A Framework for Usability Evaluation of Mobile Mixed Reality Games

Charley Gielkens

Supervisors:

Utrecht University: Dr. Herre van Oostendorp

Utrecht University: Dr. Christof van Nimwegen

Fraunhofer FIT: Richard Wetzel, M.A.

Thesis Number: IKU-3017281

Abstract

With the advent of smartphones, i.e. phones that have advanced sensors and data connectivity options, consumers gained access to Mobile Mixed Reality (MMR) technologies. One application of this is in Mobile Mixed Reality Games (MMRGs), games that combine the virtual and real world. Due to limited input and output capacities of smartphones, a high degree of usability is very important. This study focusses on how to evaluate the usability of MMRGs by introducing a framework to help decide which method should be used for which style of MMRG. The suggested methods are used to evaluate games that fit the style they are suggested for and the results thereof are compared to a newly introduced and validated set of heuristics for MMRGs. Although the suggested methods do not outperform the adapted heuristics in a statistically significant way, further validation is still advised as there are some limitations to the current research. Based on qualitative merits of the different methods an improved version of the framework is provided in the conclusion that should be used for any further validation.

About the cover

"Standing on the shoulders of Giants" is what science is about. Building on the past and looking at it in a new light. This is symbolized on the cover by the old map revealing information by means of Augmented Reality on the smartphone.

The locations that are shown are important locations during the writing of this document:

Utrecht University - My home university

Fraunhofer FIT - The institute that has kindly received me as a guest and supported my research

You Are Go! - Important practical experience with MMRGs at the streetgame festival in Berlin

Mensch und Computer - Learned a lot about HCI in general

Advances in Computer Entertainment 2011 - Gielkens and Wetzel (2011) got accepted at a workshop and learned a lot about games

Contents

Chapter 1

Introduction

Games have been played for millenia with evidence dating back to 2600 B.C. (Green, n.d.; Soubeyrand, 2000). Surprisingly enough though, there is no clear cut definition for what a game is (Adams, 2009). After comparing definitions from scientific literature in many different fields of study Salen and Zimmerman (2003) identified the following most used terms in the definition of "game":

Play - Entertainment in which one actively participates, rather than passively enjoys.

Pretending - "*The act of creating a notional reality in the mind.*" This reality is also called the magic circle.

Rules - A set of agreements, accepted for the duration of the game, about what the players can and cannot do.

Goal - That which is to be achieved to be considered the winner.

The parts of this definition apply to both classical games like tag, where everybody pretends it's bad to be "it", tabletop games and video games alike.

When looking from a historical point of view it has only very recently become possible to have a game mediated by a computer. To be exact, OXO was designed by Alexander Douglas in 1952 and was probably the first video game. By today's standards, this required a huge computer and tiny screen (35×16 pixels). Since then we have come a long way. Now, most people have a computer in their home (OECD, 2009) or a dedicated game console (Alexander, 2010) that allows them to play video games.

The first portable game console was introduced in 1977 by Mattel (Caoili, 2008). It could only play one game and it would not be for another two years, until the introduction of the Microvision by Milton Bradley, that handheld consoles could use interchangeable cartridges (Herman, 2001). Now, over thirty years later, people are playing games on their mobile phones. This started off with a simple game of Snake that came preinstalled on Nokia phones in 1997 (Nokia, n.d.), but has now evolved to visually stunning games that can make use of all the capabilities of smartphones. This includes information about your location using GPS, but also connecting to the internet and capturing images using the built-in camera. Games using these technologies to combine the real world and virtual world in some way are called Mixed Reality (MR) games. If the camera of the phone is used to project virtual objects on top of on an image of the real world, it's called augmented reality. When your location in the real world, as e.g. determined by using GPS, is an element of the game it is called a location based game.

Compared to current computers these smartphones have very limited input capabilities and displays. This can cause problems for both users and designers. For users it might be hard to fully use the device or software as the interface has to be very limited, while designers on the other hand may need to take into account different guidelines for creating a user-friendly application than their used to. With limited input/output options the usability of the software is of even greater importance. This means that it should be thoroughly tested preferably using a generally accepted methodology. The purpose of this thesis is to generate an evaluation framework for Mobile Mixed Reality Games (MMRGs), based on currently available usability evaluation techniques.

To achieve this, in chapter 2 the general concepts of MR and MMRGs will first be explained more thoroughly. Then a study of the state of the art of MMRGs will take place. Before proceeding to give a survey of usability evaluation methods, the difference between usability and playability is explained. Chapter 3 contains the research goal and an adapted set of usability heuristics for MMRGs. Next, in chapter 4 the framework will be introduced, based on an evaluation of the methods treated in the previous chapter. Following that the method for validating the framework will be described in chapter 5 and the results thereof will be shown in chapter 6. Finally the results will be discussed in 7.

Chapter 2

Background

2.1 Mixed and Augmented reality

Although the terms Augmented Reality (AR) and Mixed Reality (MR) are sometimes used interchangeably, this isn't entirely correct. Milgram and Kishino (1994) created a virtuality continuum to visualize the different forms in which reality and virtual reality can be mixed to create Mixed Reality.

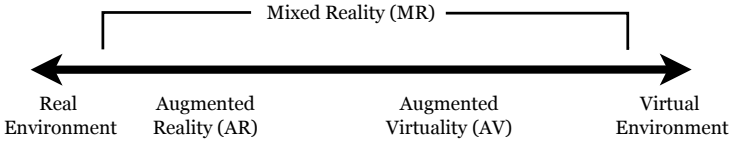


Figure 2.1: Mixed reality continuum (Milgram & Kishino, 1994).

As can be seen in figure 2.1, AR is but one part of the MR spectrum. As Henrysson and Ollila (2004) put it: "*The idea behind augmented reality (AR) is [. . .] to enhance his or her perception of the world by mixing a view of it with computer generated visual information relevant to the current context.*" This is different from Augmented Virtuality, where the main world that is perceived is generated by a computer (Milgram & Kishino, 1994) . The games on which this research will focus are generally called Mobile Mixed Reality Games, and can fall in either ends of the mixed reality continuum.

2.2 Mobile Mixed Reality Games

In order to be considered a mixed reality game, the real and virtual worlds have to be combined to some degree as per the mixed reality continuum by Milgram and Kishino (1994).

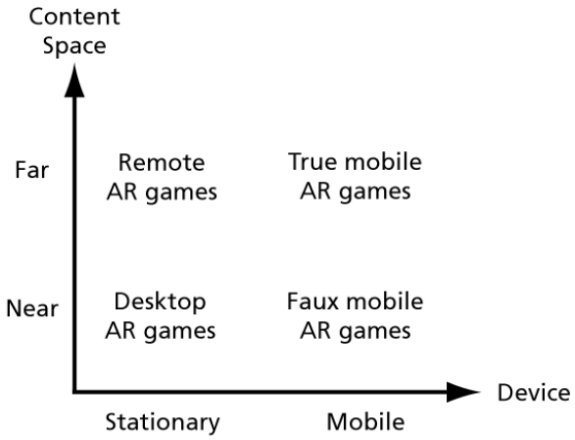


Figure 2.2: Two dimensional taxonomy of based on mobility and content space (Wetzel, Blum, Broll, & Oppermann, 2011).

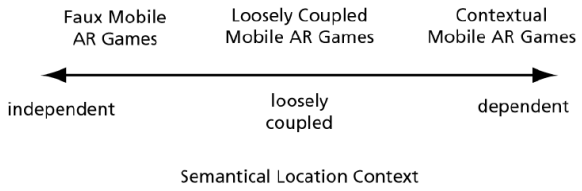


Figure 2.3: Classification of mobile AR games based on their semantic location context (Wetzel, Blum, Broll, & Oppermann, 2011).

Wetzel, Blum, Broll, and Oppermann (2011) have made a two dimensional taxonomy of Mixed Reality Games (see figure 2.2) based on mobility and content space. Mobility specifies if the device used to play

the game is stationary (like a desktop computer) or mobile (like a smartphone). Games played on mobile devices are called Mobile Mixed Reality Games. These can be subdivided again, based on how the content is provided. Faux MMRGs are played using mobile devices, but the content is still stationary because it is provided using markers, meaning it is not necessarily required to move around. True mobile games on the other hand do require players to change their location which means that a notebook is not considered as a mobile device in this context. Although they are portable, using them while walking is close to impossible.

Next to this general classification of MMRGs, one can go further still with a finer granularity.

2.3 Classifying MMRGs

When classifying MMRGs, several factors can be used. One of these proposed by Wetzel, Blum, Broll, and Oppermann (2011) is the coupling between the location where the game is played and the game content (see figure 2.3). They introduce the following kinds of AR game: Faux, Loosely Coupled and Contextual.

Faux mobile MMRGs are games that use Augmented Reality (AR) techniques and are played on a mobile device, but which have no relation to the physical place the game is played in. Contextual games on the other hand have a very strong relation to the place where they are played. Hence it takes a lot of effort to move them to a different location, if possible at all. Loosely coupled games have some relation with the location they are played at, but can also be moved relatively easily to a different location.

Gielkens (2011) introduces other ways of classifying games, like the major skill that is required (cunning or running), the amount of social interaction that is required and the persistence of the game world.

Furthermore one can also use the classic method of genres as with regular video games, e.g. role playing game or real time strategy games. These genres are not clearly defined or mutually exclusive but have historically grown as defacto standards (Adams, 2009).

The classic video game genres may not suffice as the use of MMR allows new game mechanics like checking in[1] and using peoples geographical location. This has lead to check-in games and scavenger hunts.

A special genre of games for which mobile devices and mixed reality elements are fertile breeding grounds are the so called pervasive games. This type of game blurs the boundaries that are normally present in games. These boundaries are temporal, spatial and social (Montola, 2009). This means that players can loose control over when they are playing, potentially leading to the real life and the game world getting blurred. Similarly, the spatial boundary normally denotes a playing field and which elements are part of the game, but in a pervasive game every real life item can be part of the game and the playing field may be unclear. Finally, players may not be aware who else is playing and thus non-players could unintentionally be involved in the game. This is blurring of the social boundary. The most famous example of this style of game is Killer (Montola, 2009) and Montola, Stenros, and Waern (2009) describe many more. These games needn't be mixed reality in the sense of this thesis as they can played without the use of any technology.

2.4 MMRG state of the art

The early MMRGs were often adaptations of proven concepts. There was for example a port of the famous computer game Quake to a MMR context (Thomas et al., 2002), but also Pacman (Rashid, Bamford, Coulton, Edwards, & Scheible, 2006) and Worms (Nilsen, Linton, & Looser, 2004) were adapted to this context. Sometimes the games would be a digital take on playground games like tag (Vogiazou, Raijmakers, Geelhoed, Reid, & Eisenstadt, 2006) or completely new games would be developed like The Songs of North (Lankoski et al., 2004) or TimeWarp (Herbst, Braun, McCall, & Broll, 2008). Tan and Soh (2010) give a much broader overview of the available games when they wrote their paper. They list 18 games of which only four were publicly available. Of the 18 games, seven are mobile, but of those seven only four are played on a smartphone. Furthermore, their research only looked at AR games excluding the rest of the mixed reality spectrum and they only used "*... games that are complete games.*" For this thesis games from the whole MR spectrum are interesting and it may not be a problem to have games present which have not been entirely completed. Generally speaking though uncompleted versions are

[1] Using social media to record ones presence at a certain location, e.g. a restaurant, shop or amusement park

not publicly available, meaning that although I would have at least looked into them if available there were non available.

To get a better overview of the current state of the art concerning publicly available MMRGs, an intensive study of the android market place and the internet has taken place. Research projects are left out, as these often require specialized hardware and are not publicly available. The survey has yielded a list of 29 games that were available at the time. The full descriptions of the games and screenshots can be found in appendix B and a brief summary of this study is given in table 2.1. In part, this work builds on (Gielkens & Wetzel, 2011). The reader is encouraged to look at the games in appendix B.1 in particular, as these will be referred to later on.

The 29 games that were identified were classified on the following characteristics: the way realities are mixed, whether a game focusses on cooperation or competition, the major skill required on the part of the player, game world persistence, the location coupling and the amount of social interaction that is required. Since being pervasive or not has no further bearing on the outcome of this thesis, the games will not be scored on this characteristic.

Almost all games require an active internet connection to supply the player with map or satellite images or information on the game world. Another way of mixing the real and virtual worlds is by overlaying the digital information on top of an image of the real world captured by a camera. This is a so called magic lens (Bier, Stone, Pier, Buxton, & DeRose, 1993).

Multi-player games can focus on either *competition or cooperation*, meaning that some games encourage players to outperform each other whereas other games encourage players to cooperate in order to achieve a desired goal. Single player games sometimes can be said to focus on competition by providing shared highscore lists. Cooperative single players on the other hand are a contradiction, as cooperating negates the single player status.

Persistence of the game world denotes how long it exists. Sometimes a gameworld can exist regardless of player existence, i.e. it's persistent, while in other cases it only exists when players are active.

A weak *location coupling* means that there is a link to the location at which is being played, but that it can be played anywhere without further action. No coupling means that there is no link whatsoever with the location. Finally a game that is coupled somewhat to a location can be played anywhere, but does require special action before it can be played. For example, a level or mission first needs to be created for a certain location.

Lastly the amount of *social interaction* that is required in a game will be classified as either none, a little or a lot. Activities that will be considered as social interaction are those that require two or more players to communicate verbally with each other. This means that a game where interaction between players is limited to attacking each others avatar without the possibility to communicate verbally, will be considered to have no social interaction.

2.5 Usability versus Playability

Apart from all these games for the android operating system, there are also a number of games available purely for iPhones, but also for Nintendos DS and 3DS and PS Vita. At the time of writing though, no such devices were available and hence no thorough study of the possibilities could be made.

During the short test runs with the games found in the market study, it became apparent that this technology is still in its infancy. It is for example not yet possible to create occlusion unless you have an accurate 3D model of the surroundings, batteries get drained quickly and accuracy of location determination leaves a lot to be desired.

Many of the games suffer from usability/playability issues that plagued the early video games too. The importance of usability evaluation for all video games, but lack of broad acceptance also becomes apparent from the article by Viggers (2011) in which he gives examples of failed game interfaces due to lack of usability testing.

Before discussing the different usability evaluation methods and their relevance for this research, it is important to know what playability and usability are. According to Järvinen, Heliö, and Mäyrä (2002) playability rests on four pillars:

Game	Game style					
	Map or Lens	Competition or Cooperation	Running or Cunning	Game world persistence	Location coupling	Interaction
Android Hunt	Map	Competition	Cunning	Persistent	None	Some
AR Bots	Lens	—	Cunning	One game	None	None
ConquAR	Both	Competition	Luck	Persistent	Weak	None
DJ Rivals	Other	Competition	Cunning	Persistent	None	Optional
FastFoot	Map	Competition	Running	One game	None	A lot
GPS earth defense	Map	Cooperation	Cunning	Persistent	None	Optional
GPS invaders	Map	—	Running	Persistent	None	None
Home Invasion	Map	—	Running	One game	None	None
Mister X Mobile	Map	Cooperation	Running	One game	Weak	Some
Mobeo	Map	Competition	Cunning	Persistent	Weak	Some
Nuclear	Other	—	Cunning	One game	None	None
Outbreak, Zombie Apocalypse	Map	Competition	Cunning	Persistent	Weak	Some
Parallel Kingdom	Map	Both	Cunning	Persistent	Weak	Some
Phone Bomber	Map	Competition	Cunning	Persistent	Weak	Some
PhotoFari	Lens	—	Cunning	Persistent	None	None
Portal Hunt	Lens	Both	Running	One game	Weak	Some
ScavengAR Hunt	Both	Competition	Luck	Persistent	Weak	None
SCVNGR	Map	Competition	Neither	Persistent	Weak	Unclear
Seek 'n Spell	Map	Competition	Running	One game	Weak	None
SpecTrek	Both	Competition	Running	Persistent	Weak	None
ThirdEye	Lens	Both	Cunning	Persistent	None	A lot
Tidy City	Map	Competition	Cunning	Persistent	Weak	None
Tourality/Youcatch	Map	Competition	Cunning	One game	Weak	Some
Treasure Hunters AR	Both	Competition	Cunning	Persistent	Weak	None
Underworld	Map	Competition	Cunning	Persistent	None	Optional
VuHunt	Map	Competition	Cunning	Persistent	Weak	Some
Woomba Mania	Both	Competition	Luck	Persistent	Weak	None
Zombie, Run!	Map	—	Running	One game	Weak	None

Table 2.1: Overview of the games. Names in **bold** mean they are pervasive. Based on (Gielkens & Wetzel, 2011)

Functional playability the extent to which a player is able to successfully understand an interface and use it.

Structural playability the rules of the game and dramaturgical structures implemented by the designers.

Audiovisual playability the style of graphics and audio that is used to represent the game world.

Social playability in what context is the game played and is there a feeding ground for an active community with its own culture to develop?

Nacke (2009) observes that there are many more definitions of playability. Fabricatore, Nussbaum, and Rosas (2002) state for example: "*Playability is the instantiation of the general concept of usability when applied to videogames, and it is determined by the possibility of understanding or controlling the gameplay.*" While Kücklich (2004) states that playability is "*the product of a media technology's or media text's characteristics and its user's media literacy.*" He also points out though that in the generally accepted lingo of game reviews, playability means: "*the capability to provide enjoyment for a player over an extended period of time*".

All these definitions clearly apply exclusively to games, while usability is a much broader concept as can be seen in the ISO definition: "[Usability is the] *extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*" (ISO/IEC, 1998).

In order to be able to measure usability, Sharp, Rogers, and Preece (2007, p. 20–23) break it down in to the following six goals that together form good usability:

Effectiveness How good is the application at what it is supposed to do?

Efficiency The way in which an application supports its users in performing their task.

Safety This refers to multiple elements:

- The external safety, i.e. the surroundings in which it is used
- Avoid carrying out unwanted actions accidentally
- The perceived fears by users for the consequences of possible mistakes

Utility Are the right set of tools or actions available for the task it is used for?

Learnability How easy is it to learn to work with the system?

Memorability How easy is it to remember the way the system works, once the user has learned it?

