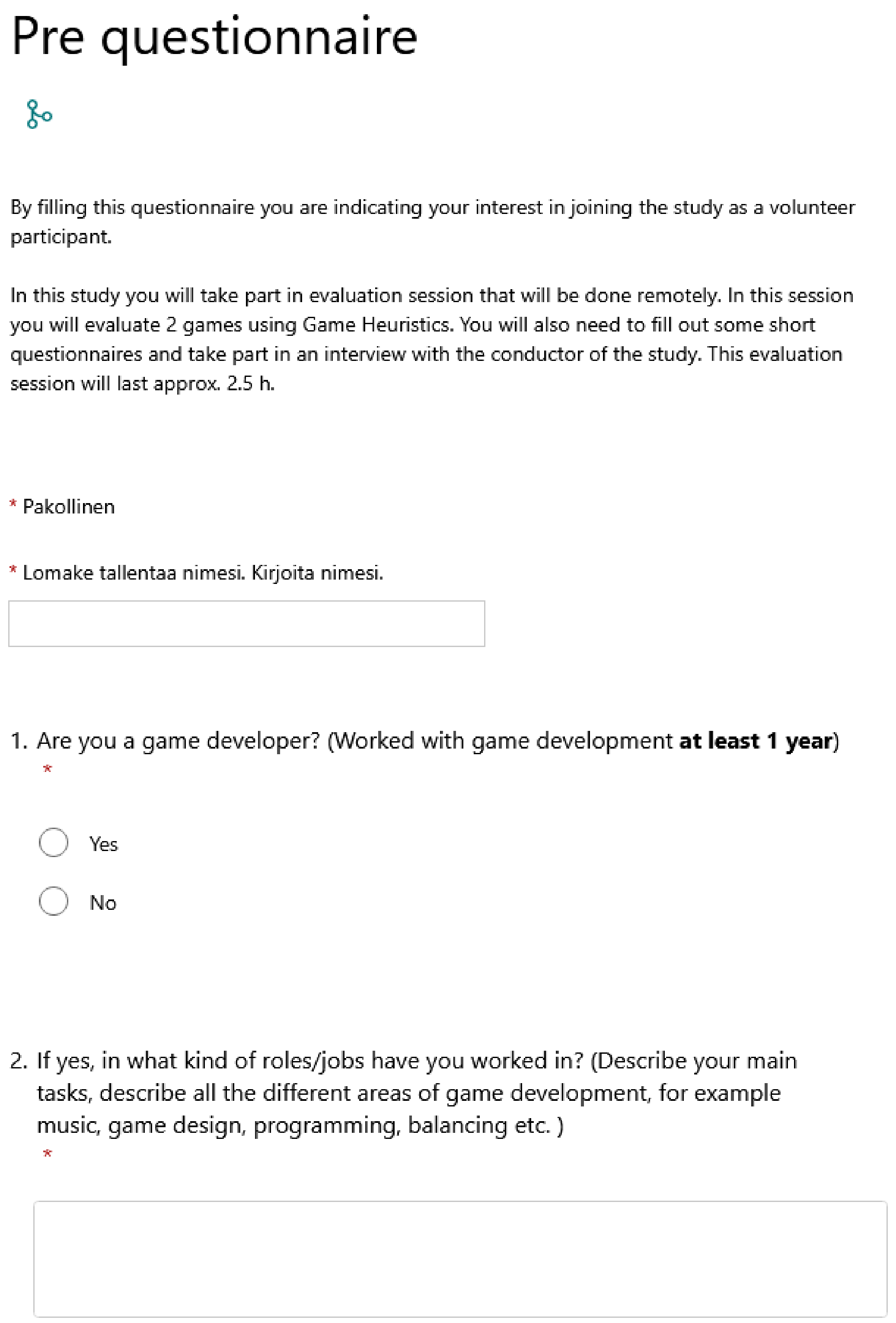
No	Heuristic	Malone[12]	Federoff[14]	HEP[15]	Korhonen & Koivisto[16]	Korhonen&Koivisto[17]	Schaffer[18]	Pinelle et al. [19]	Pinelle et al. [20]	PLAY [21]	Papaloukas et al. [22]	Omar & Jaafar [25]	Tan et al. [27]	Sweetser et al. [28]	Gu Ji et al. [7]	Al-Razgan [8]
1.	Upon initially turning the game on the Player has enough information to start playing the game and doesn't need to read a manual		X	X						X						
2.	The game and play sessions can be started quickly				X											
3.	Controls are simple, intuitive and straightforward		X							X			X	X		
4.	Navigation is consistent, logical, and minimalist				X					X		X	X			
5.	Provide users with infromation on their score/status in the game		X					X				X				
6.	Is there a clear goal in the activity? Does the interface provide performance feedback about how close the user is to achieving the goal?	X									X	X				
7.	Does the interface use audio and visual effects to arouse interest to the player	X	X													
8.	Game provides instructions, training, and help							X					X			
9.	Provide means for error prevention and recovery through the use of warning messages		X										X			
10.	Provide appropriate feedback for user actions(music, sound effects, vibration)		X	X						X						
11.	All relevant information is displayed, such as progress in the game, points, lives etc.				X		X	X		X			X			
12.	The Player should experience the menu as a part of the game.			X						X						
13.	The game supports communication and social interaction					X			X							
14.	Is there audio/visual/haptic confirmation when tapping buttons or other user interface elements															X
15.	Are the icons clear, understandable and easy to predict what they do															X
16.	Is the visual indication about which items can be selected clear?														X	x
17.	The game provides information about other players					X					X					
18.	Do the font types and sizes used allow for easy reading?												X	X		
19.	The game's interface should be intuitive and easy to use													X		
20.	Interfaces should be consistent in control, color, typography, and dialog design		X	X						X						
21.	Maximizes consistency by following the trends set by the gaming community to shorten the learning curve		X							X		X				
22.	Screen layout is efficient and visually pleasing		X		X					X		X				
23.	Device UI and game UI are used for their own purposes				X											
24.	Interruptions are handled reasonably				X											
25.	The game accommodates with the players surroundings(lighting, noise, other people etc.)				X											

Appendix B: New improved heuristic list



1.	Upon initially turning the game on the Player has enough information to start playing the game and doesn't need to read a manual
2.	The game and play sessions can be started quickly
3.	Controls are simple, intuitive and straightforward
4.	Navigation is consistent, logical, and minimalist
5.	Provide users with infromation on their score/status in the game
6.	Is there a clear goal in the activity? Does the interface provide performance feedback about how close the user is to achieving the goal?
7.	Does the interface use audio to arouse interest to the player
8.	Does the interface use visual effects to arouse interest to the player
9.	Game provides instructions, training, and help
10.	Provide means for error prevention and recovery through the use of warning messages
11.	Provide appropriate feedback for user actions with music
12.	Provide appropriate feedback for user actions with sound effects
13.	Provide appropriate feedback for user actions with vibration
14.	All relevant information is displayed, such as progress in the game, points, lives etc.
15.	The Player should experience the menu as a part of the game(style of the game and menu are similar)
16.	The game supports communication and social interaction
17.	Is there audio confirmation when tapping buttons or other user interface elements
18.	Is there visual confirmation when tapping buttons or other user interface elements
19.	Is there haptic confirmation when tapping buttons or other user interface elements
20.	Are the icons clear, understandable and easy to predict what they do
21.	Is the visual indication about which items can be selected clear?
22.	The game provides information about other players
23.	Do the font types and sizes used allow for easy reading?
24.	The game's interface should be intuitive and easy to use
25.	Interfaces should be consistent in control, color, typography, and dialog design
26.	Maximizes consistency by following the trends set by the gaming community to shorten the learning curve
27.	Screen layout is efficient and visually pleasing
28.	The idea of the game should be clear
29.	The number of ads should be reasonable and the ads shouldn't interrupt the game sessions

Appendix C: Initial questionnaire



Pre questionnaire

By filling this questionnaire you are indicating your interest in joining the study as a volunteer participant.

In this study you will take part in evaluation session that will be done remotely. In this session you will evaluate 2 games using Game Heuristics. You will also need to fill out some short questionnaires and take part in an interview with the conductor of the study. This evaluation session will last approx. 2.5 h.