Measurable indicators that can be derived from these goals are e.g. the numbers of errors made while performing a task or the time it takes to complete a task. (ibid) What they fail to take into account though is satisfaction, which is accounted for by Sauro and Kindlund (2005) who conclude that this can be measured by using any of a number of standardized satisfaction questionnaires and refer to the questionnaires by Brooke (1996), Chin, Diehl, and Norman (1988), Kirakowski (1996) and J. Lewis (1992).

More practical and easier to interpret usability goals are provided by the ISO standard:

Effectiveness Accuracy and completeness with which users achieve specified goals.

Efficiency Resources expended in relation to the accuracy and completeness with which users achieve goals.

Satisfaction Freedom from discomfort, and positive attitudes towards the use of the product

What becomes apparent when comparing both descriptions is that the relation between usability and playability is complicated. While a high level of usability requires users to understand software very easily and quickly be able to use advanced features, this is not entirely the case for playability. Moreover, games almost always rely on withholding functionality until certain conditions have been met to create enjoyment, although they should have an easy to understand (basic) interface (Kücklich, 2004). Or as Jørgensen (2004) says, "[games should be] *easy to learn but difficult to master*".

This somewhat orthogonal relation means that the evaluation of the usability of games, MMR or otherwise, needs special attention. Therefore, this is going to be focus.

2.6 Usability evaluation

Usability has been a well-established field for quite some time and many methods have been introduced to evaluate it. Well known examples are heuristic evaluation (J. Nielsen, 1990) and cognitive walkthrough (C. Lewis, Polson, Wharton, & Rieman, 1990; Polson, Lewis, Rieman, & Wharton, 1992), the latter of which can be done concurrently or retrospectively (Haak, De Jong, & Schellens, 2003).

Next to the empirical methods like heuristic evaluation and cognitive walkthrough (Gray & Salzman, 1998), automatical, formal and informal methods are also available (ibid). Recently tools have become available for automatic usability of websites, but it still isn't commonplace. Alonso-Rios, Luis-Vazquez, Mosqueira-Rey, Moret-Bonillo, and del Rio (2009) provide a list of tools that can do this as well as introducing a new tool themselves.

What now follows is a literature study on currently available usability evaluation methods, which will serve as a background for the specialized matter of evaluating the usability in a mobile context or the usability of games which will be discussed in the next sections. The methods are dealt with in no particular order.

2.6.1 Heuristics evaluation

This technique entails using a set of heuristics which denote important usability attributes common to all software (J. Nielsen, 1990). Evaluators are asked to inspect the software using these heuristics, but otherwise have a great degree of freedom (J. Nielsen, 1994b). Advantageous is that anybody can be an evaluator as no formal usability training is required. Usability experts do find more problems than non-experts though, and people with both expertise in the field of the application and usability find more problems still (Desurvire, Kondziela, & Atwood, 1992; J. Nielsen, 1992).

The generally accepted usability heuristics for productivity software are those by J. Nielsen (1990):

Visibility of system status *The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.*

Match between system and the real world *The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.*

User control and freedom *Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.*

Consistency and standards *Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.*

Error prevention *Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.*

Recognition rather than recall *Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.*

Flexibility and efficiency of use *Accelerators – unseen by the novice user – may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.*

Aesthetic and minimalist design *Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.*

Help users recognize, diagnose and recover from errors *Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.*

Help and documentation *Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.*

2.6.2 MiLE+

The MiLE+ method introduced by Triacca, Inversini, and Bolchini (2005) is a form of heuristic evaluation. The difference with the method introduced by J. Nielsen (1990) is that evaluators look at the product from two different perspectives. On the one hand you have the application independent perspective which looks at generic features like font size or contrast between content and background. On the other hand there is the application dependent perspective. When looking from this perspective the context of use is taken into account. As an example they give the availability of content in multiple languages. This is only relevant in the context of use where the users come from different countries but has no bearing on the technical functionality of the product itself.

Since its introduction very little research has been done on or with this method. Only one other paper using it was found (Bolchini & Garzotto, 2007).

2.6.3 Use Case Evaluation

Hornbæk, Hoegh, Pedersen, and Stage (2007) introduce the Use Case Evaluation method to allow usability evaluation to take place earlier in the design process, as it is much cheaper and easier to fix issues at this stage than it is at a later stage.

The method is based on the heuristic evalation method with adaptations to the heuristics to make them more suitable for context of Use Cases. Use cases describe the user interaction with a system in order to perform exactly one particular operation (Dennis, Wixom, & Tegarden, 2004).

The method was validated by comparing its results to those of a think aloud evaluation. The use case evaluation identified 22 of the 54 issues found by regular means but also predicted problems that were not detected with traditional methods.

2.6.4 Cognitive Walkthrough

The cognitive walkthrough focuses on ease of learning, particularly by exploration rather than formal instructions. Participants are generally not users but a designer and at least one expert evaluator (Wharton, Rieman, Lewis, & Polson, 1994).

When an application requires the user to perform complex operations in order to complete their tasks this method can be most useful (Sharp et al., 2007, p. 702–705), because it "*simulates a users problem solving process at each step through the dialogue, checking if the simulated users goals and memory content can be assumed to lead to the next correct action.*" (J. Nielsen, 1992) However, this also means it takes a lot of time and effort as every response needs to be recorded and analyzed later. Furthermore in depth knowledge of the relevant cognitive processes is required (Sharp et al., 2007, p. 702–705) .

As Sharp et al. (2007) point out, this method can cost a lot of time and may be hindered by social barriers. That is why Spencer (2000) introduced a streamlined version that circumvents these problems.

2.6.5 Pluralistic Walkthrough

The pluralistic walkthrough is similar to the cognitive walkthrough. Contrary to the cognitive walkthrough though, users are invited to participate (Sharp et al., 2007, p. 705–706) in the evaluation. Moreover they are allowed to voice their opinions even before the designers and experts. Also rather than discussing actions straight away participants write down their actions and only then discuss them (Bias, 1991, 1994) .

Arguments in favor of this method are the high level of user involvement and the strong focus on a detailed level on users tasks. The major drawback is that many different people are needed making coordination difficult.

2.6.6 Formal usability inspection

A formal usability inspection is a review by the interface designer and their peers of the potential task performance of the user (Kahn & Prail, 1994). It's similar to the pluralistic walkthrough in that the reviewers step through the tasks performed by the user but it differs in who does this. In a formal inspection only usability experts are involved which increases the efficiency (Kahn & Prail, 1994). In their literature study Hollingsed and Novick (2007) did not find much research on this method after 1996 (Sawyer, Flanders, & Wixon, 1996).

2.6.7 Think-aloud

The think aloud method was introduced by Ericsson and Simon (1985) according to Sharp et al. (2007, p. 335–337). Just as the cognitive walkthrough, it is meant to examine problem solving strategies. The difference between think aloud and cognitive walkthrough is that in the former (potential) users can be involved, while in the latter this is not the case. A major problem though is that participants often fall silent and thus fail to communicate what they are actually thinking (Van Velsen, Van Der Geest, Klaassen, & Steehouder, 2008). On the other hand this is a good indication they have encountered something complicated that requires their full attention. In turn this means it is worthy of your attention to see if it isn't unnecessarily complicated.

2.6.8 Instant data analysis (IDA)

This method, introduced by Kjeldskov, Skov, and Stage (2004), is an extension to the think-aloud method. The evaluations are performed with the think aloud protocol but the analysis of the data is where the two methods part ways. With IDA the analysis is done by the test monitor, the data logger and facilitator. After four to six evaluations have taken place the test monitor and data logger conduct a brainstorm session of one hour in which they discuss the problems they observed and rate their severity. The facilitator supervises this brainstorm session, asking questions and taking notes. After the brainstorm, the facilitator goes over his notes for about 1 to 1.5 hours, making a ranked list of the problems discussed in the brainstorm. Finally all three discuss this list to ensure consensus.

The difference from many of the other methods discussed here is that it does not focus on finding as many usability problems as possible but rather aims to identify the most critical ones. The method as described here can be performed in one day, from start to finish. This is a lot faster than the regular think aloud method, in which just gathering the data could cost up to several days and analyzing it a multiple of that.

2.6.9 Feature inspection

The feature inspection method requires the evaluator to not just look at the usability of an interface element, but also at its functionality. Does the function actually do what the user needs (J. Nielsen, 1994b, p.6)? Unfortunately more information could not be found about this method.

2.6.10 Consistency inspection

Wixon, Jones, Tse, and Casaday (1994) developed the consistency inspection method based on the observation that software is becoming ever more integrated, i.e. packages combine functionalities that were separate before. This means that the different components have to be the same across every system which is what this method focusses on. The inspection is done by a user interface expert who creates a document describing his findings. This is then discussed with a team consisting of developers from every component group.

2.6.11 GOMS

The GOMS method was introduced in 1983 and it stands for goals, operators, methods and selection rules (Card, Moran, & Newell, 1983). Unlike the previous methods, GOMS is a predictive model rather than an inspection method. This means that performance is measured by making use of experts and formulas instead of actual users (Sharp et al., 2007, p. 706–708).

2.6.12 Interaction logs

Using log files to track user interaction is mainly popular and easy to achieve when the product that is to be evaluated is web based because web servers write log files about its transactions anyway. In research the most popular subject to study by far is the website of a (university) library (Allen, 2002; Asunka, Chae, Hughes, & Natriello, 2009; Ghaphery, 2005; Jamali, Nicholas, & Huntington, 2005; Jeng, 2005; Spiteri, Tarulli, & Graybeal, 2010) .

The collected log files need to be analyzed in order to get information out of the data. For this many packages are available, like the open source projects webalizer[2] and awstats[3]. A more comprehensive list can be found on Wikipedia.[4] None of these tools however generate readymade usability scores or info. That is for an expert to extract.

When evaluating standalone applications one would either have to alter the program itself or use an application that captures the user input at the operating system level in order to generate log files (Okada & Asahi, 1999). A third option would be to capture the screen output and later analyze this, but this is very time consuming and inefficient compared to automatic logging.

[2] `http://www.webalizer.org/`
[3] `http://awstats.sourceforge.net/`
[4] `http://en.wikipedia.org/wiki/List_of_web_analytics_software`, accessed July 14, 2011

2.6.13 Remote testing

Next to interaction logs (Fernandez, Insfran, & Abrahão, 2011) remote usability testing can be done by sharing the screen via the internet and capturing software to see and record what a remote participant does (Thompson, Rozanski, & Haake, 2004). Thompson et al. (2004) found no significant difference in task performance although their statistics are questionable as each condition had only five participants. The amount of usability problems identified was slightly higher in the remote condition which they found to be comparable to earlier research by Tullis, Fleischman, McNulty, Cianchette, and Bergel (2002).

2.6.14 Diaries

In long-term studies it can be advantageous to use diaries as a method of collecting data, because they do not require special training, equipment or lots of resources (Sharp et al., 2007, p. 338–340). A disadvantage of filling out diaries is that the sheer act of filling one out may influence the actions that have been taken and what is written in the diary (Czerwinski, Horvitz, & Wilhite, 2004). When it is not possible to automatically log user interaction, for whatever reason, diaries are a good alternative (Tomitsch, Singh, & Javadian, 2010).

Two types of diary study are generally found in usability studies: feedback studies and elicitation studies. In the former predefined questions about events are answered, while in the latter the media captured by the participants is used to start a discussion afterwards (Carter & Mankoff, 2005).

With any type of diary sustaining the participants engagement is a challenge and with paper diaries even more so than with digital diaries (Carter & Mankoff, 2005). Typically participants get reminders or reimbursements to improve their response rate (Palen & Salzman, 2002). Tomitsch et al. (2010) confirmed the importance of good instructions as their paper diaries unexpectedly had a better participation rate than their digital diaries due to lack of proper instructions for the latter. They also point out that a good design is important.

Palen and Salzman (2002) introduce a different method for recording information in a diary: using a voice-mail diary. The strength of this method is that participants can use it more easily in a mobile context than a paper diary and are not required to carry around an extra item for the sole purpose of recording information.

Another way of capturing information for a diary is introduced by Brown, Sellen, and O'Hara (2000). They have the participants take a photograph whenever they want to note down information or bring an item with them. This is a very useful method in a multicultural study because the saying "One picture says more than one thousand words" is quite often true. Pictures can be a great tool to explain or communicate cultural differences. They also allow you to easily observe the natural contexts of use for mobile applications where it would otherwise not be feasible (Sampanes, Snyder, Rampoldi-Hnilo, & White, 2011).

2.6.15 Metaphors of Human Thinking (MOT)

Normally in human computer interaction the word metaphor is used to indicate an interface metaphor, e.g. the desktop metaphor (Johnson et al., 1989). In this case though the metaphors are not even part of the design. The metaphors are meant "*to support the evaluator-systems designer in a focused study of how well certain important aspects of human thinking are taken into account in the user interface under inspection.*" (Hornbæk & Frøkjær, 2004a) The metaphors they go on to describe are introduced to the HCI community in(Frøkjær & Hornbæk, 2002) and in (Hornbæk & Frøkjær, 2004a) is explained how to utilize them in a usability evaluation context. The metaphors and their respective considerations in usability inspection are (Hornbæk & Frøkjær, 2004b):

Habit formation is like a landscape eroded by water *Are existing habits supported? Can effective new habits be developed, when necessary or appropriate? Can the user use common key combinations? Is it possible for the user to predict the layout and functioning of the interface?*

Thinking as a stream of thought *Is the flow of users' thoughts supported in the interface by recognizability, stability and continuity? Does the application make visible and easily accessible interface elements that relate to the anchor points of users thinking about their tasks? Does the application help users to resume interrupted tasks?*

Awareness as a jumping octopus [5] *Are users' associations supported through flexible means of focusing within a stable context? Do users associate interface elements with the actions and objects they represent? Can words in the interface be expected to create useful associations for the user? Can the user switch flexibly between different parts of the interface?*

Utterances as splashes over water *Are changing and incomplete utterances supported by the interface? Are alternative ways of expressing the same information available? Are the interpretations of users' input in the application made clear? Does the application make a wider interpretation of users' input than users intend or are aware of?*

Knowing as a building site in progress *Are users forced by the application to depend on complete or accurate knowledge? Is it required that users pay special attention to technical or configuration details before beginning to work? Do more complex tasks build on the knowledge users have acquired from simpler tasks? Are users supported in remembering and understanding information in the application?*

The core difference between a regular heuristic evaluation and a MOT evaluation is the active interpretation of the complex guidelines that are offered by MOT, where in a heuristic evaluation simple guidelines with straight forward interpretations are used (Hornbæk & Frøkjær, 2004a).

Preliminary experiments suggest that MOT is a valuable addition to the usability evaluation toolbox as it performs at least as well as heuristic evaluation (Frøkjær & Hornbæk, 2008). Furthermore it identifies different kinds of problems than heuristic evaluation does. The ones found by MOT are generally more complex to repair and more severe for users. The problems found by heuristic evaluation were from a broader spectrum, but generally classified as cosmetic (Hornbæk & Frøkjær, 2004a) .

When compared to cognitive walkthrough (Frøkjær & Hornbæk, 2008; Hornbæk & Frøkjær, 2004b), MOT appears to be the better of the two as it identified 30% more problems. Compared to the think-aloud method though, there isn't really a discernible difference (Frøkjær & Hornbæk, 2008) .

A criticism on this method is that it may be hard to understand the abstract metaphors and hence the evaluator might miss important usability issues (Hvannberg, Law, & Larusdottir, 2007) .

2.6.16 Systematic Usability Evaluation (SUE)

SUE uses abstract tasks for the evaluation of interfaces and was created in the course of defining a general framework for usability evaluation (Matera, Costabile, Garzotto, & Paolini, 2002). The main idea behind the method is that an interface should be evaluated at different levels. Up to this point the authors have addressed two levels. Level one focuses on the presentation layer of the interactive application and level two focuses on features that are specific to a certain application category. For each level specialized conceptual tools, like application models, abstract tasks and usability attributes, need to be defined. An abstract task is an operational activity that is formulated independently from any one application, both in task description and its references to interface elements.

Another core idea to the method is that combining inspection and user-based evaluation yields better results than either method alone. The authors suggest that in order to optimize user resources, first an inspection should take place and the problems found should be addressed before getting to the user evaluation. To reduce the need to come together Ardito, Lanzilotti, Buono, and Piccinno (2006) introduce a web based tool to support usability inspection using this method. It uses dynamic websites to collect the data and notifies evaluators via e-mail of activity.

2.6.17 Perspective based inspection

When using the perspective based inspection method (Z. Zhang, Basili, & Shneiderman, 1999) each evaluator looks at the interface from only one perspective. A perspective consists of a specific point of view, a list of inspection questions representative of the usability issues to check for and a specific procedure for conducting the inspection. The authors have chosen for this approach because it is very difficult, even for an expert, to regard an interface from many different perspectives at once. The perspectives provided

[5] This is a partial quote of Naur (1995). The full quote is: "*The mental activity is like a jumping octopus in a pile of rags.*" It is meant to denote the locus of attention by the body of the octopus, the fringes by its arms and the changing nature of human thought by the jumping.

by the authors are:

Novice use The user's knowledge and experience do not tell the user how to use the system to achieve the goal.

Expert use The user knows how to use the system but prefers to achieve the goal efficiently and easily, or wants to achieve higher goals.

Error handling the user has a problem with the effect achieved by the previous action and needs to resolve the problem.

According to the authors, these three perspectives can be derived from answering the questions whether or not the user knows how to achieve their goal and if they continue to perform the action correctly. This means that both novice and expert use only cover correctly executed actions. What they fail to take into account though, is that it is likely that novices and experts solve problems differently too.

Z. Zhang et al. (1999) also define usability goals for the different perspectives. The goal for novice use is that the fundamental task must be achievable with the minimum knowledge. One of the goals for expert use is that the interface must be customizable to their wishes. The full list can be found in the article.

2.6.18 Pattern-based usability inspection

Usability patterns are very much like design patterns (Gamma, Helm, Johnson, & Vlissides, 1994) as both are a grouping of collaborating classes that provide a solution to a commonly occurring problem (Dennis et al., 2004), or as Schmettow (2005) puts it: "[Usability patterns] *describe well established solutions together with preconditions and rationales for ergonomic design of user interfaces.*"

Patterns have several advantages over heuristics. First, they help to identify the correct pattern for a given situation, because the required preconditions and rationales for a solution are included in their description. Furthermore, it is easier to understand them because the descriptions of both problem and solution are more problem-oriented and verbose. Finally, they give concrete solutions to problems which makes them more valuable when making design recommendations (Schmettow, 2005) .

Trying to validate the method Schmettow and Niebuhr (2007) found that they currently needed 10 evaluators to find 80% of the usability problems. This is far worse than heuristic evaluation which needs only 4 to 5 evaluators for the same result (J. Nielsen & Landauer, 1993). As they also point out, the evaluators were not experts and the sample was too small for statistical validation. Both these factors may have had an adverse effect on the outcome.

2.6.19 Interviews

Another evaluation technique are interviews (Sharp et al., 2007, p. 298–308). These can be done with one or multiple interviewees at a time. Open ended questions and semi-structured or even unstructured are most appropriate as these allow a greater degree of freedom and the possibility to react to unexpected situations.

It is also suggested (ibid) to enrich the interview sessions with extra material, like prototypes or screenshots or use e.g. diaries as the basis for an interview.

2.7 Mobile usability evaluation

Some of the methods mentioned up until now can be used in a mobile context although most were not designed with this in mind. As a result it is often very hard to use them in the field (Brewster, 2002; C. Nielsen, 1998). As an example one can look at a classic think-aloud setup. This requires multiple observers of which one is hidden from view and the recording of all the interactions of the participant. When a participant is walking around outside the boundaries of a room it quickly becomes harder if not impossible for a hidden observer to do their work. Also recording user interactions and system output on a mobile device is very complicated to do unobtrusively.

From here on the term "context" will be used quite often. It is defined by Jensen (2009) as follows: "*Context is the sum of relevant factors that characterize the situation of a user and an application, where*

relevancy implies that these factors have significant impact on the user's experience when interacting with that application in that situation."

The need for contextual usability testing with mobile devices is a point of discussion in itself. Many studies have been conducted to see if usability testing of mobile devices and applications should be done in the wild or in a laboratory setting. Unfortunately there is no consensus. For each source claiming that either of the methods is best, there is another claiming the opposite or a lack of difference. See for example (Duh, Tan, & Chen, 2006; Kallio & Kaikkonen, 2005; Kjeldskov & Stage, 2004; C. Nielsen, 1998; C. M. Nielsen, Overgaard, Pedersen, Stage, & Stenild, 2006; Waterson, Landay, & Matthews, 2002) However, all these studies focus either on the mobile device itself or basic applications like calling or sending text messages. Although all of these can be done while moving it is also a perfectly reasonable scenario to perform these actions while stationary. Many MMRGs on the other hand are specifically meant to be played while moving to a greater or lesser extent. This leads me to believe that MMRGs must also be tested in the wild, not just stationary in a laboratory setting. This is also supported by the framework for the design and implementation of usability testing of mobile applications by D. Zhang and Adipat (2005) (see figure 2.4) as the real world context can be very important.

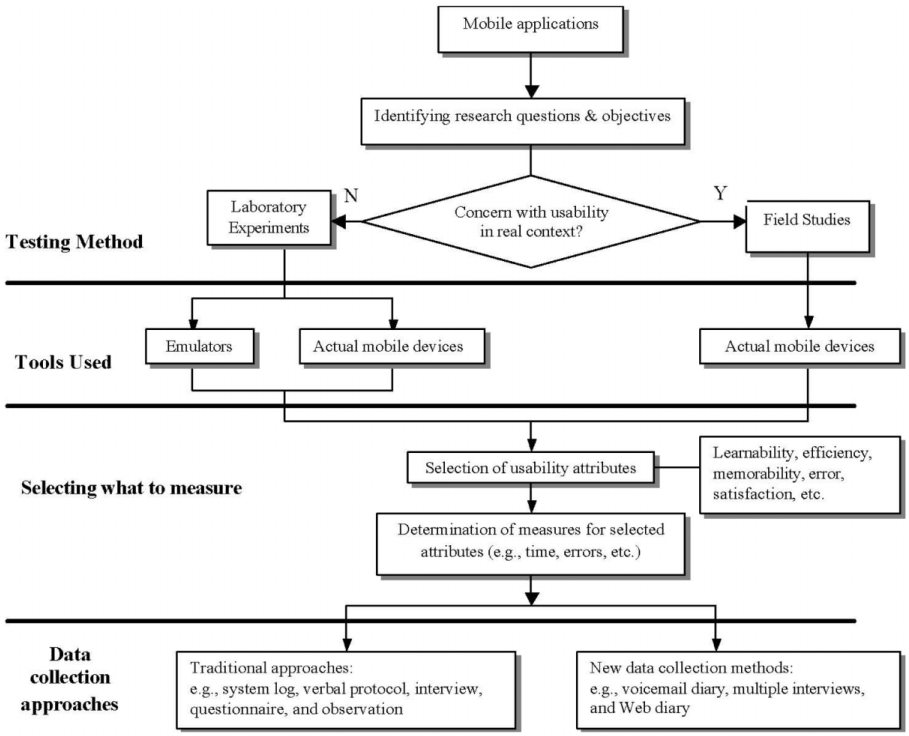


Figure 2.4: A framework for the design and implementation of usability testing of mobile applications (D. Zhang & Adipat, 2005) .

In a review of one hundred papers on mobile usability Coursaris and Kim (2011) found that the main constructs that are measured are efficiency, effectiveness and satisfactions. They also conclude that it would be very good to expand the body of literature for mobile usability by further investigating the technology beyond the interface, what user characteristics are most influential on perceived usability, the impact of task complexity and interactivity on mobile usability and what the influence is of environmental conditions on usability. Especially the last recommendation is entirely in line with the divided body of research on whether or not to evaluate the usability of mobile devices and applications in the field.

2.7.1 Contextual evaluation

Po, Howard, Vetere, and Skov (2004) introduce two methods for contextual evaluation. The first, heuristic walkthrough, combines heuristics and scenarios of use but takes place in a laboratory. Contextual relevance is introduced by means of the scenario.

The second method introduced by Po et al. (2004) is contextual walkthrough. In this method heuristics are combined with scenarios of use, but now the evaluation takes place in the correct context of use.

Although the authors report encouraging findings for both methods, a follow up study by Varsaluoma (2009) contradicts this. He has found that adding contextual information in either way made no difference to the outcome when compared to the original method.

A different method suitable for contextual evaluation is capturing the users behavior. Before smartphones became commonplace different tools were developped, for example the SocioXensor by Mulder, Ter Hofte, and Kort (2005) for PDA's, MyExperience running on Windows Mobile 5.0, by Froehlich, Chen, Consolvo, Harrison, and Landay (2007) and the ContextPhone by Raento, Oulasvirta, Petit, and Toivonen (2005) which ran on Symbian OS. It appears that of these only MyExperience has been developped somewhat further (Froehlich, 2009) though it still only runs on Windows Mobile 2005. More recently, Jensen (2009) has developed a tool called RECON to capture mobile interaction in context on Windows Mobile 5.0. For the dominant smartphone operating systems iOS, for Apples iPhone, Shepard, Rahmati, Tossell, Zhong, and Kortum (2011) developed LiveLab. Although this application is not directly designed to capture user input, it seems to be within the realm of possibilities. A drawback is that the telephone has to be "jailbroken", i.e. root access has to be acquired in a manner that does not conform with the usage policy of the operating system. In a paper that has not been accepted at the time of writing Bell, McDiarmid, and Irvine (2011) introduce an application they have created to capture user interaction on a smartphone running Android. It captures both the user action and the screen output just before, during and after the interaction.

The voice-mail diary (Palen & Salzman, 2002), pocket and web diaries (Kim, Kim, Lee, Chae, & Choi, 2002) that D. Zhang and Adipat (2005) mention as viable options for contextual data recording (see figure 2.4) still have drawbacks when applied to games. They are very likely to break the flow of the game as they require the user to stop playing and actively do something else (Adams, 2009). This is very likely to influence the results.

Other known problems with current methods are the highly dynamic nature of the environment when testing in the wild, the limited battery life of the devices being used, the large amount of time needed to perform enough evaluations for quantitative research and the difficulty of capturing user input and/or the output of the device. Some of these can be addressed to a certain degree, while others are just woven into the fiber of this type of game.

2.8 Usability of games

Games are used in a very different manner than other applications, because their purpose is orthogonal to the purpose of most other applications. Most applications are used to achieve a goal as quickly as possible, while games are used as a form of entertainment. This means that although some of the evalation methods introduced before may be valuable tools, others will be useless because of the basic assumptions that underlie them. This goes for every kind of video game, MMR or otherwise.

Pinelle, Wong, and Stach (2008a) argue that heuristic evaluation is a good way to investigate the usability of games, because it makes no assumptions on task structure and it leaves considerable freedom for the evaluator. Due to the different nature, not all of the classical heuristics by J. Nielsen (1990) may apply. Therefore Pinelle et al. (2008a) introduce specific usability heuristics for games, some of which match with the ones by J. Nielsen (1990). The Pinelle heuristics are:

Provide consistent responses to the users actions Actions should have predictable, appropriate reactions.

Allow users to customize video and audio settings, difficulty and game speed. The default settings are not suitable for everyone

Provide predictable and reasonable behavior from computer controlled units. If the AI takes over certain tasks, like pathfinding for units, make sure it does this correctly, so that user does not need to issue unnecessary extra commands.

Provide unobstructed views that are appropriate for the users current actions. One view might not be appropriate for all situations. Make sure the player can always see everything he needs to.

Allow users to skip non-playable and frequently repeated content. Cut-scenes are nice, but they interfere with the gameplay.

Provide intuitive and customizable input mappings. Quick responses are often required, so the player should not need to think about what to do. If the input is customizable support the common input devices like keyboard and mouse.

Provide controls that are easy to manage, and that have an appropriate level of sensitivity and responsiveness. Controls should respond neither too slow nor too fast. In simulations reactions should mimic the real world.

Provide users with information on game status. Without relevant information it is very hard to make appropriate decisions about how to proceed with the game.

Provide instructions, training and help. Games are often difficult to master. When appropriate, interactive training should be provided.

Provide visual representations that are easy to interpret and that minimize the need for micromanagement. Make sure visual representations do not occlude the view or force the user to do micromanagement.

They also found that each game genre violated different heuristics most frequently. For example, most issues with adventure games concern consistency, where as action games mainly have issues with the controls (Pinelle, Wong, & Stach, 2008b).

All this research however has focussed on single player games which are played on personal computers or dedicated stationary consoles. Mobile games in general bring a different way of interacting with the system, requiring the introduced heuristics to be reevaluated. Korhonen and Koivisto (2006) introduce a list of heuristics suitable for mobile games and Wetzel, Blum, Broll, and Oppermann (2011) take it one step further and introduce usability guidelines specifically for MMRGs:

Make the technology part of the game When the technology is not part of the story, it hurts the willing suspension of disbelief.

Keep the interaction simple Interaction should always be easy, but for MMRGs even more so because out there players do not have access to playing aids.

Take display properties into account Screens of mobile devices are far from perfect: they are small and outside the reflection of sunlight may make them hard to read.

Take tracking characteristics into account In MMRGs the players position and that of virtual objects is tracked. Every method available has strong and weak points that must be taken into account.

Avoid occlusion rich areas Correct occlusion can only be achieved with accurate virtual models. These may exist for buildings, but for other things it may be a problem

Design seamfully and for disconnection When playing outside it is more than likely that players will encounter areas with either bad connectivity to the internet or a bad GPS signal. This should not cause problems for the player.

These guidelines touch upon important subjects, but are only suitable to evaluate the usability of the MMR part of the game. With nothing else than these guidelines you will be hard pressed to find more general usability issues that could also occur with interfaces on static devices.

Some of these guidelines show overlap again with the ones by Pinelle et al. (2008a) and J. Nielsen (1990). What they do not take into account though is the highly social nature of some MMRGs. Although all these guidelines are valid for multiplayer MMRGs, separate heuristics are necessary that apply to the elements that are associated with the multiplayer elements. Pinelle, Wong, Stach, and Gutwin (2009) introduce heuristics that do just this:

Simple session management Allow the player to either start or find and join appropriate games.

Flexible matchmaking Provide features to match players based on interests.

Appropriate communication tools provide communication features that accommodate the demands of gameplay.

Support coordination Allow players to coordinate actions in appropriate game modes.

Meaningful awareness information Provide meaningful information about the players.

Identifiable avatars Use noticeable and distinct avatars that have intuitive information mappings.

Training for beginners Allow novices to practice without pressure from experts.

Support social interaction Provide support for both planned and opportunistic social interactions.

Reduce game-based delays Reduce temporal dependencies between players in order to minimize interaction delays.

Manage bad behavior Provide technical and social solutions for managing unwanted behavior.

Chapter 3

Research goal

3.1 Social relevance of MMRGs

Or more plainly, the answer to the question: "Games are good and well, but aren't they for kids? What does this bring for adults?"

Apart from simply providing entertainment, mixed reality games can be used for all sorts of purposes like helping stroke victims recover (Burke et al., 2010), providing novel and interesting ways for team building exercises (Bulman, Crabtree, Gower, & Oldroyd, 2006), introducing players to their surroundings (Wetzel, Blum, Feng, Oppermann, & Straeubig, 2011), serving as a motivator for physical exercise[1] or as a teaching tool (Facer et al., 2004).

As an example, I will now describe a play session of the game Mister X (see section B.1.1) I experienced during the street game festival You Are Go! in Berlin[2]. The players in this session were adults, with estimated ages ranging between 20 - 50 years which matches closely with the findings presented in The Economist (*All the worlds a game*, 2011).

> We stood together in an open space, the six of us. The first time we met was five minutes ago. For most of us, we had been in Berlin no longer than 24 hours, and were completely unfamiliar with the area. After I set up the game and the other five players joined the game, one of us was appointed "Mister X". As soon as this happened he started running for his life, trying to avoid us at all costs. Only after two minutes were we allowed to follow him. In an attempt at tactics we tried to surround our Mister X by splitting up and heading in the direction of his last known location which was shown on the map.
>
> By the time the first person got there he, of course, was long gone. Baffled we stood there, because from every street leading to the place we were coming. How could he have passed us?
>
> Half a minute later Mister X's position was updated on the map and it became clear he must have passed us by hiding somewhere, somehow. Not withstanding our last botched attempt at tactics, we tried the same thing again. Him being completely on the other side of the map gave him quite some confidence, as I saw him stopping at our bags to have a quick drink and snatch a new sweater in an attempt to hinder recognition. If we had entered our phone numbers into the game, this would've been the point at which I called the other detectives. Unfortunately we didn't do that, so all that I could do was run for dear life in an attempt to catch Mister X. What followed was a frenzied chase first through empty streets, then into a crowd where I lost him. Suddenly a dodgy stranger with a hood and sunglasses on tried to sneak past me, but started running when he noticed I was looking at him. No doubt about it this was our Mister X, the posture fit. Although at this point my breath was limited, the thrill of the chase gave me the adrenaline kick I needed to push through and follow him for another couple hundred of meters until his breath gave out before mine.

[1] `https://www.zombiesrungame.com/`
[2] `http://invisibleplayground.com/#you-are-go`

> After this exciting game we all returned to our base, and the other detectives rejoined us and we shared "war stories" about the game how we saw (or didn't) Mister X when he was hiding on the other side of a car when two detectives walked past. What also received a lot comment though, was the unclarity of the icons used for the special items we could all use. None of the other five could guess from the icons what they would do and some didn't even notice they had these items available to them.

This example shows a couple of things. First, there is the physical exercise involved in the game. As one of players in a different session remarked: "Wow, when I saw Mister X I just started running, because I really wanted to catch him. I was all excited about it, even though I don't know why." This shows that a game can be a great stimulator to get some exercise.

Secondly, in this example we all got to know the surroundings in which we played quite well. We discovered all kinds of nooks and crannies in which we could hide or could use as a short cut.

Furthermore, the recounting of the game and the chases are a great way to bond even though one of the six was on the other team.

Lastly, it became apparent the designers of the game had not spent a great deal of attention on the usability of their game. The interface elements that would have shown the user their special items sometimes went unnoticed, because the contrast between it and the background was too low to be seen in bright sunlight. If it was found at all, the icons representing the special powers did not speak for themselves. The users either had to find out by trial and error, or by accessing a help file whose existence was not very clearly indicated. The first method isn't very efficient, as the effects of the items are generally only visible to the other team and the latter is bad because in a game that focuses on paying attention to your surroundings and running it is very bad practice to stop and first try to find a help file and then spend more time to reading it.

Another reason why research like this is relevant, is the size of the gaming industry in general. Gartner[3] estimated the total spending on video games to top 74 billion US Dollars in 2011, which puts it on par with the GDP of countries like Ghana, Guatamala and Kenya according to the CIA World Factbook (CIA, 2012).

3.2 Research goal

From the previous section it becomes clear that as with any interface, the usability testing of MMRGs is no superfluous luxury. What is shown in the literature section though is that although research has been done on the usability of games and mobile devices separately, research where the two meet is still very limited. Especially when mixed reality is added to the mix. Hence, there are currently no appropriate, generally accepted techniques available to evaluate the usability of MMRGs. The methods that do exist are either likely to break the flow of the game or do not take the dynamic nature of these games into account. Based on that, the research goal is formulated.

Research goal: *Create a standardized framework for the evaluation of the usability of MMRGs, based on currently available usability evaluations methods.*

3.3 Approaching the research goal

From the extensive study of games by Gielkens (2011) several archetypes emerged. There are games that focus a lot on running (like Mister X Mobile and Portal Hunt), while others focus much more on how cunning you are (like Tidy City). Some games can last only minutes (like Seek 'n Spell) while others can go on indefinately (like Parallel Kingdom). Of course the well known archetypes of single player (Tidy City) and multiplayer (Portal Hunt) are also present. Each archetype or genre has its own idiosyncrasies that influence how one can best observe players.

Since every usability evaluation method also has its own strengths and weaknesses that make it more applicable for certain situations than for others, the main hypothesis to be tested to accomplish the research goal is:

[3] `http://www.gartner.com/it/page.jsp?id=1737414` accessed April 8, 2012

Hypothesis 1 *Depending on the style of the mobile mixed reality game, different usability evaluation methods will be more suitable*

A method is considered suitable if it finds a similar amount of issues or more as a heuristic evaluation as this is the preferred method for evaluating the usability of regular video games at this point (Desurvire, Caplan, & Toth, 2004; Pinelle et al., 2008a). Also taken into consideration will be the severity of the identified issues (J. Nielsen, 1995).