* Pakollinen

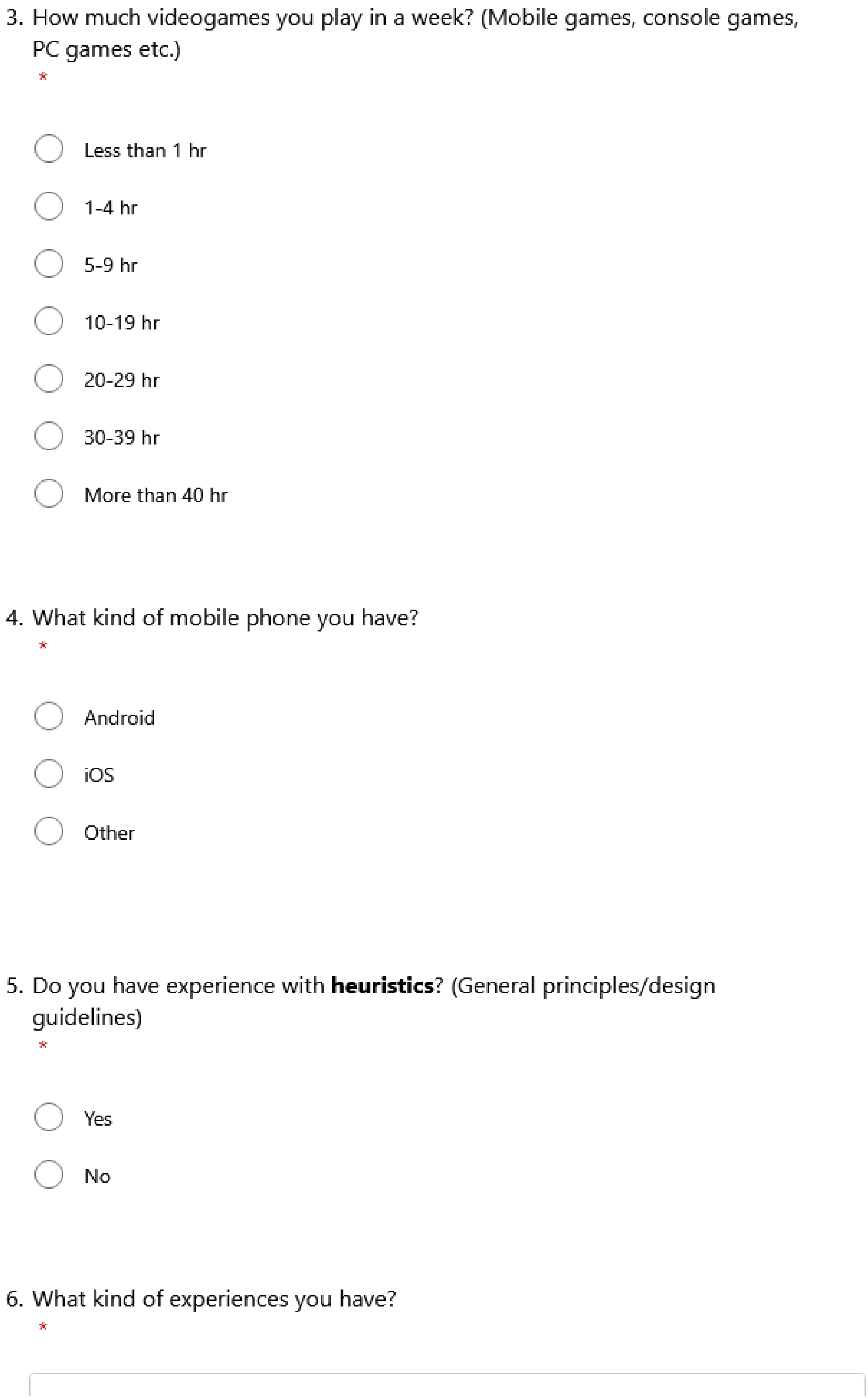
* Lomake tallentaa nimesi. Kirjoita nimesi.

1. Are you a game developer? (Worked with game development **at least 1 year**) *

○ Yes

○ No

2. If yes, in what kind of roles/jobs have you worked in? (Describe your main tasks, describe all the different areas of game development, for example music, game design, programming, balancing etc.) *