In order to test this hypothesis all the evaluation methods mentioned in section 2.6 will be evaluated for their suitability in relevant scenarios for the evaluation of MMRGs. Based on the outcome thereof a framework will be created suggesting the most appropriate methods for distinct situations. These suggestions will then be formulated as hypotheses which in turn will be tested.

If any of the newly formed hypotheses prove to be clearly wrong, the framework will be revised accordingly and presented in the discussion.

3.4 Defining usability for MMRGs

So far the only definitions of usability that have been discussed apply to regular software. These definitions however do not line up with the goals and usage of games. Games are meant to entertain and not to provide the quickest solution for a task, therefore games need their own definition for usability.

Pinelle et al. (2008a) give a good definition for regular games: "*game usability* [is] *the degree to which a player is able to learn, control, and understand a game*". They also stress that this definition strictly ignores the issues of entertainment, engagement and storyline as these are strongly tied to the artistic and technical elements of the game.

As this definition was created based on literature from before MMRGs became anything other than academic exercises (Desurvire et al., 2004; Federoff, 2002), it does not allow for their idiosyncrasies. The foremost of these is the dynamic environment. To take this into account I propose the following, slightly modified, definition: *Game usability is the degree to which a player is able to learn, control, understand and safely play a game in the environment it was designed for.* Again entertainment, engagement and storyline are left out of the equation for the same reasons as stated by Pinelle et al. (2008a).

Based on this definition, it is possible to determine the usability goals that need to be tested by an evaluation method for MMRGs. These are learnability, memorability, effectiveness and safety. In this context, safety should mainly focus on the external safety rather than on the ability to prevent errors or to recover from them.

Also taken into account is the environment a game is designed for, because this can influence the safety goal. For example, Seek 'n Spell can safely be played in a park, but not somewhere with busy roads. This isn't really a problem, as a place with busy roads is often easily identified as being unsuitable to play this game because having to run across them is clearly a bad idea.

As a direct result from not taking into account entertainment, engagement and storyline, satisfaction will not be considered a usability goal. Although usability can contribute to the level of satisfaction, it is also a real possibility that a player is not at all satisfied by a game because, for example, they do not like the plot, the style of graphics or the pace of the game.

Also not directly taken into account are weather conditions. This is by design, rather than by omission. Precipitation and low temperatures can cause problems not related to the usability of the game itself, but to that of the device. Although the influence of direct sunlight is also an issue that partly relates to the hardware, it can be partially alleviated by good design. None the less, it is not explicitly mentioned, as it will most likely pop up under either learnability or effectiveness because not being able to properly see the interface has a negative impact on these usability goals.

Having a clear idea of what usability should be for games, it is now possible to proceed with determining the heuristics that are most applicable to this situation that will be used as a benchmark.

3.5 Heuristic evaluation

Heuristics for the usability of games in general (Desurvire et al., 2004; Pinelle et al., 2008a, 2009), mobile games (Korhonen & Koivisto, 2006) and mixed reality games (Wetzel, Blum, Broll, & Oppermann, 2011) are available, but all of them cover only certain elements that are important. None of them however cover the whole. Therefore, these lists will be combined. This however also means that this list will have to be validated again in order to be of use in this context. For that purpose a second heuristic evaluation will be done using the heuristics by Pinelle et al. (2008a) and Pinelle et al. (2009).

3.5.1 Adapted heuristics for MMRGs

In order to get a list of heuristics for MMRGs, the literature on game usability heuristics was studied. The lists of heuristics for games (Desurvire et al., 2004; Pinelle et al., 2008a), multiplayer games (Pinelle et al., 2009), mobile games (Korhonen & Koivisto, 2006) and mixed reality games (Wetzel, Blum, Broll, & Oppermann, 2011) show a certain overlap which was to be expected. First, this overlap was removed by comparing the lists and removing duplicate entries and merging similar ones. The resulting list was then analyzed for heuristics that are irrelevant for mobile mixed reality games and that did not fit with the definition of usability introduced earlier. These were then removed, resulting in the following list of heuristics. The heuristics based on literature are not explained further as they have been described in section 2.8.

General usability heuristics

The general heuristics that are applicable to any MMRG, whether it is single or multiplayer.

1. Audio-visual representation supports the game and are easy to interpret (Desurvire et al., 2004; Korhonen & Koivisto, 2006; Pinelle et al., 2008a; Wetzel, Blum, Broll, & Oppermann, 2011)
2. Provide unobstructed views that are appropriate for the users' current situation (Korhonen & Koivisto, 2006; Pinelle et al., 2008a)
3. Device UI and game UI are used for their own purposes (Korhonen & Koivisto, 2006)
4. Provide users with information on game status (Pinelle et al., 2008a)
5. The player understands the terminology (Korhonen & Koivisto, 2006)
6. Navigation is consistent, logical and minimalist (Desurvire et al., 2004; Korhonen & Koivisto, 2006)
7. The game gives immediate and consistent feedback on the players actions (Desurvire et al., 2004; Korhonen & Koivisto, 2006; Pinelle et al., 2008a)
8. Provide intuitive input mappings that are easy to manage and have an appropriate level of sensitivity and responsiveness (Korhonen & Koivisto, 2006; Pinelle et al., 2008a; Wetzel, Blum, Broll, & Oppermann, 2011)
9. The player cannot make irreversible errors (Korhonen & Koivisto, 2006)
10. The player does not have to memorize things unnecessarily (Korhonen & Koivisto, 2006)
11. The game contains instructions and help, so that the user does not need to look things up (Desurvire et al., 2004; Korhonen & Koivisto, 2006; Pinelle et al., 2008a, 2009)
12. The player can turn the game easily off and on, and save games in different states either by choice or by temporarily loosing connectivity (Desurvire et al., 2004; Wetzel, Blum, Broll, & Oppermann, 2011)

Heuristic number 6 in the referenced literature only takes into account navigation within the interface of the game, not the real world. For an MMRG this is also important to consider, therefore the following heuristic is introduced:

13. Real world navigation takes into account the type of game and is logical – Navigation through the real world should also be logical and take into account both the size of the game world and the type of game. In a game that focuses on running, it may be acceptable to send the player across the game world time and again but in a puzzle game this can hinder people that would otherwise be able to play it.

In the list above, there is a glaring lack of heuristics concerning safety of the users as there is hardly anything to be found about that in literature. In a recent, extensive literature study on mobile usability Coursaris and Kim (2011) only found two studies that mentioned safety at all. Safety was not in the context of usability though, but as reason to have a mobile phone (Kurniawan, 2008; Palen, Salzman, & Youngs, 2001). Moreover, there are even studies that ignore safety altogether when mobile devices are concerned (Ji, Park, Lee, & Yun, 2006). A study about a mobile health care application (Kjeldskov, Skov, Als, & Høegh, 2004) briefly mentions safety, but this is more with regard to the correctness of data handling and thus of the patient than the external safety of the user themselves.

Related to the usability aspect of external safety Duh et al. (2006) suggest using see through optical devices rather than hand held displays for AR purposes. Although this is a good solution from a technical point of view it is not from a practical point of view as these devices are far from commonplace and hard to use with smartphones. Therefore this will currently not be considered as a viable option.

To cover the lack of available heuristics regadering safety three will now be introduced and explained.

14. Display a short warning message about physical safety – If the game can put the player in dangerous situations, like crossing busy roads or running into objects, display a short and entertaining warning. The length should be minimized, because else people are like to just ignore it.
15. Take into account the characteristics of the environment for which the game is designed – If the game is meant to be played in the streets, design it so that the user does not need to look at a map constantly while navigating the real world at the same time at high speed. On the other hand, in an open space like a park this may not be problem.
16. Safeguard the players legal safety – If the area for which the game is designed has certain laws or regulations that can get the player into trouble, this should be taken into account to the extent that players are not forced to break them. Make sure for example that items do not become inaccessible because they are in places which are not freely accessible or that players have to get from one place to another in a time that's only possible when breaking the speed limit.

The heuristics leave room for interpretation and also assume common sense on the part of the player. Although games may not force a player to run across a busy street without looking, they could still do it. In a game like Mister X you could possibly see Mister X running on the other side of the road and in an attempt to catch him you run across without checking for traffic. A situation like this relies on common sense. Unfortunately, anecdotal evidence from the TimeWarp project [Richard Wetzel, private communication 2011] shows that when players are very engaged with the game they are playing they become oblivious to the world around them and are likely to switch off common sense.

The heuristics presented here match the usability goals introduced in section 3.4 as follows:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Learnability	X				X	X		X			X					
Memorability					X	X		X		X						
Effectiveness	X	X	X	X	X	X	X					X	X			
Safety			X						X			X	X	X	X	X

Table 3.1: Schematic overview of what usability aspects are measured by which heuristic

Since the list of adapted heuristics has not been validated before, this also needs to be done before being able to use it. In order to do so, the following hypotheses are postulated:

Hypothesis 2 *The heuristics introduced in this thesis will yield a greater number of issues than the heuristics for usability of games by Pinelle et al. (2008a).*

Hypothesis 3 *The heuristics introduced in this thesis will yield more severe issues than the heuristics for usability of games by Pinelle et al. (2008a).*

Multiplayer usability heuristics

A game is considered multiplayer when players can interact with each other in some form via the game, but not when there is just a highscore list to compare results. Still, not every heuristic may apply to every multiplayer game. Although Parallel Kingdom clearly is multiplayer, there is no session management or matchmaking going on for example.

Rather than generating a new list of heuristics the ones introduced by Pinelle et al. (2009) will be used. Although they were originally intended for regular multiplayer video games, they are also relevant for MMRGs.

Chapter 4

Creating the framework

4.1 Problems that will and will not be addressed

Currently when evaluating the usability of MMRGs, some or all of the following problems can be encountered:

1. Time it takes to perform enough evaluations
2. Battery life time
3. The dynamic nature of the environment complicates data recording
4. Capturing input and output is difficult on mobile devices
5. Data recording can break the flow of the game
6. On some devices there may be legal issues concerned with making it suitable for usability testing

The battery life limits the amount of tests that can be performed in a row. Usability evaluation should not influence this, save for when the device is also used to capture data as for example Bell et al. (2011) or Mulder et al. (2005) do or when used to record diaries (Salen & Zimmerman, 2003). Currently the simplest way to alleviate this problem without having to wait for the phone to recharge, is to have spare batteries.[1] Going any further than this is outside the scope of this project.

Many MMRGs are meant to be played in a highly dynamic environment. Some, like Mister X, will even not come to full fruition when played in a less dynamic environment like an enclosed campus. Since influencing the environment is generally not an option, the data recording techniques will have to be considered carefully to fit the situation.

To avoid breaking the usage policy on iPhones the easiest solution is to either use a method that does not require this or to simply not use iPhones but smartphones running Android.

In creating the framework, the other problems will be used as guides when making decissions.

4.2 Evaluation of evaluation methods

Every usability evaluation method has its strengths and weaknesses which determine its appropriateness in certain situations. Diaries are for example suitable to capture real user data over an extended period of time, while heuristic evaluation is not fit for this.

Similarly not every method can be used to evaluate a game (Pinelle et al., 2008a) or perform an evaluation in a context sensitive manner. For a usability evaluation method to be suitable for games it has to provide measurements towards the correct usability goals: learnability, memorability, effectiveness and safety. It is also important that the method does not assume that tasks always are completed in the same order or at all. While this may be the case for some games it definitely is not the case for every game.

In regular evaluations it is preferable if they can be conducted as early in the development process as possible, e.g. with just a paper prototype (Laitinen, 2008; J. Nielsen, 1994a) . Although this can also apply to MMRGs, it can only be used to validate if the meaning of an interface element is clear. Although the emotional part of the game experience is not directly part of the usability evaluation, it

[1] Not possible with every device

can influence the cognitive processes of players (Bartolic, Basso, Schefft, & Glauser, 1999; Schutte & Schuettpelz, 2001; Seibert & Ellis, 1991) and thus the way they interact with the interface. Therefor it is important to also test it using a working prototype, because it impossible to generate the entire experience of playing a video game with a paper prototype (Laitinen, 2008) .

The other avenue to explore is the one of contextual evaluation. To determine if a method is suitable for contextual evaluation, the method and frequency of data recording are the most important determinants. For example, if a game is played over a longer period of time (days) it is not desirable to have an observer follow you around, but answering some questions every time you play is not a big problem. On the other hand, when a game is action packed and lasts only a short time you wouldn't want to fill out a list of questions after every action you take, but having an observer wouldn't be that much of a issue.

In table 4.1 all the evaluation methods mentioned in chapter 2.6 are evaluated on their applicability for the evaluation of games, evaluation in a contextual setting and some general information represented by the following questions.

- What usability aspects do they measure?
- How much time does it take to perform one evaluation?
- How much time does it take to analyze the gathered data?
- How many participants are needed?
- Who should the participants be? I.e. users, designers, usability experts or something else.
- How many people are needed to run the evaluation?

From table 4.1 we can conclude that heuristic evaluation, think aloud, interaction logs, remote testing, diaries, perspective based inspection and instant data analysis are suitable to test MMRGs as they can be used to perform evaluations of games and with the context in mind. Though pattern based evaluation is also potentially suitable, no usability patterns for games could be found but only design patterns for gameplay elements (Björk & Holopainen, 2004; Davidsson, Peitz, & Björk, 2004) .
MOT could also be used, but the consequences of the metaphors need to be re-evaluated to make them suitable for games. For example, the habits of input schemes need to be supported, but habits created in the game may need to be broken with once in a while to keep the game challenging.

Laitinen (2006) has also shown that expert evaluations have face validity and that contrary to the domain of regular software it makes no difference for the outcome if the evaluator is an expert in the field of games. This makes looking for experts easier and cheaper.

The methods that remain as viable options are think aloud (TA), diaries, instant data analysis (IDA) interviews and interaction logs (IL) when combined with another method to capture information about safety. Think aloud can be used concurrently or retrospectively. If not otherwise specified, TA means concurrently. Retrospective think aloud will be abbreviated as RTA

4.3 ARGUEMENT

Now that it is clear which methods can potentially be of use for evaluating the usability of MMRGs, they need to be examined for which types of games they can be used.
Based on this information it is possible to determine which of these methods is suitable for a game.
A distinction is made between "1 game" and "1 session". A game is the collection of all sessions that are played without resetting the score of one or multiple players. A session is the time a player is consecutively in the magic circle and is either ended by the player engaging in another task or by the end of the game.

As identified by Gielkens (2011) one way of classifying games is by the major skill that influence the ability to win the game. The two focus skills mentioned are running (as with Portal Hunt) and cunning (as with Tidy City). This division also needs to be kept in mind when selecting an evaluation method to use. When a lot of running, or physical activity in general, is involved, it becomes difficult to perform other tasks concurrently. Both for the participants and the evaluators.

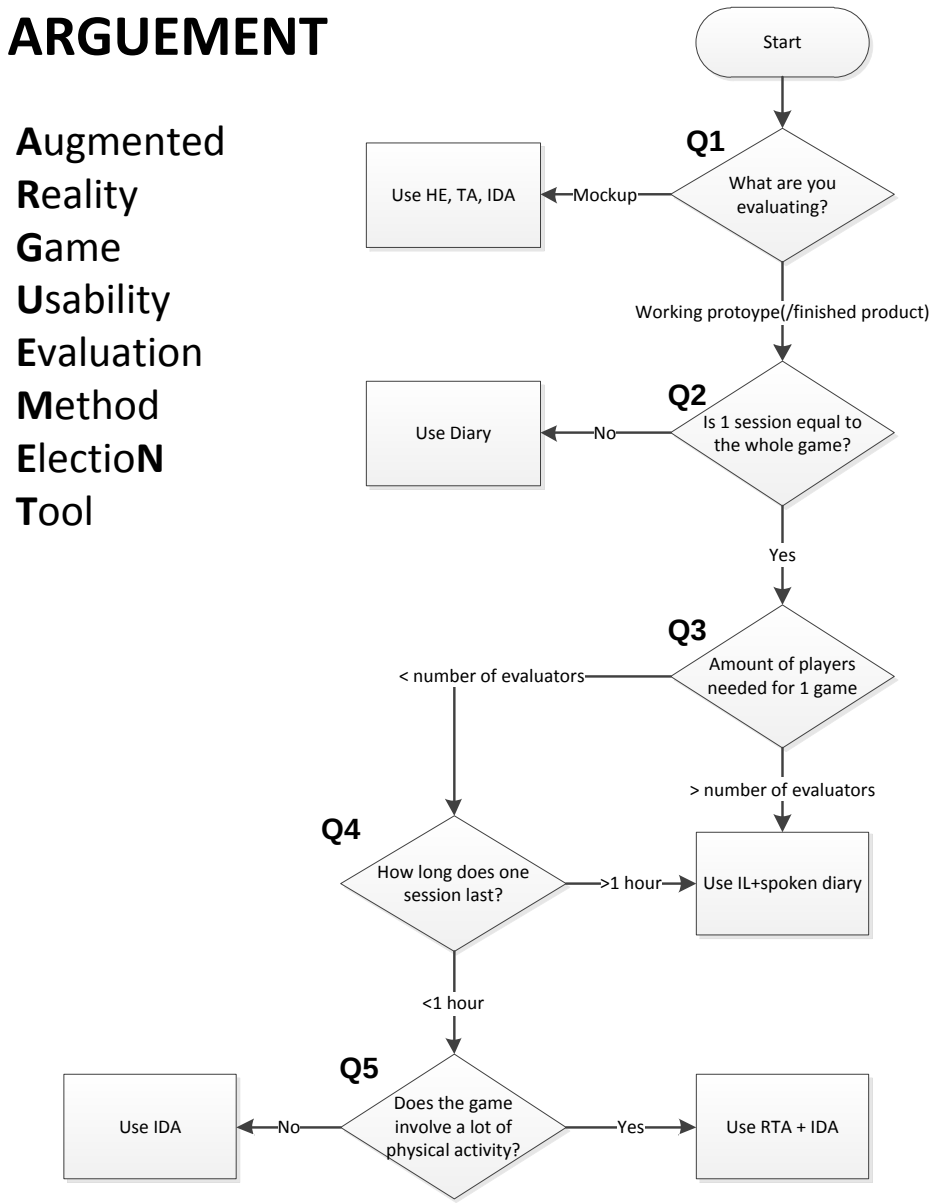


Figure 4.1: Augmented Reality Game Usability Evaluation Method ElectioN Tool. A flow chart to determine the most suitable usability evaluation method for a game

Another important factor in determining the suitability of an evaluation method is how long one play sessions lasts. Sessions of some games, e.g. Parallel Kingdom, can take an unlimited amount of time as there is no game mechanic that suggests any limit. In other games, e.g. SpecTrek or Seek 'n Spell, the player has to set a time limit ranging from minutes to hours. Sometimes game sessions may last even less than a minute, for example for a game like Photofari you may see something that matches a pattern you need, even though you're not currently playing, upon which you decide to quickly snap a picture and get on with what you were doing.

In the games that run for a limited amount of time, direct observation is a workable option because in the games that last only for a limited amount time, one session is almost always equal to the whole game. This means that the player is likely to encounter most or all of the interface elements. This in contrast

to the games that consist of multiple session, where you might not be able to access a some interface elements until you have performed certain actions. In this case, following the players/participants for direct observation is likely to take too much time.

In figure 4.1 the decision process for selecting an appropriate usability evaluation method is modeled in a flow chart. The decisions are made with the problems that need to be solved in mind, this means that although more methods than the ones given may be successfully applied, these may entail problems that are trying to be avoided.

Although there is no consensus in the literature about the need for a contextual usability evaluation, there is no question in the diagram to take the preference of the evaluator on this into account. This is by design, as MMRGs are required to be used in a mobile context. Evaluating them just in a lab session would limit the evaluation to generic usability issues that can be found in normal games too, like icons with unclear meanings. Issues unique to MMRGs will be hard or impossible to discover in a lab session. As an example one could think of the visibility under direct sunlight or navigational issues.

A non contextual evaluation can thus be used as a starting point on a first version like a mockup. It is strongly advisable though to also perform an evaluation in a context sensitive manner to check for issues arising in a real use situation.

In the following sections the decisions made in figure 4.1 will be explained, starting from the top down. Every suggestion will be able to cover all the relevant usability aspects, when applied properly.

4.3.1 Q1 – Are you evaluating a mockup or working prototype?

When performing an evaluation very early on in the design process, it is very likely that a mockup is used rather than a working prototype. For this kind of situation heuristic evaluation or instant data analysis are best suited. Both methods require only very few (3 – 5) participants and can be performed quickly. If thoroughness seriously outweighs the temporal benefits of instant data analysis, think aloud is also a valid method.

The mockup path can be extended further, but will not be unique to MMRGs therefore this will not be done here.

4.3.2 Q2 – Is one session equal to the whole game?

When working with a working prototype, or end product even, the first consideration is how much content there is in game, i.e. if one session equals the entire game. If this is not the case and players need multiple sessions to access the entire interface, the best technique is to use a combination of interaction logs and a diary. This way you can give the participants the chance to play over a longer period of time while still monitoring their performance.

One caveat in answering this question is that in multiplayer games where one sessions equals the whole game there is generally one player who sets up the session via an interface the other players do not have access to. This can be circumvented either by playing multiple sessions after each other set up by different players or letting all the players set up a game in turn and only actually start playing the last one.

Also important to note is that the diary method does enable the evaluator to analyze all the relevant usability aspects, but only if and when they create a diary that contains the right questions.

Although (retrospective) think aloud could also be used, this would require participants to return multiple times to play and yield many videos that would have to be painstakingly analyzed. All in all this would require that both participants and multiple usability experts have a lot of time available, making it a less favorable option.

Hypothesis 4 *When evaluating an MMRG that has an elaborate interface that cannot be fully accessed within one play session, using a diary will yield a higher number of usability issues than a heuristic evaluation.*

Hypothesis 5 *When evaluating an MMRG that has an elaborate interface that cannot be fully accessed within one playsession, using a diary will yield more severe usability issues than a heuristic evaluation.*

4.3.3 Q3 – How many players do you need for one game?

If one session is equal to the whole game, the first step is to see how many players are needed to play the game. If this is more than there are researchers available to monitor the players in real time for methods like IDA, using interaction logs and a spoken diary allows the participation of many people at once. In order for the participants to not have to exit the game application, it is worth considering to give the users a separate recording device so that they do not need to switch between applications.

To get a feeling for the dynamics of the game it may be interesting to follow one player around to see how they interact and interview them afterwards. As Jenova Chen points out in (Isbister & Schaffer, 2008) you have to be extra careful as players do not react the same way outside as at home and they shouldn't think that you are the maker of the game as this can influence their responses too.

One could also consider using retrospective think aloud at this point, but this will cost a lot of time since every participant will have to go through this process. That means that some participants will have to wait around for a long period of time, in which their memories of the events will degrade and may get contaminated by communicating with eachother. The evaluators could minimize the evaluation time they need afterwards by using IDA, but it would be much easier and less time consuming to simply log the participants actions and analyze these.

Hypothesis 6 *If a game requires more players than there are evaluators available, using interaction logs and and an audio diary as an evaluation technique will yield a higher number of usability issues than a heuristic evaluation.*

Hypothesis 7 *If a game requires more players than there are evaluators available, using interaction logs and and an audio diary as an evaluation technique will yield more severe usability issues than a heuristic evaluation.*

4.3.4 Q4 – Does one session on average take ≤ 1 hour?

In order to minimize the time it takes to perform an evaluation, it is important to take into account the time each individual session takes as some evaluation methods require many times the time it takes to perform an evaluation to process the results thereof. Although the duration of one session doesn't directly have any consequences on the suitability of any method, it does when taking into account that minimizing the time an evaluation takes is a goal of this tool.

When sessions take a long time, e.g. more than an hour, doing a think aloud exercise of any kind is going to become difficult. Retrospectively the participants will probably have a hard time remembering what happened at start and concurrently will only be somewhat possible using IDA. That method though, requires that 4 to 6 sessions are performed on a day and afterwards analyzed. When one session takes over an hour, this simply will not fit in one regular working day and is therefore not a viable option.

4.3.5 Q5 – Does the game involve a lot of physical activity?

Some games, like Mister X or Seek 'n Spell, require the player to do a lot of running. This can influence the ability to perform other tasks, like thinking aloud both because players may have low endurance and because thinking aloud is intense cognitive task. A good way to circumvent this is by using interaction logs, capture the input and output and let the player perform a retrospective think aloud. Because using retrospective think aloud approximately doubles the time it takes to gather data (Hoonhout, 2008) it is more appropriate to process the gathered data the way it is done in instant data analysis. This way, the time added on top of the pure play time is minimized.

Should the game focus on cunning, rather than running the evaluation time can be minimized by using IDA. This way, both participants and experts need to be available for a much shorter time.

Hypothesis 8 *When evaluating an MMRG that is slow paced and can be played by one player, using think aloud and instant data analysis as an evaluation technique will yield a higher number of usability issues than a heuristic evaluation.*

Hypothesis 9 *When evaluating an MMRG that is slow paced and can be played by one player, using think aloud and instant data analysis as an evaluation technique will yield more severe usability issues than a heuristic evaluation.*

Hypothesis 10 *When evaluating a fast paced MMRG using retrospective think aloud with instant data analysis as an evaluation technique will yield a higher number of usability issues than a heuristic evaluation.*

Hypothesis 11 *When evaluating a fast paced MMRG using retrospective think aloud with instant data analysis as an evaluation technique will yield more severe usability issues than a heuristic evaluation.*

4.3.6 Other considerations

Amaya et al. (2008) state that interaction between players can be a useful source of information, as especially friends and acquaintances are very likely to communicate about the game they are playing even without instructions to think aloud. This is why they suggest, if at all possible, to recruit groups of friends or acquaintances to test social games. This can extend to MMRGs where the players can operate in groups using one or more devices, even though recording the communication requires extra effort. From what Amaya et al. (2008) explain it seems worth the trouble, especially if combined with interaction logs.

In the framework it is suggested that instant data analysis and retrospective think aloud are used. These methods have not yet been used to evaluate games, but the method they are directly based on, think aloud, is (Desurvire et al., 2004). Since the only differences are how the data is analyzed, respectively when the thinking aloud takes place it is a reasonable assumption that these methods are also suitable for games.

It is also suggested to use interaction logs and diaries, even though there is no evidence in literature that they are actually suitable to evaluate the usability of games. Both diaries and logs have been used in research concerning games, but not specifically for usability (Cheong, Jhala, Bae, & Young, 2008; Cummings & Vandewater, 2007; Ducheneaut & Moore, 2004; En & Lan, 2010; Thawonmas & Iizuka, 2008) . In order to validate the claims an empirical study is setup.

	Usability Aspects										
	L[a]	M[b]	E[c]	S[d]	Evaluation time	Analysis time	# participants	Who participates	# evaluators	Context	Games
Heuristic evaluation	Y	Y	Y	N	Hours	Hours	3–5	Usability Experts	1	Debatable	With the correct heuristics
Cognitive walkthrough	Y	N	N	N	Hours	Days	≥ 2	Designer and $\geq$ 1 expert	1	No	No
Pluralistic walkthrough	Y	N	N	N	Hours	Days	≥ 3	Designer, $\geq$ 1 expert, $\geq$ 1 user	1	No	Yes
Think aloud	Y	Y	Y	Y	Hours	Days	$\geq$ 10× number of conditions[2]	Anyone	3	Yes, but it's complicated	Yes
Formal inspection	N	N	Y	N	Hours	Hours	4–8	Usability and domain experts	1	No	No
Feature inspection	N	N	N	N	Hours	Hours	unknown	unknown	unknown	unknown	no
Consistency inspection	N	N	N	N	Hours	Hours	1 per development team	developers	1	No	No
GOMS	N	N	N	N	Hours	Hours	0	—	1	No	No
Interaction logs	Y	Y	Y	N	Hours to months	Hours	Many	Anyone	≥ 1	Yes	Yes, partly
Remote testing	Y	Y	Y	Y	Hours to days	Hours to days	Depends on data capture method	Anyone	≥ 1	Yes	Yes, partly
Diaries	Y	Y	Y	Y	Days to months	Days	$\geq$ 10× number of conditions	Users	≥ 1	Yes	Yes

a: Learnability; b: Memorability; c: Effectiveness; d: Safety

Y denotes "Yes", N denotes "no" and C denotes "could be"

Table 4.1: Evaluation methods summarized together with their suitability for contextual evaluation and the evaluation of games.

[2]More is better

	Usability Aspects										
	L[a]	M[b]	E[c]	S[d]	Evaluation time	Analysis time	# participants	Who participates	# evaluators	Context	Games
MOT	Y	Y	N	Y	Hours	Hours	unclear	experts	1	With adaptations	With adaptations
SUE	C	C	C	C	Days	Hours	3 – 7	Anyone	1	No	No
MiLE+	Y	Y	Y	Y	Unclear (estimate: hours)	Unclear (estimate: hours)	Unclear	Experts	Unclear	Partly	No
Perspective based inspection	Y	Y	Y	Y	Hours	Hours	3	Experts	1	Debatable	With the correct heuristics
Pattern based inspection		unclear			Hours	Hours	10	Experts	1	Yes, with the right patterns	Yes, with the right patterns
Instant data analysis	Y	Y	Y	Y	Hours	Hours	3–5	Anyone	4	Yes, but it's complicated	Yes
Use case evaluation	Y	Y	Y	Y	Hours	Hours	4	experts	1	No	Yes
Interviews	C	C	C	C	Upto an hour	Upto an hour	3–10	Users	1	yes	yes

a: Learnability; b: Memorability; c: Effectiveness; d: Safety
Y denotes "Yes", N denotes "no" and C denotes "could be"

Table 4.1: continued

Chapter 5

Method

The validation process took place in two stages. First ARGUEMENT was reviewed by three usability experts along with the first version of chapter 4 and some extra notes to alleviate flaws I received feedback on beforehand. Some valid questions were raised but these stemmed from the added explanation that was insufficient rather than from the model itself. Taking this information into account, the explanations were altered to what can now be read in chapter 4.

Following this initial validation the suggested methods are used to evaluate the usability of a game that suits the conditions for which it has been suggested. The results of this evaluation are compared to the results of a heuristic evaluation as this is a generally accepted way of benchmarking a new method (Hornbæk & Frøkjær, 2004b; Schmettow & Niebuhr, 2007; Z. Zhang et al., 1999).

A comparison will be made on the raw number of issues, the number of issues per evaluator and the median severity rating. When comparing the raw numbers, it is also taken into account how many unique issues each method has found. Matching issues is done using fuzzy matching, i.e. generalized issues are matched to specific issues. For example, if one method identified three buttons that did not respond as separate issues and another method identified this as the general issues "buttons do not respond" these will be matched. This can result in a discrepancy between the amount of issues that match on both lists.

The raw amount of issues identified per game and method will be compared using a χ^2-test, but also qualitatively. A more robust validation would be by means of comparing the mean amount of issues identified per method and game across multiple evaluation sessions and evaluators, but this would require resources far beyond that of a master thesis. A general rule of thumb is that one needs at least 30 samples to be able to perform a meaningful statistical analysis and the simplest session, IDA, will require 4 to 6 participants. The more complicated ones, RTA with IDA and Interaction Logs + Spoken Diaries, will require about 10 people each per session and the evaluation of diaries requires at least 10 people per evaluation session as well. This would mean that in this case $(5 + 10 + 10 + 10) * 30 = 1050$ participants would be needed, not even looking at the number of usability experts that are required.

In addition to the statistical validation also qualitative evaluation of the methods is presented.

To validate the adapted heuristics a second heuristic evaluation will be performed using the heuristics by Pinelle et al. (2008a) and Pinelle et al. (2009). The adapted heuristics will considered valid if they perform at least as good as the heuristics by Pinelle, i.e. if hypotheses 2 and 3 can not be rejected. First both heuristic evaluation sessions will be described and following that the validation process for the other methods suggested in ARGUEMENT is discussed.

5.1 Adapted heuristics

For this method three experts were used. Two were external usability experts who had recently finished the same master as the author of this thesis and the third expert is the author himself.

Both external experts were familiar with the concept of heuristic evaluation, so this was not explained to them further. They were also to some degree familiar with the concept of augmented reality, but not with mixed reality and what constitutes a mobile mixed reality game. Therefore they were sent the same summary of my related work as the three experts that reviewed ARGUEMENT to make them more familiar with the matter.

# Experts	3
# Participants	0
Location	Schloß Birlinghoven campus
	Eindhoven
Games	Parallel Kingdom
	Portal Hunt
	Tidy City

Table 5.1: Summary for adapted heuristics

Rather than handing them the relevant pages of chapter 3.5, they were handed an extended version with explanations of the heuristics and without the references for legibility. (See appendix C). After finishing each evaluation, they were asked to compare and merge their lists. The items on the merged list finally were rated using the method described by J. Nielsen (1995) and in section 5.5.

As this method will serve as part of the benchmark every game that is tested with the suggested methods is also tested with this method. I.e. this method will be used to test Parallel Kingdom, Portal Hunt and Tidy City.

Due to time restrictions each game was evaluated for no longer than 60 minutes followed by 20 minutes of comparing notes and 20 minutes of rating. The evaluation sessions mostly took place on the Fraunhofer campus at Schloß Birlinghoven[1]. Due to technical problems with setting up the evaluation for Portal Hunt, part of the evaluation was done at a later date in Eindhoven[2] as this was easier for the two external experts. For this last part of the evaluation a mission was setup in Eindhoven with similar characteristics as the one available at the Schloß Birlinghoven campus.

5.1.1 On the sessions per game

To allow the reader some insight into the evaluations of the different games using the Adapted Heuristics, each session is now described briefly.

Parallel Kingdom

Although changing your location is not a requirement for the game, the evaluation did take place outside as both the GPS connection and internet connection were generally more reliable there. This also helped in checking if everything was clearly readable in sun light, which proved to be not the case.

Because the evaluation took place within a rather strict time limit, it is very likely that not every screen and interface element has been thoroughly checked. Nevertheless, a fair amount of issues has come to light.

Portal Hunt

This evaluation session was plagued with technical difficulties in setting up the evaluation, therefore it was done in two stages. The first stage was evaluating as much of the interface as possible without actually being able to play the game, as at this point only missions several hundred kilometers away were available. This was the case because setting up a test level required special rights that weren't available straight away and the mission editor was also highly unusable. In the end, getting the hang of the editor the first time around and getting the game to actually run cost about 3 days.

The second stage of the evaluation was done several weeks later and involved evaluating the parts of the interface that now could be used. As expected, several more issues were added to the existing list.

Tidy City

The session took place on completely unfamiliar territory for the external experts, so solving the riddles was quite difficult. Nevertheless, they did succeed in solving several of them within the imposed time limit.

To simplify recollecting all the issues that were identified, all issues were noted on paper while playing and later digitized. After this was done, the experts went over the application with the heuristics once more to check nothing had been missed.

[1] http://goo.gl/FbCmy
[2] http://goo.gl/4Sen8

5.2 Pinelle Heuristics

For this method three external usability experts were recruited. They all were familiar with the concept of heuristic evaluation and received no further instructions concerning that. They also had a good idea of what constituted augmented and mixed reality, so they received no further instructions on this either. In order to get them acquainted with the heuristics that were to be used, they were sent the relevant papers two weeks in advance and instructed to study them. From this point on the evaluation was equal to the one using the adapted heuristics.

# Experts	3
# Participants	0
Location	De Uithof, Utrecht
Games	Parallel Kingdom
	Portal Hunt
	Tidy City

Table 5.2: Summary for Pinelle heuristics

In order to meet the experts busy schedules, this evaluation was planned in Utrecht at the university campus on a Sunday[3]. All three experts were familiar with the surroundings in which the evaluations took place.

To draw up the lists of issues and rate them, the computers in the university library were used.

Like the adapted heuristics, these heuristics will be part of the benchmark for the suggested methods. Therefore all three games that will be used in the other methods are evaluated with this method as well.

5.2.1 On the sessions per game

To allow the reader some insight into the evaluations of the different games using the Pinelle Heuristics, each session is now described briefly.