3. How much videogames you play in a week? (Mobile games, console games, PC games etc.)
*

○ Less than 1 hr

○ 1-4 hr

○ 5-9 hr

○ 10-19 hr

○ 20-29 hr

○ 30-39 hr

○ More than 40 hr

4. What kind of mobile phone you have?
*

○ Android

○ iOS

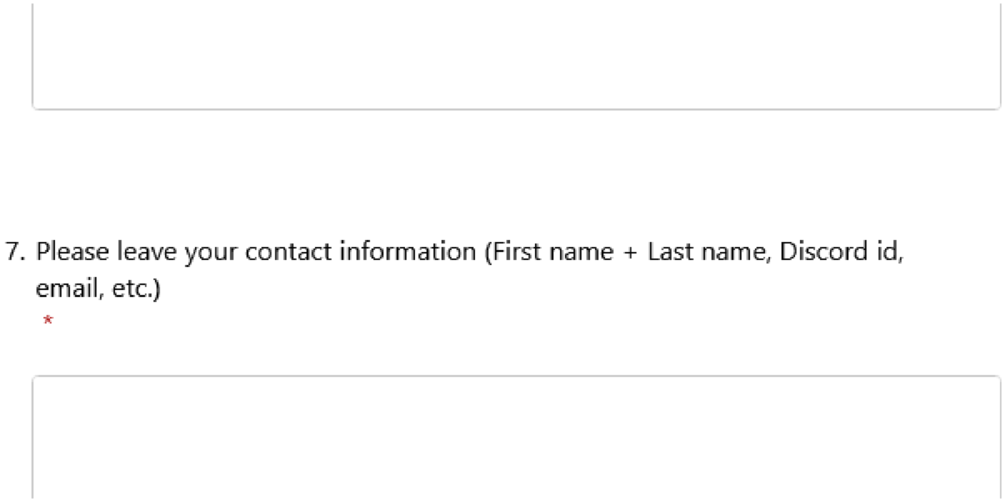
○ Other

5. Do you have experience with **heuristics**? (General principles/design guidelines)
*

○ Yes

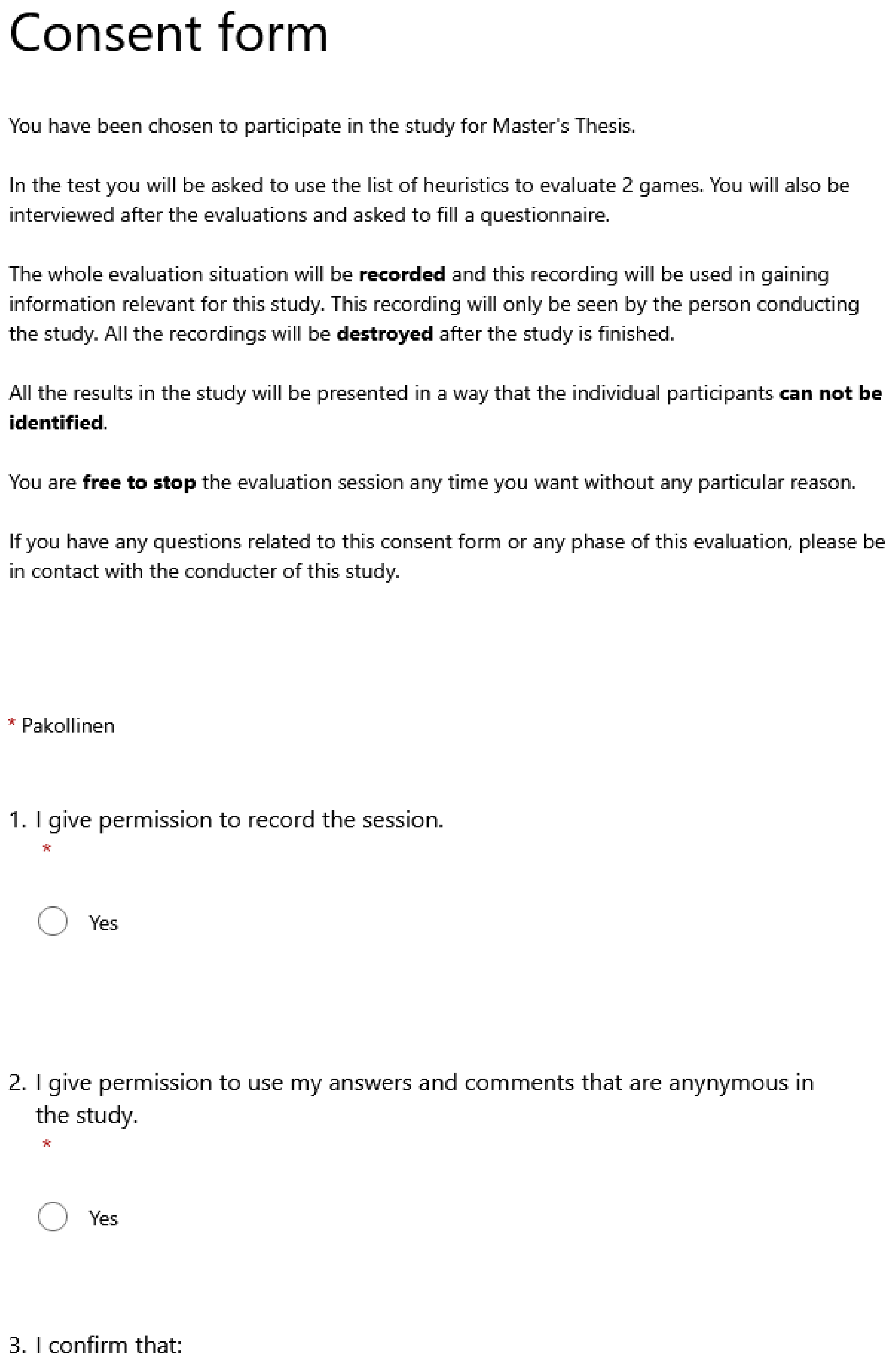
○ No

6. What kind of experiences you have?
*



7. Please leave your contact information (First name + Last name, Discord id, email, etc.)
*

Appendix D: Consent form



Consent form

You have been chosen to participate in the study for Master's Thesis.

In the test you will be asked to use the list of heuristics to evaluate 2 games. You will also be interviewed after the evaluations and asked to fill a questionnaire.

The whole evaluation situation will be **recorded** and this recording will be used in gaining information relevant for this study. This recording will only be seen by the person conducting the study. All the recordings will be **destroyed** after the study is finished.

All the results in the study will be presented in a way that the individual participants **can not be identified**.

You are **free to stop** the evaluation session any time you want without any particular reason.

If you have any questions related to this consent form or any phase of this evaluation, please be in contact with the conducter of this study.

* Pakollinen

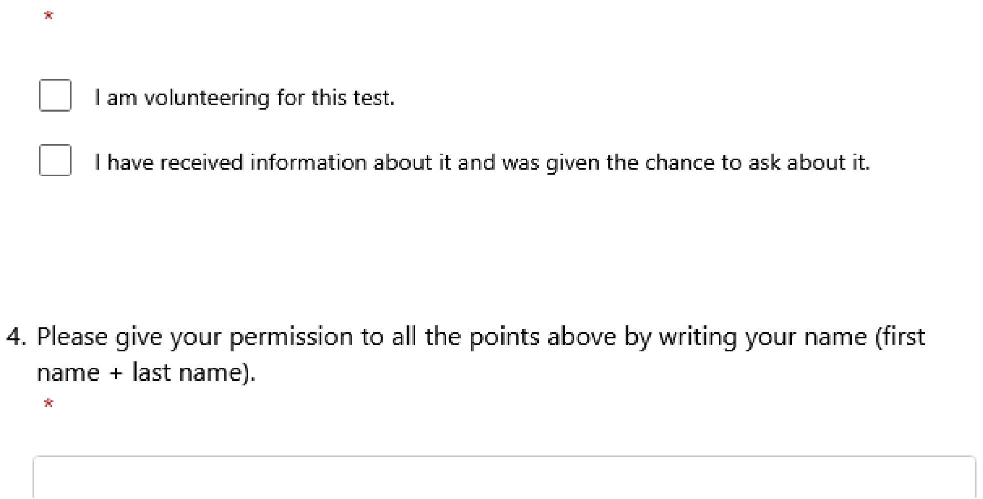
1. I give permission to record the session.
   *

   ○ Yes

2. I give permission to use my answers and comments that are anynymous in the study.
   *

   ○ Yes

3. I confirm that:



*

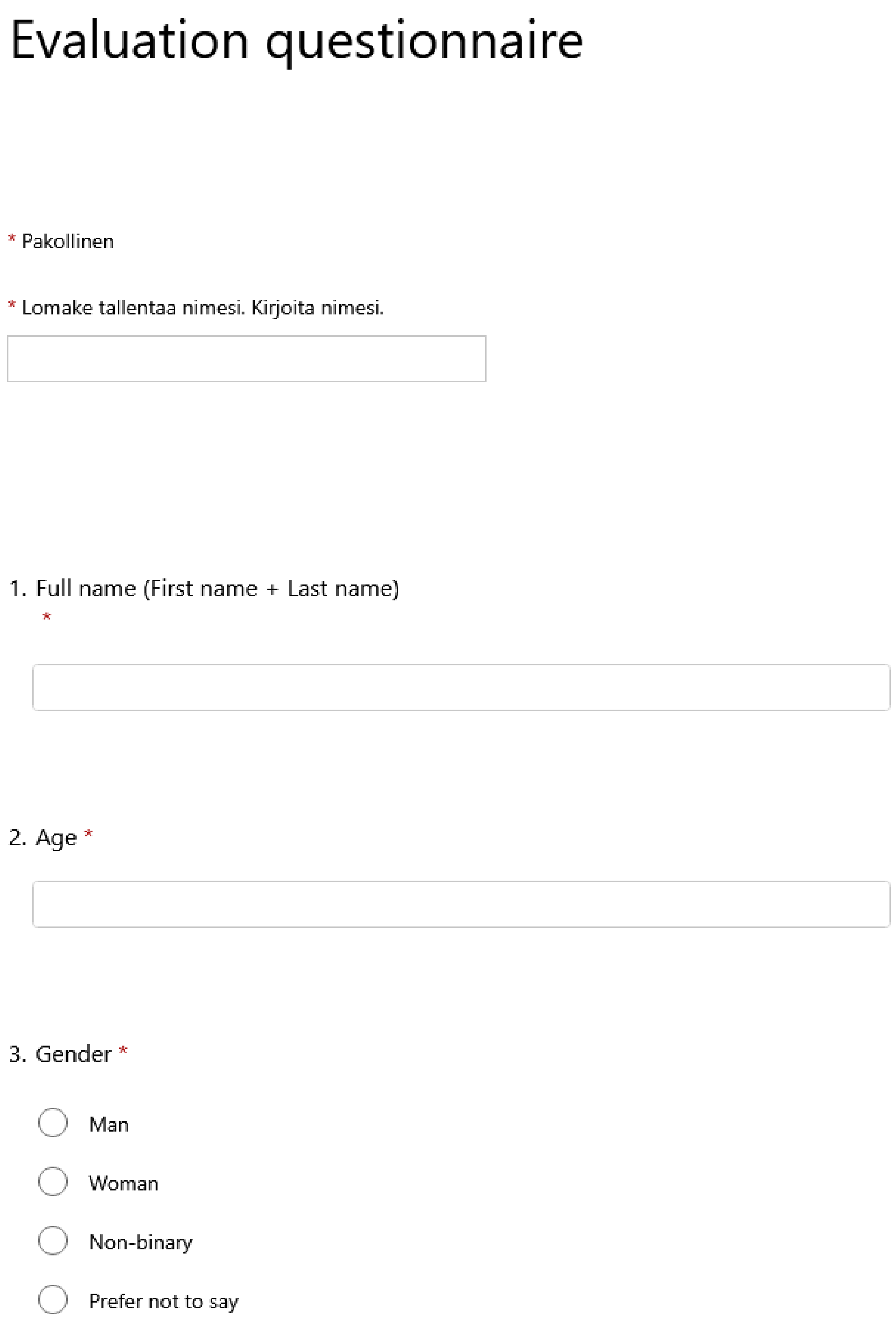
□ I am volunteering for this test.

□ I have received information about it and was given the chance to ask about it.

4. Please give your permission to all the points above by writing your name (first name + last name).

*

Appendix E: Larger questionnaire



Evaluation questionnaire

* Pakollinen

* Lomake tallentaa nimesi. Kirjoita nimesi.