Parallel Kingdom

Most notable in the evaluation session for this method were the severe problems connecting to the server. Although all the evaluators had a good mobile data connection, they all had to try and log in multiple times before they actually succeeded. One of the experts was completely unable to register an account, because the game indicated that a different account was already active on this device and it was not possible to register another account on that device. To allow her to evaluate the game somewhat at least, the author gave this expert his own phone and let her play with his own account.

Due to the weather, i.e. pouring rain, this evaluation took place indoors. Although this at first hampered GPS reception somewhat, this was easily solved by moving closer to a window.

Portal Hunt

As no game was created at this location before, one was set up to be similar to the one used for the other evaluations.

The game was played on a nearly empty parking lot. Amongst other things, the experts had great difficulties trying to figure out how to catch portals. After trying to figure out for themselves how it's done for about ten minutes, they were told how to actually do it.

Tidy City

At the Utrecht University campus a mission was already available, because it was created as part of the article by Gielkens (2011), therefore this one was used rather than setting up a new one.

The mission was created without taking into account that certain parts would only be accessible during office hours. Because this session took place in the weekend, two riddles became unsolvable within the time frame of the evaluation. As a direct consequence, it also became impossible to finish the whole mission.

[3] `http://goo.gl/xft3m`

5.3 Validating the suggested methods

Some participants took part in multiple sessions, but none were used to evaluate the same game with multiple methods. The participants that were used multiple times were all part of the B-IT lab course "Mixed Reality Games for Mobile Devices" at B-IT Bonn-Aachen International Center for Information Technology.[4]

5.3.1 Instant Data Analysis (IDA)

Selecting the game

This method is suggested for games that do not involve running and require only a limited amount of players. Single player games are for example very suited for this method. Therefore, Tidy City (section B.1.7) is an ideal candidate to be used for this method. It is a slow paced game, designed to be played by one or two people using the same device.

# Experts	3
# Participants	6
Location	Schloß Birlinghoven campus
Game	Tidy City

Table 5.3: Summary for IDA

On the evaluation method

IDA focuses on the players thinking aloud and two observers watching what happens and discussing their observations at the end of the day. As part of the interface is also determining which mission to play, the participants were allowed to choose which mission they would play as many different missions were available.

Although the game can also be played with multiple people working together on one device, having two people perform a think aloud exercise concurrently will most likely influence results and complicate data recording. Therefore only one participant was allowed to play at a time.

On the session

The session took place on the enclosed Fraunhofer campus at Schloß Birlinghoven, with a very limited amount of traffic. None of the participants had any trouble keeping their safety in mind when traffic was concerned.

Allowing the participants to freely play the game while thinking aloud was a good way to understand what they were doing. Although all of them were to a degree familiar with the method, some participants needed to be reminded during the sessions to keep thinking aloud.

Participants

Six participants played Tidy City for 15 to 30 minutes while thinking aloud. The participants were recruited from the B-IT lab course. Their average experience with games in general was rated as 5 on a 10 point scale, i.e. they sometimes play video games but not daily. Their experience with mobile games was much lower, averaging at 3.83 on a 10 point scale.

Some of the participants were familiar with the campus, because they already were doing an internship there. Others had only visited a small part of the campus as part of the B-IT lab course.

5.3.2 Retrospective Think Aloud with Instant Data Analysis

Selecting the game

This combination is suggested for games that involve a lot running and no more participants than evaluators. A game that suits this definition is Portal Hunt (section B.1.4), although it can also be played with more people than there are evaluators currently available.

# Experts	3
# Participants	4
Location	Schloß Birlinghoven campus
Game	Portal Hunt

Table 5.4: Summary for RTA/IDA

[4] http://www.b-it-center.de

On the evaluation method

In order to enable the retrospective element user interaction somehow has to be recorded. Preferably this is done with software that runs in the background so that player does not notice it. Alas this type of software is highly specialized and not publicly available. Creating such a piece of software falls outside the scope of this thesis and therefore a slightly more overt method of recording was chosen.

Participants were handed a smartphone to which a webcam is attached. This recorded all user actions and system reactions to a laptop they had to carry in a backpack. A schematic drawing is shown in figure 5.1. The reasoning, design process and final implementation of this setup will be explained in section 5.4.

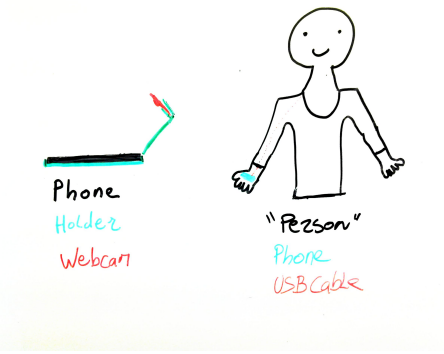


Figure 5.1: Left: general layout of the phone, webcam and rig. Right: Person holding the rig, with a usb cable running from the webcam under their shirt to the laptop on their back (not drawn)

In order to minimize the influence of time on memories, only one player per two evaluators per session was used for this method. The others were used for evaluating the next method, in order to save time in the process of evaluating the framework itself. Mixing the two methods should not cause problems, as every player has the same role and interface.

Since performing retrospective think aloud effectively doubles the time that is needed to perform an evaluation, sessions were ended after about 30 minutes. The data generated by the participants will be evaluated using IDA to reduce the amount of time it takes to process it.

On the participants

In total 4 participants (3 male, 1 female) were used for this method. Their mean age was 27 years and they were very experienced smartphone users (mean 9.25). With games in general their experience was much lower though (6.25), as was their experience with mobile games (mean 6.5) and augmented reality games (mean 6.75). Their areas of expertise varied, two reported computer science, one human-computer interaction and augmented/virtual reality and one medicine/neurology.

Three participants were recruited from colleagues at Fraunhofer FIT and care was taken that none of them had prior experience with this game. One participant was a visiting friend who had heard about the game, but didn't have any actual experience with it.

Two of the participants were clearly more talkative by nature than the other two, as they were talking a lot when playing without being told to do so. This was also reflected in their ability to think aloud retrospectively.

On the sessions

The play sessions for this method took place on the Fraunhofer campus at Schloß Birlinghoven, which meant that there was very little to no traffic. There were however other physical obstacles likes trees, bushes, buildings and other people.

In general participants in this condition seemed to find it quite hard to tell us afterwards what they were doing and thinking. The ones that were chatty by nature and had already made comments during the play session seemed to have an easier time than the others.

5.3.3 Interaction Logs and Audio Diary

In order to validate this combination, the extra participants of the previous method were used. Data recording took place using the same rig to attach a webcam to a smartphone and record to a laptop. As the webcams provide microphones and the evaluator will have to watch the entire video anyway to detect the interactions, the participants were instructed to say anything they particularly liked, disliked or did

# Experts	3
# Participants	4
Location	Schloß Birlinghoven campus
Game	Portal Hunt

Table 5.5: Summary for IL+Audio Diary

not get. This makes it easier for the participants, as they didn't have to handle multiple devices concurrently.

One issue that was not expected came to light when watching the videos of the experiments. When the phone received a certain kind of signal, this caused a lot of noise in the recording. This is the same kind of nice one hears when a phone is placed near speakers and it gets some kind of signal.

On the participants

Five participants were recruited from a pool of colleagues and visiting students from the B-IT lab course. One participant did not fill out a form with information on themselves, so the information presented here reflects only the four others.

All participants were male with an average age of 35, though this is highly skewed by one participant who was 55 years of age. All participants had a background in computer science with varying specializations. Also all of the participants owned a smartphone of some sort. This resulted in a high experience with smartphones (average 9.25). Their experience with games in general, mobile games and augmented reality games was much lower though. These averaged at 3.75, 3.50 and 4.50 respectively.

On the sessions and method

Listening to the comments people made while playing was very revealing, though their actions were maybe even more revealing. One participant for example ran into a tree, because he was focused on the game. He, however, did not comment on this apart from laughing. This is just one example of how capturing the user input and system output using the rig and a laptop generated more information than using a software solution would allow. Watching the videos also provided a good opportunity to see participants interacting with the game and explore the interface in a different way than one would do themselves.

5.3.4 Written Diaries

Selecting the game

# Experts	1
# Participants	9
Location	Anywhere
Game	Parallel Kingdom

Table 5.6: Summary for Written diary

This combination is suggested for games that are played over an extended period of time in which multiple sessions make up the whole game. A good example of such a game is Parallel Kingdom (section B.1.2).

On the method and session

Because response rates in diary studies have been shown to dwindle over time, a somewhat larger group of participants is invited to participate in this evaluation process.

Their instructions are to play the game at least once a day for at least 5 to 10 minutes, or more often if they feel like it. After each play session they were to fill out digital diary with a predefined set of questions and room to make unstructured comments. The diary will be available as a website, designed in such a way that it is also easily accessible from a smartphone. This way, participants who want to fill out the diary on the go can still do so. In appendix A a screenshot of the diary can be found. The choice for digital and structured diaries is based on literature that suggests that this increases response rates.

The game does not give the player any clear goals, other than a short tutorial. Adding goals for the purpose of this study has been considered, but rejected. Doing so could mess with the original game dynamics and the participants are very likely to report difficulties concerning the lack of integration between quest list and game.

Although interaction logs would probably be very helpful to see where the participants have difficulties, the unavailability of software to record them thwarts any plans for actually using interaction logs in this situation. Asking participants to always carry the rig with camera and laptop around, just in case they might play would influence the natural way of interacting too much. Therefore in this validation process, interaction logs can not be tested and are left out of the equation.

Participants

For this study participants were recruited from the B-IT lab course, as well as via Facebook. This resulted in a group of 9 participants who were instructed to play at least once a day for at least 5 to 10 minutes

(a) Headmounted cam (b) Lego prototype (c) Glasses

Figure 5.2: The three constructions tested to record interaction with the device

and fill out the diary afterwards for two weeks. They were also told that playing longer or more often was also allowed, as long as they filled out the diary after every session they played.

Of the participants 2 were female and the other 7 male. When asked about their area of expertise, 6 reported computer science, 1 reported math, 1 mobile applications and 1 mixed reality. The mean age of the participants was 25.11 years and all but one had completed at least a university bachelor.

5.4 Development of interaction logging tool

To decide what the best way to record the user interaction is three methods of recording it were tested. First a head mounted rig was tested (fig. 5.2a), secondly a Lego prototype of the phone holder suggested in section 5.3.2 was used (fig. 5.2b) and lastly special glasses with a camera integrated in the frame between the eyes were tested (fig. 5.2c). From these tests, it became clear that the phone holder with webcam was by far the best solution. This provided a stable, focused shot of the phone, while with the other two the view was very unstable, hard to focus and quite often the phone and hence the interaction were not in view of the camera.

As several different phones would be in use at the same time, more prototypes were constructed using Lego to suit every phone in use. After the prototypes were completed, they were converted to 3D computer models using Google SketchUp (fig. 5.3a and appendix D) and then printed using a 3D printer (fig. 5.3b). This way, the dimensions of the holder could be fine tuned to prevent the phone and webcam from shaking around when the participants moved. To fix the phone and camera in place and still be able to get them out again easily, a hybrid solution with Lego was made.

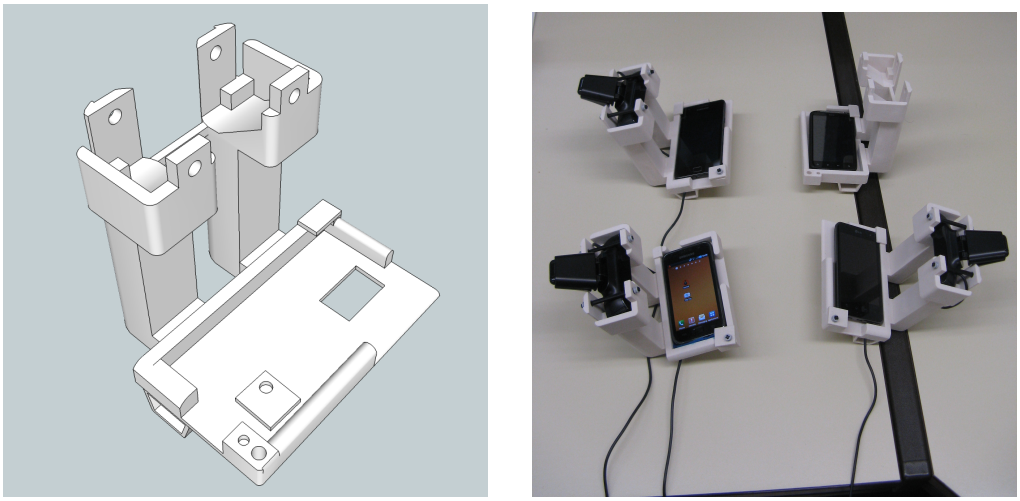
The phones used in the evaluations are a Samsumg Galaxy S, Samsung Galaxy S2 and LG Optimus Speed all running Android 2.3. The webcams used were all Logitech HD Pro Webcam C910's and the devices used to record their output were an Acer Aspire One A150, a Fujitsu Stylistic and a Dell Latitude X1. The first ran Ubuntu 11.04 and used the opensource program Cheese[5] to record video input, while the latter two both ran Microsoft Windows XP and used Microsoft Windows Movie Maker to record the videos.

5.5 Post evaluation

Each method generated a list of usability issues. These items were then rated on their severity from 0 to 4, with the following meanings per number (J. Nielsen, 1995):

0. This is not a usability problem
1. It is a problem, but only cosmetic. Fixing is only desirable if the time and budget allows.
2. It's a minor problem and fixing it has a low priority

[5] `http://projects.gnome.org/cheese//index`



(a) 3D model (b) Printed versions

Figure 5.3: The computerized and printed 3D models of the Lego prototypes for aiming the webcam at the phone.

3. It's a major problem and should be fixed
4. This is a catastrophe, with this in place the game can not be released

To get a more accurate measure, the rating was done by three evaluators (J. Nielsen, 1995). For the instant data analysis (retrospective) think aloud methods this was done while discussing the notes and for the other methods three evaluators rated each issue and the rounded average is used as final rating.

Furthermore, each issue is assigned to the heuristics defined in section 3.5.1 to investigate which kinds of errors are determined more by which method.

Lastly, the lists of usability issues per game were compared to determine how many unique issues were identified using each method.

5.6 Testing the hypotheses

The hypotheses concerning the severity of the issues found (i.e. 5, 9, 11, 7 and 3) will be statistically tested using a Mann–Whitney test to compare the medians of the different methods. Although this will only say something about the differences between these test-sessions, rather than the methods as a whole it can still be informative to determine if there is anything worth pursuing.

The hypotheses concerning the raw amount of issues found (i.e. 4, 8, 10, 6 and 2) can only be validated using a χ^2 test based on the frequencies per evaluation.

Chapter 6

Results

6.1 Introduction

In the following sections the summarized results of the different evaluations will be presented per game. The full lists of usability issues that were identified per method and game can be found in appendix E, for readability purposes. To get a general impression of what a play session could look like, one is described for each game using a persona. In each description some usability issues that were discovered will also be mentioned.

Each game has been evaluated using the methods suggested in ARGUEMENT (fig. 4.1), the adapted heuristics as introduced in section 3.5.1 and the heuristics by Pinelle et al. (2008a) and Pinelle et al. (2009).

6.2 Parallel Kingdom

6.2.1 An average play session

Persona

Name: Jake Watson
Age: 27
Education: University
Specialty: Computer Science
Owns a smartphone which he uses frequently, but has not much experience with games.

To start the game, Jake taps the red-flag icon of Parallel Kingdom on the home screen of his phone. After several attempts to login he makes it into the game and starts contemplating what to do now. Remembering that in his last session he was trying to gain access to high level crafting skills but hadn't quite completed that yet he sets out to find some enemies to slay.

After having killed some deer in the area where he started, he walks his dog to find a new spot. In the new spot he gets a message that a dungeon is nearby and he sees several deers and boars which he decides are now more interesting. After moving closer to one of the boars, he attacks it and he sees its health decreasing. Suddenly though, this stops. He tries walking about a bit, which does seem to work. Returning to boar he tries to finish what he started, without success. Although he told his avatar to attack the boar nothing happens. Then Jake notices that there is a message at the top of the screen, saying that there was a connection error and he has to restart the game.

Slightly discouraged by this, he does so. After first having problems connecting again, this time because of bad GPS, he manages to get back into the game. After finishing off the local wild life he tries to find the dungeon that was supposedly in the vicinity by randomly walking around. When he finally finds it he comes to the conclusion that he needed some oil in order to be able to enter the actual dungeon. He remembers that there was an oil well near his block hut, so using the travel menu he goes there. Upon gathering the required items, he tries to find the dungeon again but is unable to do so as there was no way to see where it was or revisit that place. Somewhat annoyed by this he leaves the game and goes on to work on thesis.

6.2.2 Heuristic Evaluation - Adapted

In total 35 usability issues were found using the adapted heuristics. The median rating of these issues was 3 on the 0 - 4 scale, indicating that there were quite a few severe issues. Unique issues identified using this method were e.g. the lack of auditory feedback and the obstructed views caused by enemies

standing on top of each other and so making it impossible to see how many there are. The full list of issues can be found in appendix E.1.1 and an overview of the amount of issues in per severity rating table 6.1.

Severity	Count
0	0
1	6
2	11
3	17
4	2
Total:	36
Median:	3

Table 6.1: Number of issues per severity rating for the adapted heuristics

Taking into account that there were three evaluators involved, this means that on average each evaluator identified 11.66 issues.

6.2.3 Heuristic Evaluation - Pinelle

Using the heuristics for single player games (Pinelle et al., 2008a) and multiplayer games (Pinelle et al., 2009), the three evaluators identified 29 usability issues with a median rating of 3. Unique issues identified in this sessions are e.g. that the avatar used for the monk in the tutorial is not at all clear and that it is impossible to see how strong an enemy is before you attack them. The full list of issues identified can be found in appendix E.1.2, whereas the summary is presented here in table 6.2.

Per evaluator on average 10 issues were identified.

Severity	Count
0	0
1	3
2	9
3	13
4	4
Total:	29
Median:	3

Table 6.2: Number of issues per severity rating for the Pinelle heuristics.

6.2.4 Diary study

This session yielded 58 diary entries, which averages to 6.44 entries per participant. The most active participant made 13 entries and the least active only 1. Not every entry showed a usability issues, some entries indicated multiple issues and other entries revealed issues that had already been identified by earlier entries. The net result of the 58 entries was 41 usability issues.

If every participant had filled out the diary every day, many more entries would have been created and possibly many more issues identified. Multiple participants reported they disliked the game so much they would have stopped playing, were it not that they had promised to help me.

With the 9 participants in the play session this method averages 4.55 unique issues per participant.

The median severity rating of the identified usability issues was 2. One of the unique issues identified by this method is the fact that the player sometimes is presented the possibility to perform actions he is not allowed to perform. Also not identified by other methods, is the issue that sometimes information can become unavailable once you have clicked it away. In table 6.3 an overview of amount of issues is given for this method, whereas in appendix E.1.3 the full list of issues and their ratings can be found.

Severity	Count
0	0
1	7
2	15
3	16
4	3
Total:	41
Median:	2

Table 6.3: Number of issues per severity rating for the diary study

6.2.5 Comparison

Severity

Using a Kolmogorov-Smirnov (Field, 2009) test it was determined that the severity ratings for all the methods were significantly not normal. With $D(36) = .28, p < .05$ for the adapted heuristics, $D(29) = .25, p < .05$ for the Pinelle heuristics and $D(41) = .23, p < .05$ for the diary method. This means a non-parametric test like the Mann-Whitney U test needs to be used rather than a t-test (Field, 2009).

Method	# issues	Normalized # issues	Median severity
Adapted heuristics	36	12.00	3
Pinelle heuristics	29	9.66	3
Diary	41	4.55	2

Table 6.4: Summary of the results per method

Although it seems that a difference exists in performance when looking at the amount of issues per severity rating (see figure 6.1), a comparison using a Mann-Whitney test reveals otherwise.

Severity ratings do not differ significantly between the Adapted heuristics ($Mdn = 3$) and Pinelle heuristics ($Mdn = 3$), $U = 460, z = -.878, ns$. The same goes for the difference between severity ratings of the adapted heuristic ($Mdn = 3$) and diary method ($Mdn = 2$), $U = 706.50, z = -.343, ns$ and the Pinelle Heuristics ($Mdn = 3$) and diary method ($Mdn = 2$), $U = 500.00, z = -1.198$.

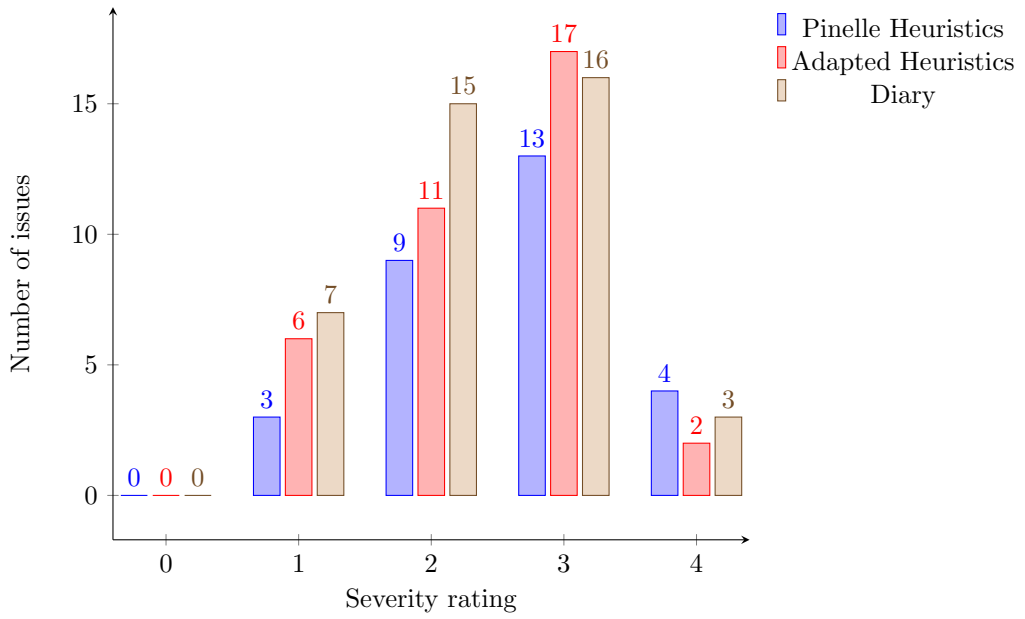


Figure 6.1: Bar chart showing the amount of issues per severity rating per method for Parallel Kingdom.

Raw amount of issues

By means of the diary method 12 issues more (41.37%) were identified than when the Pinelle heuristics were used and 5 more issues (13.88%) more were identified when compared to the adapted heuristics as can be seen in table 6.4.

Using a χ^2 test to compare the raw amount of issues, no significant difference was found between the adapted heuristics and the diary method ($\chi^2(1) = -0.324, p = .568$).

In order to see the difference in issues found all three the lists of issues were compared to one another. The list of issues generated by using the adapted heuristics contained 36 items, of which 11 could be matched to items on the diary method list. Conversely though, the diary method list contained 41 issues of which only 5 could be matched to issues on the adapted heuristics list.

Comparing the lists of issues identified by the adapted heuristics and the original ones, shows that both lists contain 4 items of respectively 36 and 22 items that are on both. When the list of issues generated using the original heuristics is compared to the ones generated using the diary method, the heuristic list contains 6 out of 22 issues that are also on the diary list and the diary list contains 6 out of 41 issues that are on the other one too.

heuristic #	adapted	pinelle	diary
1.1	6	4	4
1.2	**7**	4	4
1.3	0	2	0
1.4	3	2	2
1.5	0	1	**7**
1.6	2	2	**9**
1.7	2	2	3
1.8	1	1	2
1.9	2	1	1
1.10	2	1	1
1.11	1	2	4
1.12	3	0	2
1.13	1	1	1
1.14	0	0	0
1.15	0	0	0
1.16	0	0	0
0	1	1	1
2.1	1	1	0
2.2	0	0	0
2.3	1	2	0
2.4	0	0	0
2.5	1	1	0
2.6	2	1	0
2.7	0	0	0
2.8	0	0	0
2.9	0	0	0
2.10	0	0	0

Table 6.5: Number of issues per adapted heuristic, per method by which they were identified

Normalized amount

Comparing the normalized amount of issues one can see that the diary method identified approximately 7 issues per person less than the adapted heuristics.

Type of issues

Looking at the amounts of issues per heuristic in table 6.5 we can see that the adapted heuristics seem to be most suited to identify issues with providing unobstructed and relevant views (7), although the other methods also identify a number of these issues (4 each).

The diary method seems to be by far the best method identify issues when it comes to players understanding the terminology (heuristic 1.5) and the usability of the navigation in game (heuristic 1.6). For these kinds of issues the diary method identified 7 and 9 issues respectively, while the heuristic evaluations only identified $0 - 2$ issues for each.

Although there is not a difference of 3 or more, the heuristic methods do seem to be able to detect multiplayer issues better than the diary method. The latter didn't manage to reveal any, where the former did reveal some.

6.3 Portal Hunt

6.3.1 An average playsession

Persona

Name: Giselle Summers
Age: 24
Education: University
Specialty: Mathematics
Owns a smartphone which she uses frequently, plays games regularly and has some experience with AR.

Giselle starts Osmo4 and calls the session leader to complain that her phone broke down. She shows him a black screen, to which he responds that she has to press the menu button of the phone, select load file and navigate to the correct file and load it. Once she has done this, she repeatedly taps the input field for username. When nothing happens she looks frustrated and taps the password field repeatedly again, but to no avail. Again she calls out to the session leader for assistance. He explains that she has to tap menu again and select keyboard.

Once this was cleared up she succesfully logged in using the provided credentials although there was some doubt which server was to be used.

Now she needed to select a team to join, but the large list confused her and once again he asked the session leader which team she had to join. He told her to join team "Power Puff Girls", which she did at once. Now she was presented with a list of missions of which she selected the most likely one.

On her screen appeared a map with several moving green, red and yellow spheres and somewhere in the corner a red cross. After pondering over what this could mean she started walking around. In the mean time the cross had changed its color to green and some of the spheres had moved outside of the map. Giselle assumed these were no longer part of the game and moved to the closest sphere, a green one.

While walking towards the sphere, she switched to the augmented reality view. This presented her with three similarly sized spheres partially overlapping eachother. As she got closer though, one of them seemed to disappear and the other two became bigger until one of them left the screen and the other filled it.

Assuming that the way to catch a portal is by walking through it, she kept moving forward till the cameraview didn't show the portal anymore. Slightly confused about if she had or had not caught the portal, she tapped the overview screen. To her amazement she had not gained any points, but surely she had walked right through the portal?

Once again Giselle switches back to the AR view, turns around and sees the portal still floating infront of her. Determined to catch it this time, she runs right through it hoping that this might sort some effect. Alas, it did not. Thinking to her self that this was an advanced game, she tries to wave the portal in to the phone with her hand with little success once again.

Hoping to find instructions, she clicks the help button. Unfortunately it does not respond, not even after repeatedly tapping it. Frustrated, she once again calls out to the session leader, asking how she can capture a portal. Once he has explained this to Giselle she moves back to the portal she was trying to capture and succeeds.

Now she moves on to the next portal, a red one. When at the appropriate place she taps it, but nothing happens. Annoyed, she taps it a couple more times with the same result. Giselle decides to try the help button once more and now it does show some information that makes it clear why she isn't able to catch that red portal.

6.3.2 Heuristic Evaluation - Adapted

The concept of Portal Hunt was appreciated by the evaluators, but the implementation was found to be mediocre at best. This is mainly motivated by the amount of usability issues (50) that has been found and their severity. (Median 3) See table 6.6 for the overview and appendix E.2.1 for the full list of issues. Most astounding to the experts was that on one occasion the phone completely rebooted of its own accord and that the battery lasted not much more than 30 minutes.

Severity	Count
0	0
1	0
2	19
3	18
4	13
Total:	50
Median:	3

Table 6.6: Number of issues per severity rating for the adapted heuristics

Looking at a per evaluator basis, 16.66 issues were identified.

Apart from the issues taken into account here, four more issues were identified with the process of installing the game and registering an account. As this was not part of the game and none of the other evaluation methods were used to look at this process these issues are not used in the count as presented here.

6.3.3 Heuristic Evaluation - Pinelle

As with the adapted heuristics, the evaluators thought the idea was nice but the implementation was not as good as it could have been. Using the heuristics by Pinelle et al. (2008a) and Pinelle et al. (2009), 39 usability issues were identified with a median severity rating of 3. This means that on average each evaluator identified 13.00 issues.

Severity	Count
0	0
1	1
2	15
3	19
4	4
Total:	39
Median:	3

Table 6.7: Number of issues per severity rating for the Pinelle heuristics.

Unique issues identified in this session are e.g. the blinking of a camera on one phone and the lack of a training level for new players.

6.3.4 Retrospective Think Aloud evaluation using IDA

After four play sessions of which one participant was used to perform a retrospective think aloud exercise, the notes were compared and made in to a single list. Following that, the list was discussed with a third evaluator and several issues were proposed to be merged in to one, as several similar issues were noted separately. E.g. it was reported that a great deal of buttons did not respond as different issues and the long, irregular loading times also were reported for multiple problems. For the list included in appendix E.2.3 the issues have not been merged, because this would possibly skew the results as with the other methods this has also not been done. In total 39 issues were reported, with a median rating of 3. Per participant this method has revealed 9.5 issues.

Severity	Count
0	1
1	5
2	13
3	11
4	9
Total:	39
Median:	3

Table 6.8: Number of issues per severity rating for RTA/IDA

Using this method it came to light, amongst other things, that it was not clear for everybody that in the "mission select" screen one had to tap the mission and that participants thought the number of portals was different between the AR and map view.

6.3.5 Audio diary & interaction logs evaluation

Using this method, a total number of 56 issues have been identified. See table 6.9 and appendix E.2.4 for respectively the break down per severity level and the individual issues. As with retrospective think aloud list of issues, there are some issues here too that could be grouped in a more general issue. The main examples are here as well buttons that aren't responding and long loading times that can cause an issue.

Severity	Count
0	0
1	3
2	11
3	32
4	10
Total:	56
Median:	3

Table 6.9: Number of issues per severity rating for audio diary & interaction logs

On a per participant basis this method revealed 11.2 issues.

Issues that have not been caught using the other methods are e.g. the fact that the game is called "Portal Hunt" but in the game what you actually try to catch are called "spheres" and that when the login screen is partially covered by the keyboard, it is impossible to scroll. Also, someone ran into a tree. See figure 6.2

Figure 6.2: Short clip of a recording from Audio diary/interaction log session (click to play)

6.3.6 Comparison of methods

Severity

Even though there is quite a difference in the amounts of issues found per severity rating, this does not reflect in the median severity rating, which is 3 for all methods (see table 6.12).

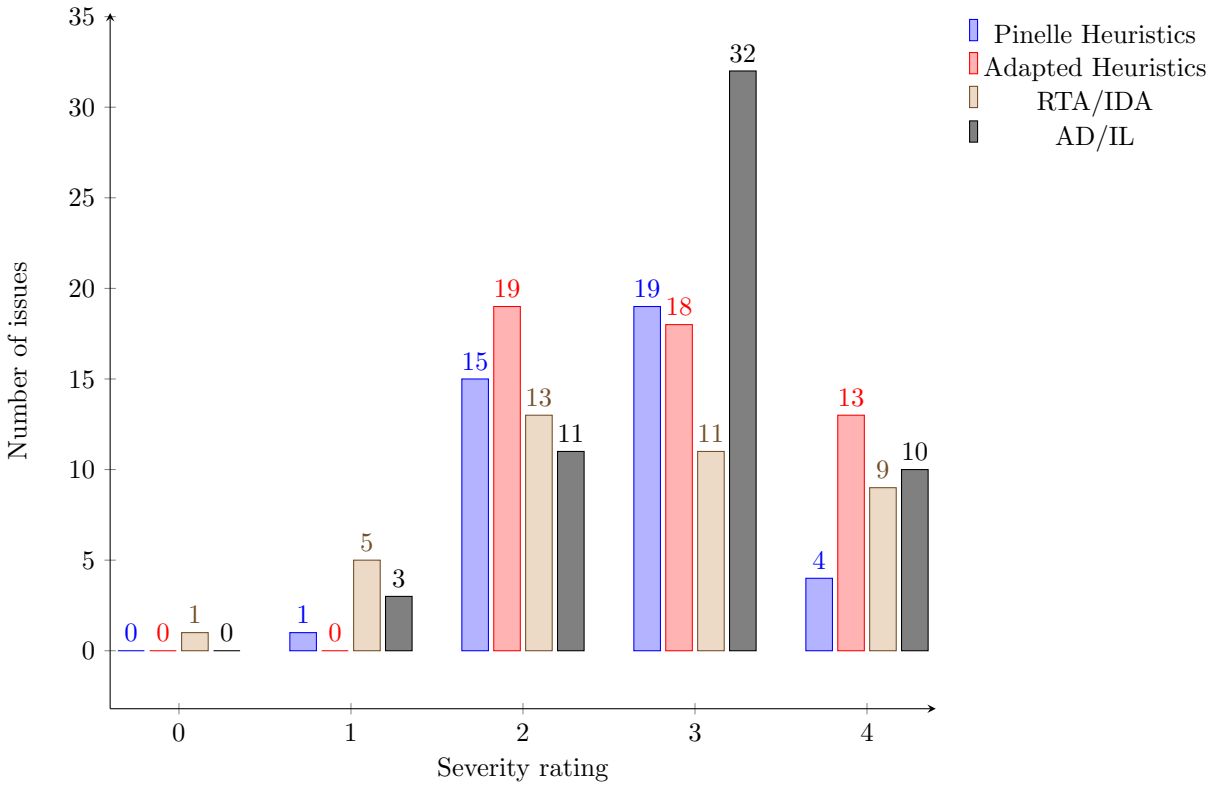


Figure 6.3: Bar chart showing the amount of issues per severity rating per method for Portal Hunt

Although the median is the same for each method, the distribution that lead to this median can still differ. To test this, first the normality of the distributions was analyzed using a Kolmogorov-Smirnov test. The distribution turned out to be significantly non-normal (see table 6.10 which necessetated the use of non-parametric test like a Mann-Whitney U test.

Tests of Normality							
		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Group	Statistic	df	Sig.	Statistic	df	Sig.
recoded	Adapted heuristics	.245	50	.000	.795	50	.000
	Pinelle heuristics	.273	39	.000	.828	39	.000
	RTA	.188	39	.001	.899	39	.002
	ADIL	.315	56	.000	.827	56	.000
a. Lilliefors Significance Correction							

Table 6.10: SPSS output for tests of normality for each data set

Pair	U	z	Sig.
Adapted & Pinelle Heuristics	848.500	−1.124	*ns*
Adapted & RTA	823.00	−1.322	*ns*
Adapted & ADIL	1364.50	−.242	*ns*
Pinelle & RTA	726.50	−.600	*ns*
Pinelle & ADIL	903.00	−1.357	*ns*
RTA & ADIL	907.00	−1.577	*ns*

Table 6.11: Outcome of the Mann-Whitney test for each pair of methods

Method	# issues	Normalized # issues	Median severity
Pinelle heuristics	39	13.00	3
Adapted heuristics	50	16.66	3
RTA	39	9.50	3
IL/AD	56	11.20	3

Table 6.12: Summary of the results per method for Portal Hunt

heuristic #	adapted	pinelle	adil	rta
1.1	7	2	12	7
1.2	4	4	5	1
1.3	2	1	0	0
1.4	3	4	5	6
1.5	1	0	6	2
1.6	2	1	3	1
1.7	4	2	4	4
1.8	6	5	9	8
1.9	1	0	0	0
1.10	3	1	0	0
1.11	4	9	1	3
1.12	3	0	2	0
1.13	1	1	3	3
1.14	1	0	0	0
1.15	0	2	1	1
1.16	0	0	0	0
0	3	1	1	2
2.1	1	1	1	1
2.2	1	1	0	0
2.3	1	1	0	0
2.4	2	2	0	0
2.5	1	1	0	0
2.6	1	1	2	1
2.7	0	0	0	0
2.8	1	0	0	0
2.9	0	0	0	0
2.10	0	0	0	0

Table 6.13: Amount of issues per adapted heuristic, per method by which they were identified in Portal Hunt

Before the test could actually be performed, the ratings had to be transformed because the test can not handle non-positive numbers and a 0 is not positive.[1] The transformation consisted of increasing every rating by 1.