1. Full name (First name + Last name) *

2. Age *

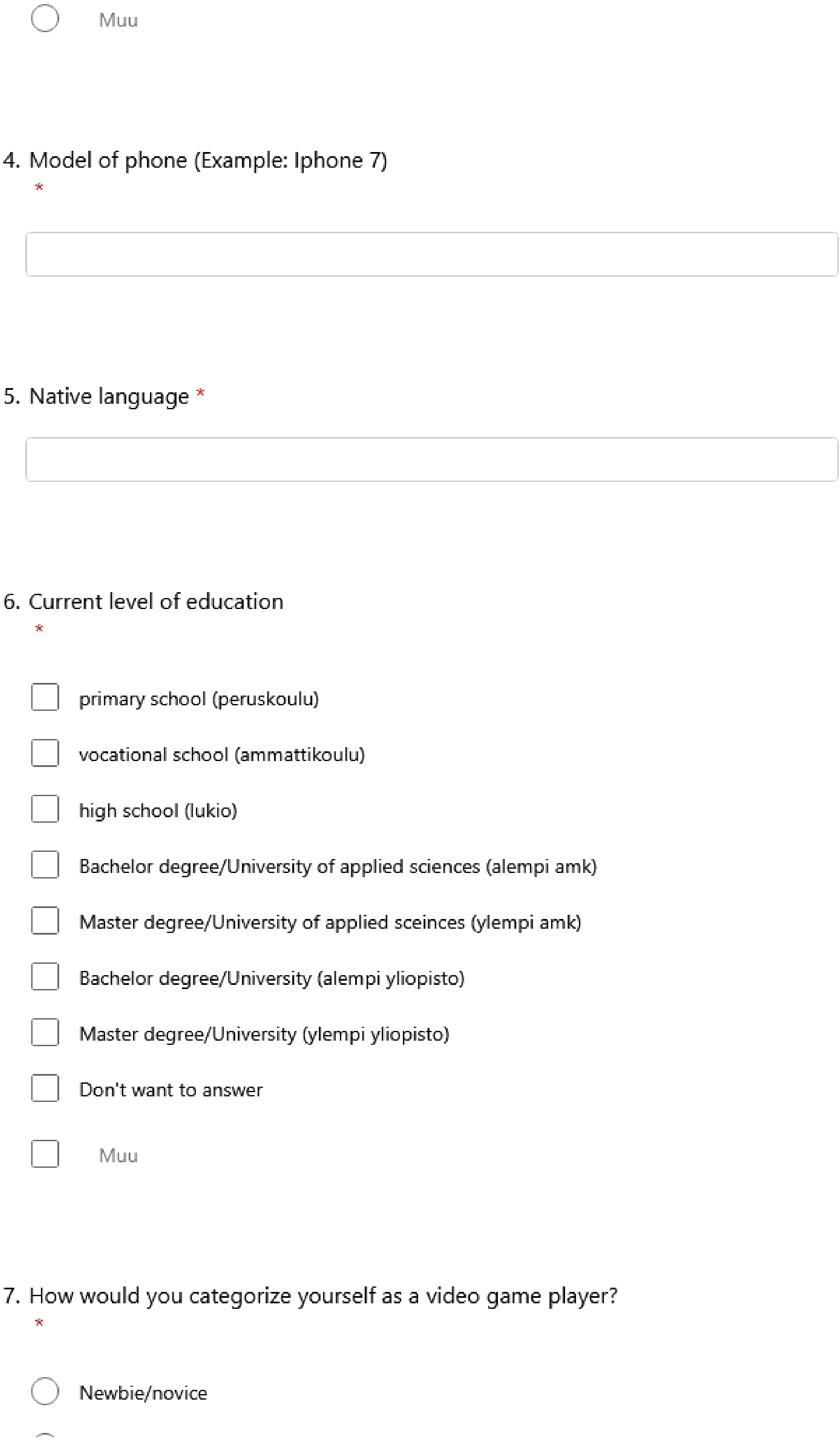
3. Gender *

○ Man

○ Woman

○ Non-binary

○ Prefer not to say



○ Muu

4. Model of phone (Example: Iphone 7)
*

5. Native language *

6. Current level of education
*

□ primary school (peruskoulu)

□ vocational school (ammattikoulu)

□ high school (lukio)

□ Bachelor degree/University of applied sciences (alempi amk)

□ Master degree/University of applied sceinces (ylempi amk)

□ Bachelor degree/University (alempi yliopisto)

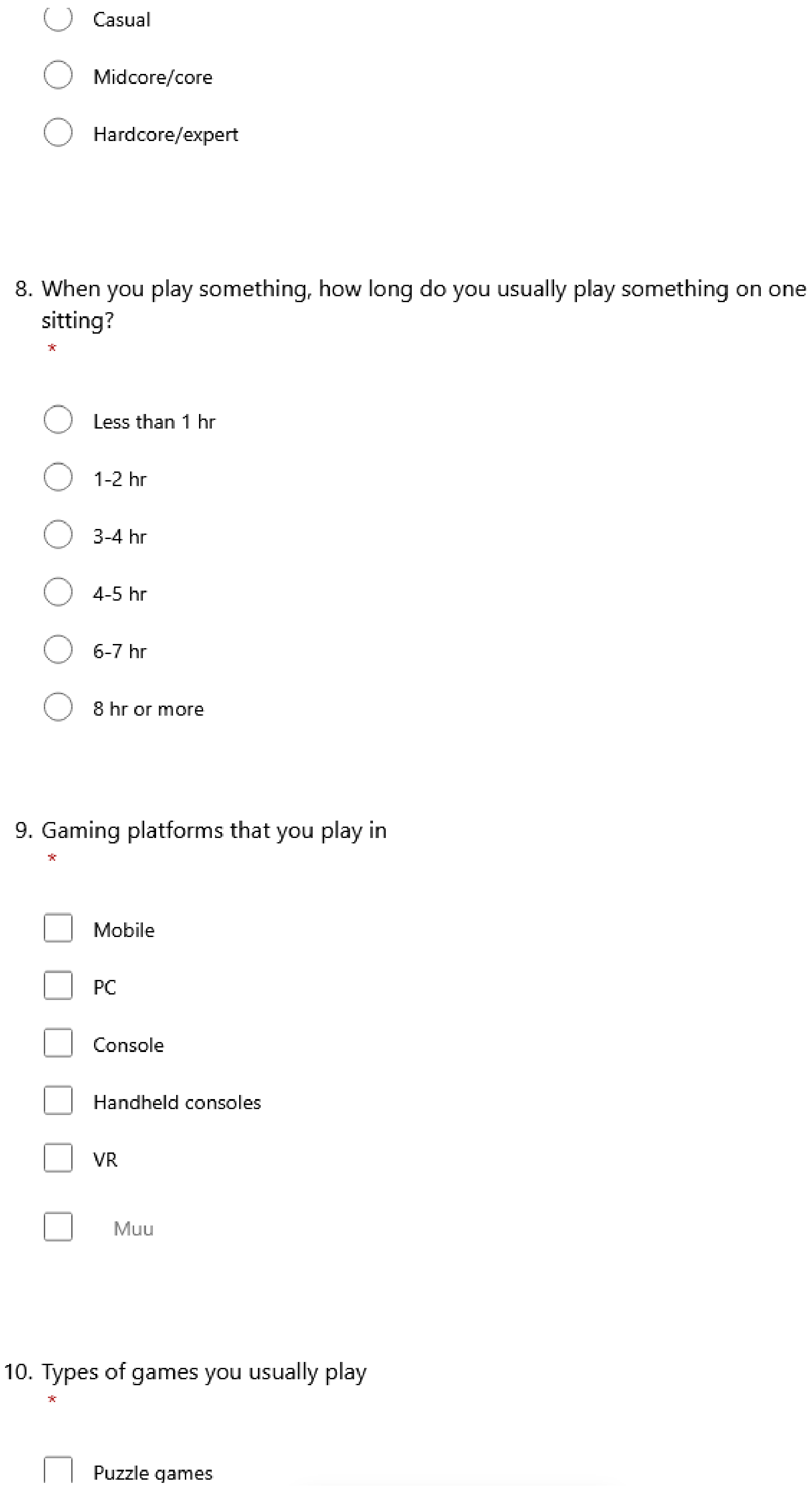
□ Master degree/University (ylempi yliopisto)

□ Don't want to answer

□ Muu

7. How would you categorize yourself as a video game player?
*

○ Newbie/novice



○ Casual

○ Midcore/core

○ Hardcore/expert

8. When you play something, how long do you usually play something on one sitting?
*

○ Less than 1 hr

○ 1-2 hr

○ 3-4 hr

○ 4-5 hr

○ 6-7 hr

○ 8 hr or more

9. Gaming platforms that you play in
*

□ Mobile

□ PC

□ Console

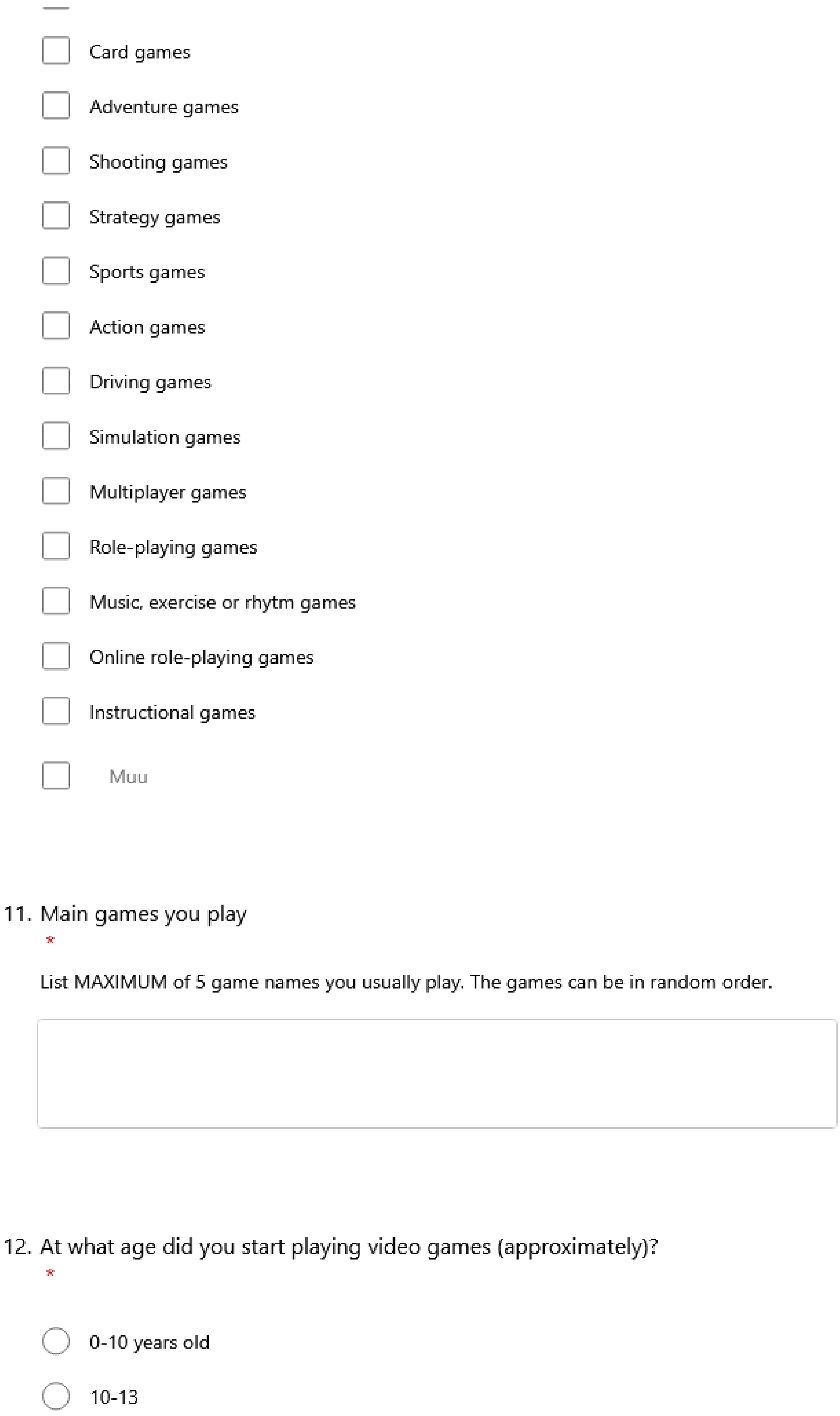
□ Handheld consoles

□ VR

□ Muu

10. Types of games you usually play
*

□ Puzzle games



□ Card games

□ Adventure games

□ Shooting games

□ Strategy games

□ Sports games

□ Action games

□ Driving games

□ Simulation games

□ Multiplayer games

□ Role-playing games

□ Music, exercise or rhytm games

□ Online role-playing games

□ Instructional games

□ Muu

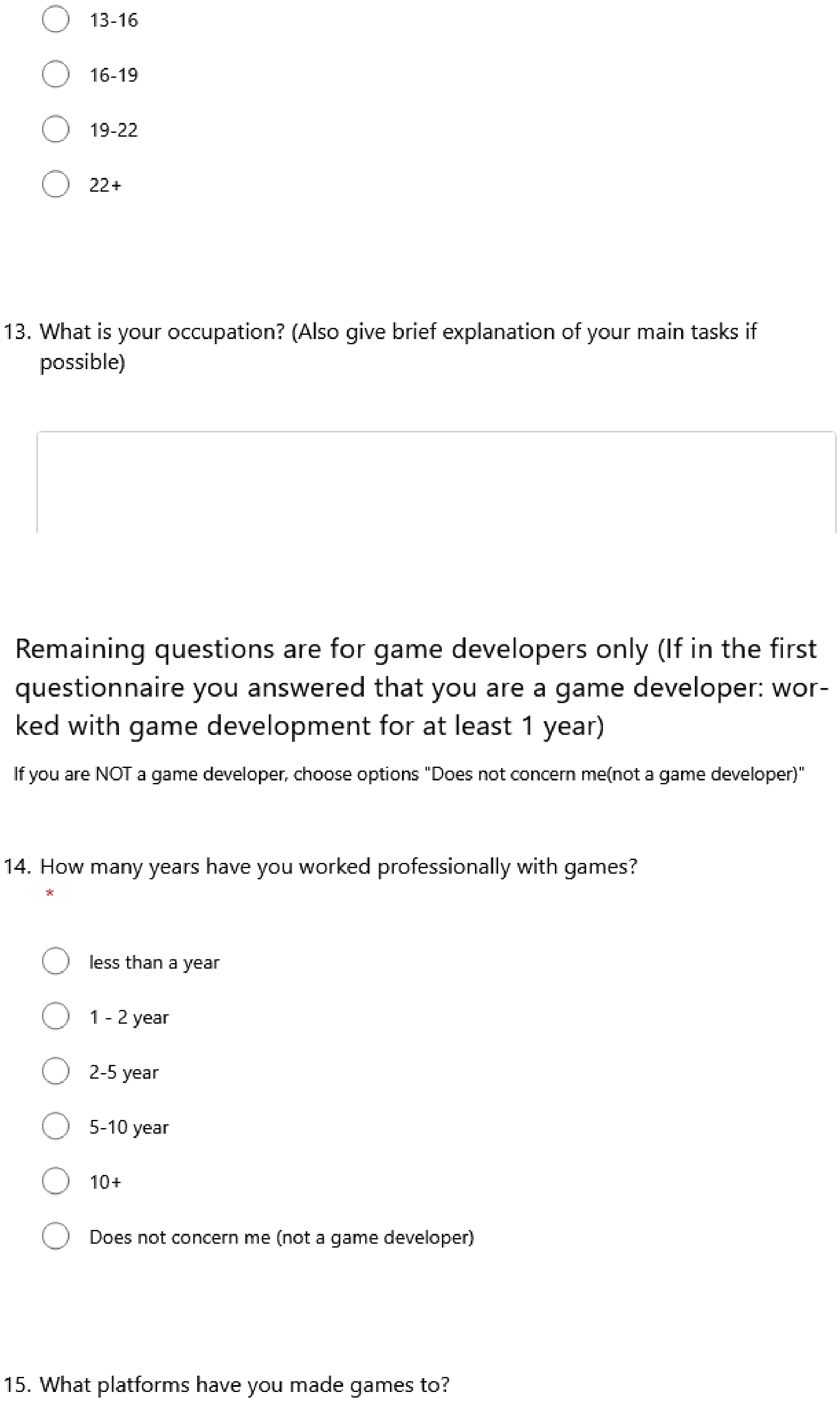
11. Main games you play
*

List MAXIMUM of 5 game names you usually play. The games can be in random order.

12. At what age did you start playing video games (approximately)?
*

○ 0-10 years old

○ 10-13



○ 13-16

○ 16-19

○ 19-22

○ 22+

13. What is your occupation? (Also give brief explanation of your main tasks if possible)

Remaining questions are for game developers only (If in the first questionnaire you answered that you are a game developer: worked with game development for at least 1 year)

If you are NOT a game developer, choose options "Does not concern me(not a game developer)"

14. How many years have you worked professionally with games? *

○ less than a year

○ 1 - 2 year

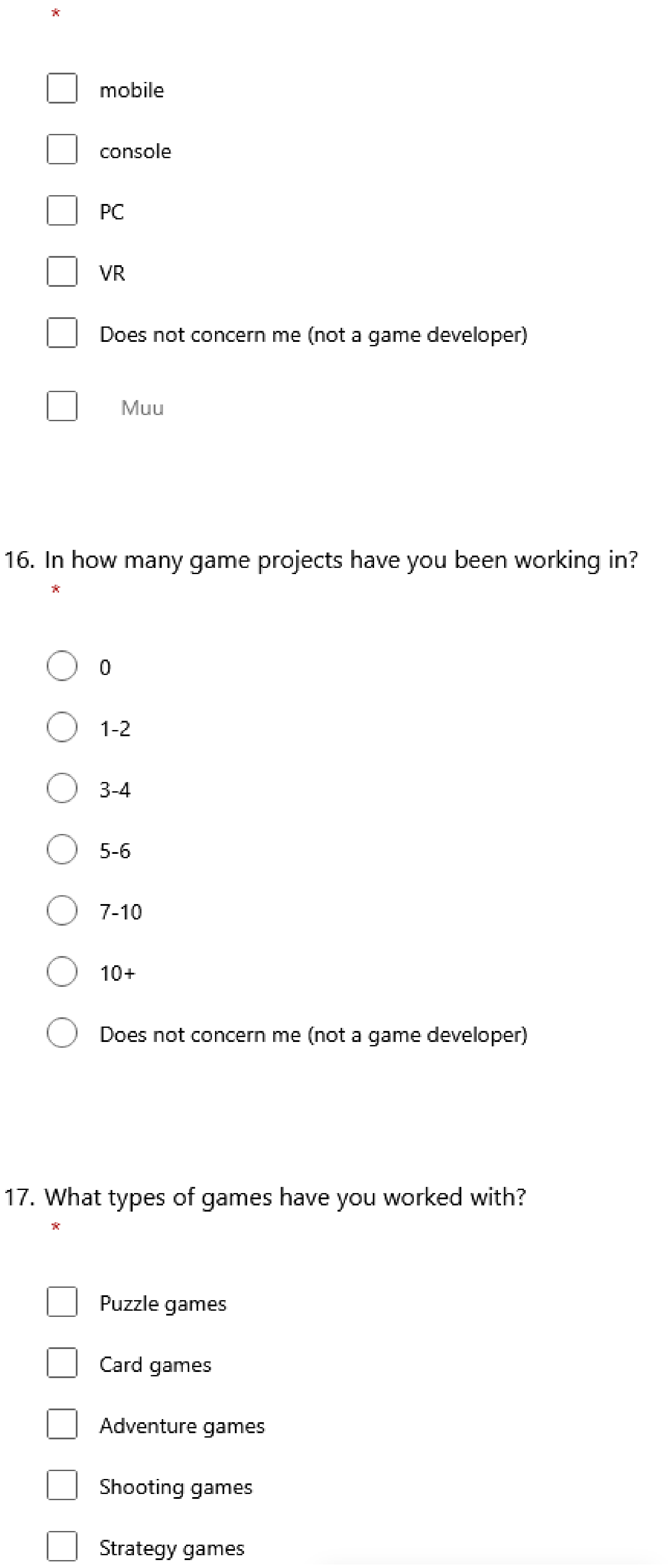
○ 2-5 year

○ 5-10 year

○ 10+

○ Does not concern me (not a game developer)

15. What platforms have you made games to?



*

□ mobile

□ console

□ PC

□ VR

□ Does not concern me (not a game developer)

□ Muu

16. In how many game projects have you been working in?

*

○ 0

○ 1-2

○ 3-4

○ 5-6

○ 7-10

○ 10+

○ Does not concern me (not a game developer)

17. What types of games have you worked with?

*

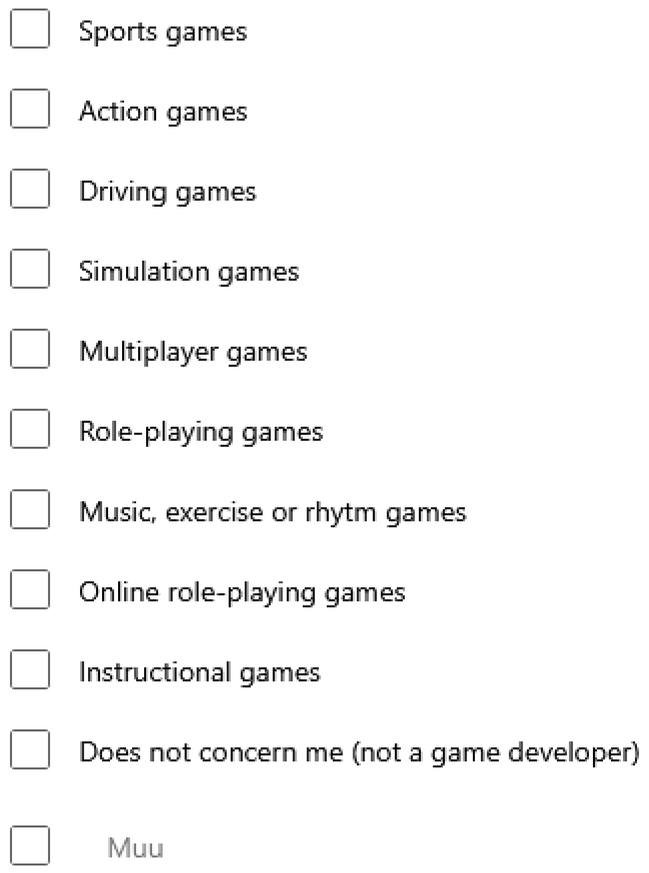
□ Puzzle games

□ Card games

□ Adventure games

□ Shooting games

□ Strategy games



□ Sports games

□ Action games

□ Driving games

□ Simulation games

□ Multiplayer games

□ Role-playing games

□ Music, exercise or rhytm games

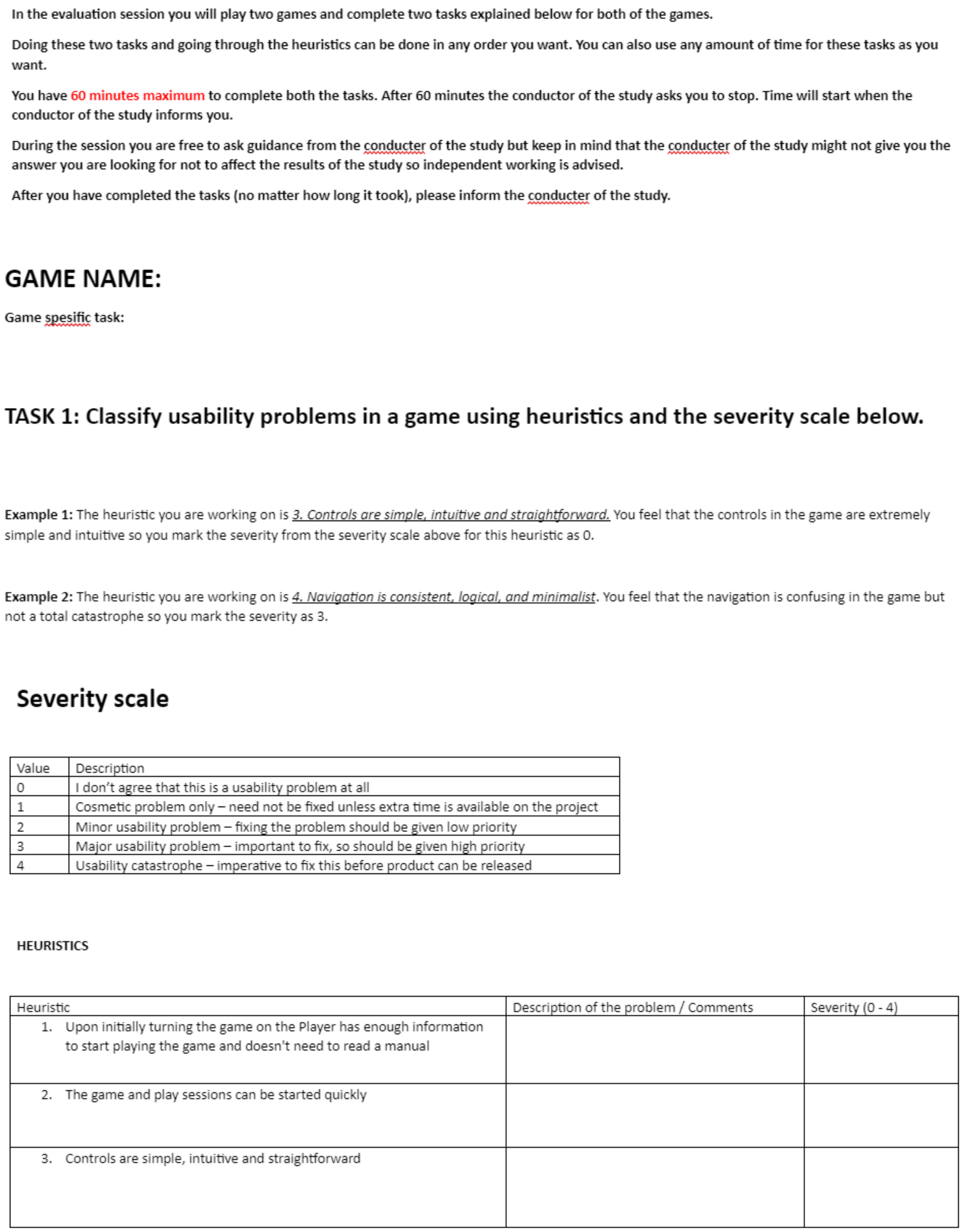
□ Online role-playing games

□ Instructional games

□ Does not concern me (not a game developer)

□ Muu

Appendix F: Evaluation form



In the evaluation session you will play two games and complete two tasks explained below for both of the games.

Doing these two tasks and going through the heuristics can be done in any order you want. You can also use any amount of time for these tasks as you want.

You have 60 minutes maximum to complete both the tasks. After 60 minutes the conductor of the study asks you to stop. Time will start when the conductor of the study informs you.

During the session you are free to ask guidance from the conducter of the study but keep in mind that the conducter of the study might not give you the answer you are looking for not to affect the results of the study so independent working is advised.

After you have completed the tasks (no matter how long it took), please inform the conducter of the study.

GAME NAME:

Game spesific task:

TASK 1: Classify usability problems in a game using heuristics and the severity scale below.

Example 1: The heuristic you are working on is *3. Controls are simple, intuitive and straightforward.* You feel that the controls in the game are extremely simple and intuitive so you mark the severity from the severity scale above for this heuristic as 0.

Example 2: The heuristic you are working on is *4. Navigation is consistent, logical, and minimalist*. You feel that the navigation is confusing in the game but not a total catastrophe so you mark the severity as 3.

Severity scale

Value	Description
0	I don't agree that this is a usability problem at all
1	Cosmetic problem only – need not be fixed unless extra time is available on the project
2	Minor usability problem – fixing the problem should be given low priority
3	Major usability problem – important to fix, so should be given high priority
4	Usability catastrophe – imperative to fix this before product can be released

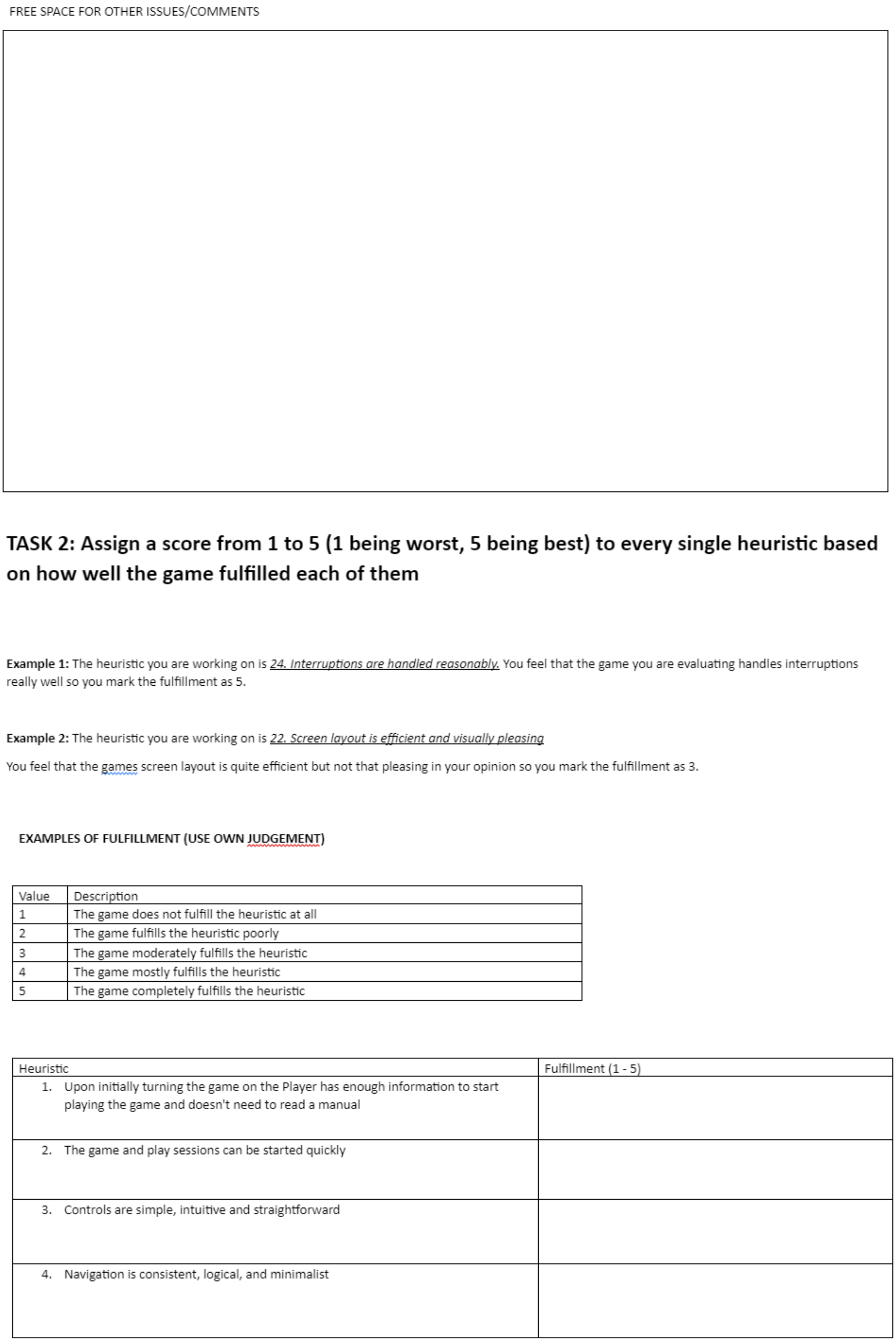
HEURISTICS

Heuristic	Description of the problem / Comments	Severity (0 - 4)
1. Upon initially turning the game on the Player has enough information to start playing the game and doesn't need to read a manual		
2. The game and play sessions can be started quickly		
3. Controls are simple, intuitive and straightforward		

4. Navigation is consistent, logical, and minimalist		
5. Provide users with information on their score/status in the game		
6. Is there a clear goal in the activity? Does the interface provide performance feedback about how close the user is to achieving the goal?		
7. Does the interface use audio and visual effects to arouse interest to the player		
8. Game provides instructions, training, and help		
9. Provide means for error prevention and recovery through the use of warning messages		
10. Provide appropriate feedback for user actions(music, sound effects, vibration)		
11. All relevant information is displayed, such as progress in the game, points, lives etc.		
12. The Player should experience the menu as a part of the game.		
13. The game supports communication and social interaction		
14. Is there audio/visual/haptic confirmation when tapping buttons or other user interface elements		
15. Are the icons clear, understandable and easy to predict what they do		
16. Is the visual indication about which items can be selected clear?		

17. The game provides information about other players		
18. Do the font types and sizes used allow for easy reading?		
19. The game's interface should be intuitive and easy to use		
20. Interfaces should be consistent in control, color, typography, and dialog design		
21. Maximizes consistency by following the trends set by the gaming community to shorten the learning curve		

22. Screen layout is efficient and visually pleasing		
23. Device UI and game UI are used for their own purposes		
24. Interruptions are handled reasonably		
25. The game accommodates with the players surroundings(lighting, noise, other people etc.)		



FREE SPACE FOR OTHER ISSUES/COMMENTS

TASK 2: Assign a score from 1 to 5 (1 being worst, 5 being best) to every single heuristic based on how well the game fulfilled each of them

Example 1: The heuristic you are working on is *24. Interruptions are handled reasonably.* You feel that the game you are evaluating handles interruptions really well so you mark the fulfillment as 5.

Example 2: The heuristic you are working on is *22. Screen layout is efficient and visually pleasing*

You feel that the games screen layout is quite efficient but not that pleasing in your opinion so you mark the fulfillment as 3.

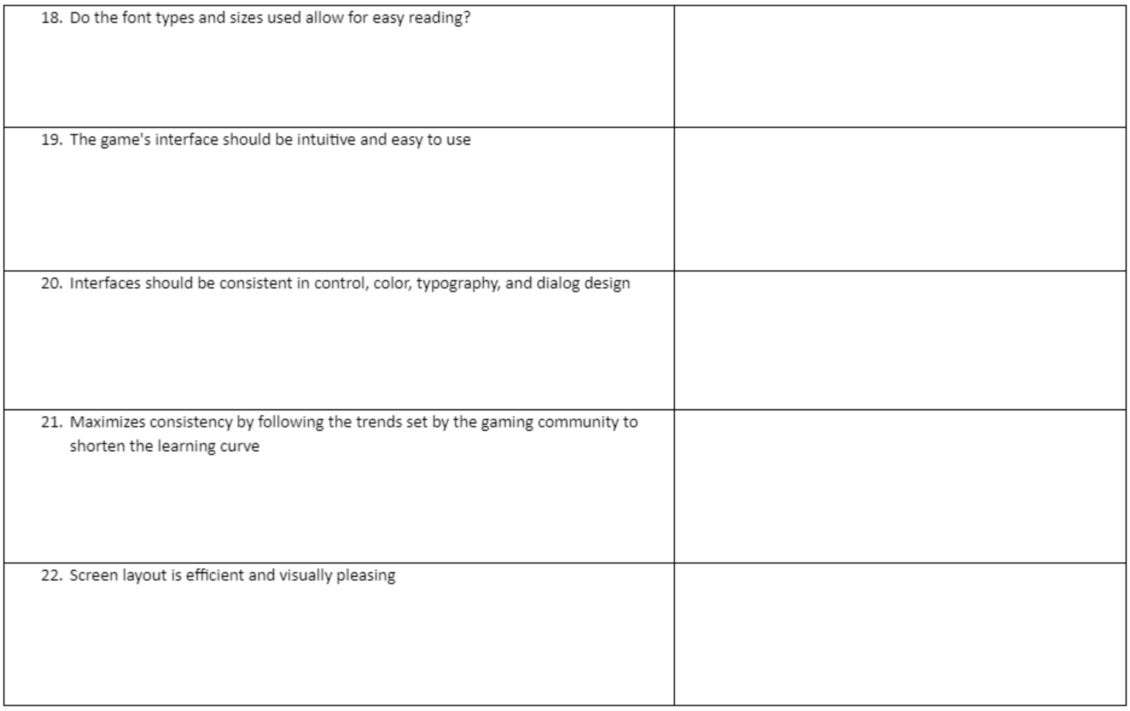
EXAMPLES OF FULFILLMENT (USE OWN JUDGEMENT)

Value	Description
1	The game does not fulfill the heuristic at all
2	The game fulfills the heuristic poorly
3	The game moderately fulfills the heuristic
4	The game mostly fulfills the heuristic
5	The game completely fulfills the heuristic

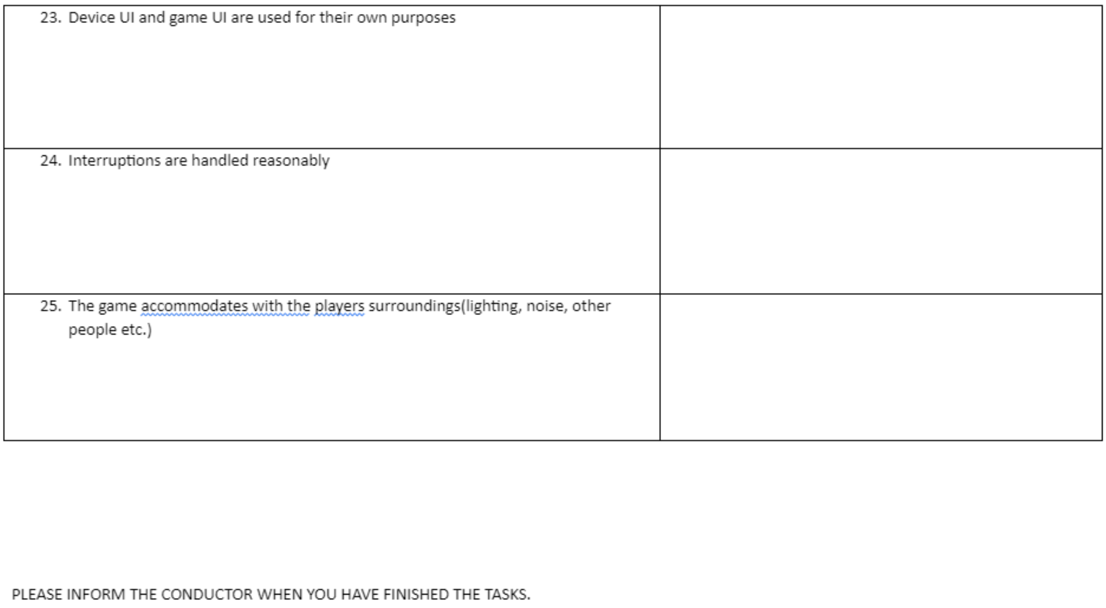
Heuristic	Fulfillment (1 - 5)
1. Upon initially turning the game on the Player has enough information to start playing the game and doesn't need to read a manual	
2. The game and play sessions can be started quickly	
3. Controls are simple, intuitive and straightforward	
4. Navigation is consistent, logical, and minimalist	

5. Provide users with information on their score/status in the game	
6. Is there a clear goal in the activity? Does the interface provide performance feedback about how close the user is to achieving the goal?	
7. Does the interface use audio and visual effects to arouse interest to the player	
8. Game provides instructions, training, and help	
9. Provide means for error prevention and recovery through the use of warning messages	
10. Provide appropriate feedback for user actions(music, sound effects, vibration)	
11. All relevant information is displayed, such as progress in the game, points, lives etc.	

12. The Player should experience the menu as a part of the game.	
13. The game supports communication and social interaction	
14. Is there audio/visual/haptic confirmation when tapping buttons or other user interface elements	
15. Are the icons clear, understandable and easy to predict what they do	
16. Is the visual indication about which items can be selected clear?	
17. The game provides information about other players	



18. Do the font types and sizes used allow for easy reading?	
19. The game's interface should be intuitive and easy to use	
20. Interfaces should be consistent in control, color, typography, and dialog design	
21. Maximizes consistency by following the trends set by the gaming community to shorten the learning curve	
22. Screen layout is efficient and visually pleasing	



23. Device UI and game UI are used for their own purposes	
24. Interruptions are handled reasonably	
25. The game accommodates with the players surroundings(lighting, noise, other people etc.)	

PLEASE INFORM THE CONDUCTOR WHEN YOU HAVE FINISHED THE TASKS.