Using the transformed data to perform the Mann-Whitney test confirmed that none of the distributions differed significantly from each other. See table 6.11.

Raw amount of issues

Using the adapted heuristics as a benchmark and the results from table 6.12, the audio diary plus interaction logs identified 12% (6) more issues, whereas both the Pinelle heuristics and retrospective think-aloud technique with instant data analysis identified 22% (11) fewer.

A χ^2-test revealed that there was no significant difference between the either the adapted heuristics and RTA/IDA ($\chi^2(1) = 1.359, p = .243$) or the adapted heuristics and Interaction logs & Audio Diary ($\chi^2(1) = 0.339, p = .560$).

Comparing the lists of issues identified by means of the Pinelle heuristics and adapted heuristics shows that the former list contains 12 issues also identified by the adapted heuristics, whereas the list created using the adapted heuristics contains 13 issues identified by means of the Pinelle heuristics.

The list of issues generated by using the adapted heuristics also contains 13 items that appear when the retrospective think aloud was applied and 11 issues that were identified by means of interaction logs and audio diaries.

Normalized amount of issues

After normalizing the results based on the number of participants needed to achieve the results, the differences between methods change somewhat. On a per participant basis the Pinelle heuristics identified about 2 issues more than the audio diary plus interaction logs. Retrospective think aloud identified 43% fewer issues per person (7) than the adapted heuristics, which is the greatest difference of all methods.

Type of issues

In table 6.13 one can see that the combination of audio diary and interaction logs reveals many more issues related to audio-visual representations than all three other methods.

Although it is not a difference of three or more, the heuristics do appear to be better at detecting issues concerning the mixing of the device and game UI (heuristic 1.3), as they have identified some issues related to this but the other two methods haven't. The same can be said for issues which force the player to memorize things unnecessarily (heuristic 1.10).

6.4 Tidy City

6.4.1 Average play session

Persona

Name: Joe Matterson
Age: 25
Education: University
Specialty: Computer Science
Owns a smartphone which he uses frequently, plays video games regularly and is quite experienced with AR.

Joe goes outside and starts the game. Before actually beginning to play he creates his own account, which he subsequently uses to log in.

After waiting for a bit until the "Browse Missions..." loading bar fills but nothing happening, he realizes it isn't a progress bar but a button. After tapping it he sees some overlapping text and a bunch of icons close to each other. He zooms in on the location so that he can read the mission names. After selecting an appropriate one, he downloads it and starts playing.

At first Joe has a bit of trouble determining which way he is going, because he didn't see the compass and he is used to his satellite navigation system which has a map that rotates to match the direction he is looking in. After getting used to this though, he managed to get the hang of it and successfully navigate towards the unsolved riddles.

[1] Mathematicians do not agree on this, but the ones that made SPSS are of this opinion so I'll have to deal with it accordingly

He tries picking up several riddles but fails because they are too far away. He walks closer and as the blue circle around him increases he stops to try and pick up the riddle. Unfortunately he is still too far away and moves closer still, until he succeeds in picking a riddle up.

Viewing the other tabs he comes across the showroom, but is slightly confused about its purpose. Ignoring this, Joe moves forward to pick up the next riddle. In the mean time, he has had an idea about the first riddle so he selects it and taps "solve" in order to start solving this riddle. To his surprise and dismay though, he is informed that he is at the wrong place and has just lost one point. Thinking it over he figures out what he did wrong and leaves the first riddle for later, because the solution is quite far away.

After a little while Joe arrives at the place where the next riddle should approximately be. He grabs his smartphone and looks if he is close enough. To his dismay he can't find the blue dot representing him. Even when pressing the "jump to me" button, the screen doesn't move. "Maybe the GPS reception is just very bad.", Joe thinks to himself and starts moving again. Then his avatar suddenly appears from behind the riddle he was aiming for and he picks up the riddle.

After solving most of the riddles, Joe is left with two very difficult riddles he simply isn't able to crack. He looks for some extra hints, but failing to find these he gives up.

6.4.2 Heuristic Evaluation - Adapted

The experts generally really liked Tidy City. During the evaluation the focus was on the general game interface and not the individual level design. It was noted on several occasions however, that in a game like this the level designer can also cause usability issues by placing riddles in weird places, especially violations of the heuristics A.13 through A.16 may be caused by them rather than by the game design.

Even though there are some usability issues concerning the auditive feedback, it was generally well received. Especially the pickup sound was appreciated.

What was also nicely done, was the colour selection on the icons representing the new and solved riddles. Even though it's yellow and green, they are still clearly distinguishable for people with yellow-green colour deficient eye sight.

Unique issues identified in this sessions are e.g. that the meaning of the "showroom" tab may not be clear when nothing is in it and that the auditive feedback when solving a riddle is somewhat minimalistic.

In total this evaluation identified 31 usability issues with a median severity of 2. This means that per evaluator on average 10.33 issues were identified.

Severity	Count
0	0
1	4
2	17
3	9
4	1
Total:	31
Median:	2

Table 6.14: Number of issues per severity rating for the adapted heuristics

6.4.3 Heuristic Evaluation - Pinelle

The experts in this session rightfully noted that the satellite view that was displayed was several years out of date, because some roads and buildings that had been around for over a year now weren't on the map. They also recognized that this was through no fault in the game design and is something that simply can't be corrected. Generally the experts found the concept interesting and despite their familiarity with the area of the mission, they still found it exciting and challenging.

Examples of issues identified in this session but not in other sessions are the fact that riddles can become impossible to solve or pick up because of the weather or buildings and that the position of the register button under the login field can be interpreted as directly registering using the data you gave in.

21 usability issues were identified using the heuristics by Pinelle et al. (2008a) with a median severity rating of 2. This averages to 7 issues per expert.

Severity	Count
0	0
1	2
2	12
3	6
4	1
Total:	21
Median:	2

Table 6.15: Number of issues per severity rating for the Pinelle heuristics

6.4.4 Think Aloud/Instant Data Analysis

After careful discussion of the notes by both observers, a list of issues was created. The full list contained 43 issues which were rated on a scale of 0 - 4. This resulted in 10 items being rated as not a usability problem. Reasons for this were e.g. that they related more to playability (people couldn't figure out the riddle) or were not a problem of the game, but the mission editor (broken missions

Severity	Count
0	10
1	17
2	12
3	4
4	0
Total:	43
Median:	1

Table 6.16: Number of issues per severity rating for IDA

were allowed to be set playable). Another reason was that one participant didn't understand an icon which was the defacto standard. Although that of course is bad, changing that would cause more trouble than it solves.

Regardless of whether the issues rated 0 are included, the median value of the severity rating is 2. On average, 7.16 issue per participant was discovered.

Issues that were identified by this method, but not by the other two are e.g. that it is not clear what the distance between the riddle and solution is and that it is confusing to have a "'login"' button in the menu when you are already logged in.

6.4.5 Comparison

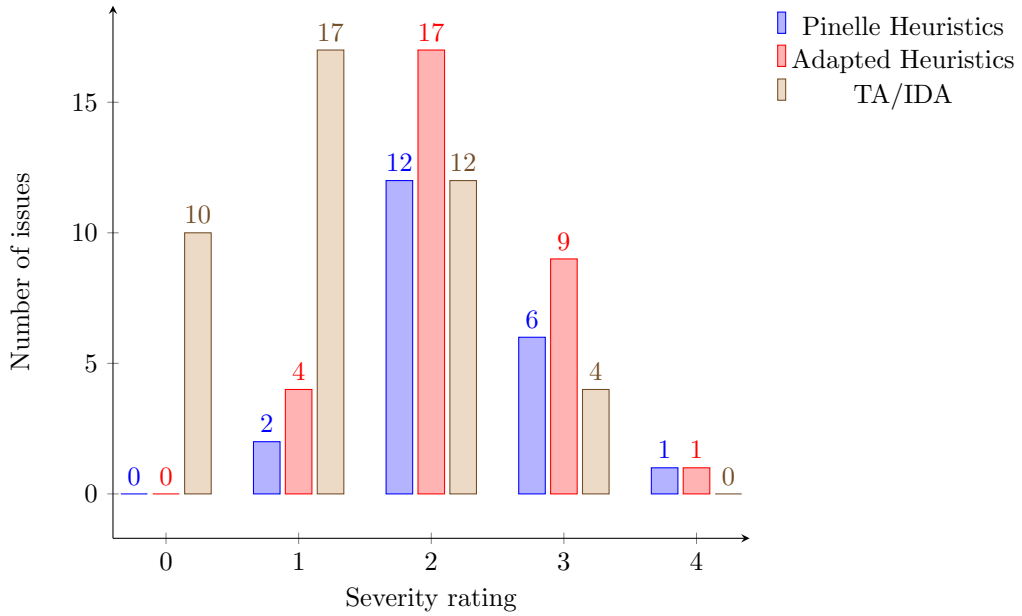


Figure 6.4: Bar chart of number of issues per severity rating per method for Tidy City

Severity

By means of a Kolmogorov-Smirnov test the normality of distribution was tested for every method and was found to be significantly not normal. For the Adapted heuristics this was $D(31) = .30, p < .01$, the Pinelle heuristics with $D(21) = .32, p < .01$ and Instant Data Analysis (IDA) $D(43) = .22, p < .01$. Based on this information, it can be concluded that a non-parametric test is necessary to analyze the differences.

Against expectations based on figure 6.4, the Mann-Whitney test revealed no significant differences between the median severity rating for the issues when the adapted ($Mdn = 2$) and Pinelle heuristics ($Mdn = 2$)were compared. $U = 313.50, z = -.250, ns$ The difference between both heuristic methods and IDA on the other hand, was significant, with $U = 286.00, z = -4.372, p < .01, r = -.508$ for the adapted heuristics ($Mdn = 2$) and IDA ($Mdn = 1$) and $U = 181.00, z = -4.046, p < .01, r = -.505$ for the Pinelle heuristics ($Mdn = 2$) and IDA ($Mdn = 1$).

Raw amount of issues

Method	# issues	Normalized # issues	Median severity
Pinelle heuristics	21	7.00	2
Adapted heuristics	31	10.33	2
IDA	43	7.16	1

Table 6.17: Summary of the results per method for Tidy City

Of all three methods used for this game, the Pinelle heuristics performed by far the worst (see table 6.17) with 47.62% (10) fewer issues than the adapted heuristics and 104.76% (22) fewer issues than were identified with Instant Data Analysis. The difference between the adapted heuristics and Instant Data

Analysis was much smaller, only 12 issues or 38.71%.

By means of a χ^2-test it was determined that the detected differences were not significant ($\chi^2(1) = 1.945, p = .163$) when comparing the adapted heuristics to IDA.

Interestingly the overlap in the amount of issues identified is rather small. The list of issues identified by means of both heuristic evaluations only show 4 items that occur on each. Comparing the list created using IDA shows a similar count, as it contains 7 items that were also identified by adapted heuristics and 3 that were identified using the Pinelle heuristics. Those two sets of overlap however, do not overlap each other.

Normalized amount of issues

Looking at the normalized amount of issues, the adapted heuristics show a much greater hit rate per participant with over 10 issues per evaluator, while both the Pinelle heuristics and Instant Data Analysis method only identify 7 issues per participant. Since only one datapoint is available for each method this can not be statistically corroborated.

Type of issues

After classifying all the identified issues per method table 6.18 can be produced. In it one can see how many issues per heuristic are identified.

Because the information gathered here is based on only one observation for each method, it is impossible to statiscally verify if there is a difference between the methods. Therefore a difference between the two best methods has to be chosen to denote the when a result is interesting or not. This boundary is somewhat arbitrarily put on a difference of 2 issues, so if the difference is 3 or more it is considered interesting.

In table 6.18 it becomes clear that the adapted heuristics themselves are apparently most suited to detect issues that concern the provision of unobstructed and relevant views, as with this method 6 issues like this are identified whereas the Pinelle heuristics only caught 1 issue and IDA 3.

On the other hand we can see that IDA seems to be more suited to detect problems regarding heuristics 1.4 ("Provide users with information on game status"), 1.5 ("The player understands the terminology") and 1.6 ("Navigation is consistent, logical and minimalist"). The Pinelle heuristics do not catch these issues at all and the adapted heuristics hardly.

The heuristic evaluation however, seems more suitable to detect problems that may not occur very frequently. When the experts are given a list of heuristics, they will attempt to find problems matching the description of the heuristic. A good example of this are the heuristics 1.14, 1.15 and 1.16. In a HE, the experts can think of situations where this would be needed. Due to the nature of the place where the IDA sessions took place, there was no problem with these heuristics.

heuristic	adapted	pinelle	ida
0	0	2	0
1	5	5	4
2	6	1	3
3	0	1	0
4	1	0	4
5	1	0	8
6	2	1	7
7	2	0	1
8	2	0	0
9	0	0	0
10	2	3	2
11	4	5	6
12	1	0	0
13	2	0	2
14	1	0	0
15	2	2	1
16	1	0	1
2.1	0	0	0
2.2	0	0	0
2.3	0	0	0
2.4	0	1	0
2.5	0	1	0
2.6	0	0	0
2.7	0	0	0
2.8	0	0	0
2.9	0	0	0
2.10	0	0	0

Table 6.18: Amount of issues found per adapted heuristic for each method in Tidy City

6.5 Heuristics comparison

As can be seen in table 6.19 the MMRG heuristics seem to consistently identify more issues than the heuristics by Pinelle et al. (2008a). Compared over the whole the MMRG heuristics identified 28 issues more, or 31.46%. This number includes the issues that were classified as multi-player issues and conformed with heuristics by Pinelle et al. (2009)

	MMRG	Pinelle et al.
Parallel Kingdom	36	29
Portal Hunt	50	39
Tidy City	31	21
Total	117	89

Table 6.19: Number of issues identified per method per game, including multiplayer issues

which were used in both evaluations. Taking those issues out of the equation we end with the numbers in table 6.20. This increases the absolute difference to 30 issues and the relative difference to 39.47%, further strenghtening the suggestion that the heuristics introduced here are a valuable improvement.

	MMRG	Pinelle et al.
Parallel Kingdom	31	24
Portal Hunt	44	33
Tidy City	31	19
Total	106	76

Table 6.20: Number of issues identified per method per game, excluding multiplayer issues

	MMRG	Pinelle et al.
Parallel Kingdom	4	1
Portal Hunt	5	3
Tidy City	7	2
Total	16	6

Table 6.21: Number of issues identified related to the added heuristics per game

As encouraging as these results look, upon inspection with a χ^2 test it was discovered that the differences were not significant both for the data including multi-player heuristics ($\chi^2(2) = 0.235, p = .889$) and the data excluding multi-player heuristics ($\chi^2(2) = 0.408, p = .815$) when looking at the differences between individual games. However, if only the total of identified issues are taken into account a slightly different picture is painted. If the issues related to the multiplayer heuristics are included the results are still not significant ($\chi^2(1) = 3.805, p = .051$). When they are not taken into account, the difference is significant ($\chi^2(1) = 4.945, p = .026$)

Drilling down further and looking only at the completely new heuristics, i.e. 12 - 16, some interesting things come to light (see table 6.21). Although the added heuristics were not available to the evaluators using the Pinelle heuristics, they still managed to identify some issues that could be classified as relating to them. Unfortunately the difference in these numbers can't be tested statistically using the χ^2 test per game, as the assumption that the expected frequency has to be at least 5 is violated. Using just the totals though, a significant difference is identified ($\chi^2(1) = 4.545, p = .033$).

Chapter 7

Discussion and Conclusion

7.1 General remarks

Some issues span every method, because they touch almost every game. Every way of evaluating the games requires people to play the game, since even with a heuristic evaluation you will need to test the real world navigation and map related issues in the field. That means participants will have to go outside and when the weather is bad, i.e. cold or wet, this may hinder or completely prevent the games from being played. It also can influence the results of the evaluation sessions.

One could try to circumvent this by passing fake GPS data to the device, but this is rather tricky, may not always be possible at all and could cause problems that would not occur in normal usage scenarios.

How much a game taxes a battery can also be quite a problem. This was specifically noted during the heuristic evaluation using the adapted heuristics, when one of the experts noted that his battery didn't last much more than 30 minutes. Considering that many elements of the device were being used, like the GPS, camera and network connection plus a higher processor load to use it all, this was to be expected but none the less it was a problem.

In the previous chapter it became apparent that when looking at the normalized amount of issues the adapted heuristics were always better than the suggested other method. This was to be expected as every other method requires more participants than are needed for a heuristic evaluation to identify approximately the same percentage of the total amount of issues. Therefore if the normalized amount would be equal the heuristic evaluation would have to have performed far worse than would normally be expected.

7.2 Adapted versus standard heuristics

7.2.1 Quantitative comparison

Looking at the difference in raw amount of issues identified per game, there was no significant difference in performance for both set of heuristics. However, if only the total amount of issues identified by each method which do not relate to the multiplayer heuristics by Pinelle et al. (2009) a significant difference is found in favor of the adapted heuristics. The same goes for when looking only at the completely newly introduced heuristics (12 - 16).

The lack of significant difference per game indicates that the newly introduced heuristics perform at least a good as tried and tested heuristics, while the significant difference in the completely new heuristics concerning navigation and safety indicate that they allow for the detection of new and relevant types of issues.

Hypothesis 2 *The heuristics introduced in this thesis will yield a greater number of issues than the heuristics for usability of games by Pinelle et al. (2008a).*

Seeing how the differences are significant when looking only at the generic issues and also when looking at the unique MMRG heuristics, hypothesis 2 is accepted. This means that the adapted heuristics will

be used as benchmark for the amount of issues.

Interestingly enough, the overlap between the two sets of heuristics was only minimal suggesting that both can be seen as a valuable tool.

Hypothesis 3 *The heuristics introduced in this thesis will yield more severe issues than the heuristics for usability of games by Pinelle et al. (2008a).*

As described in the previous chapter, no significant difference whatsoever exists between the median severity ratings of both sets of heuristics which leads to the rejection of hypothesis 3. For the rest of the analysis this means that both sets of heuristics will be used for comparing the median severity.

7.2.2 Qualitative comparison

Because both methods are a heuristic evaluation, there is no difference in the basic way they are performed. The heuristics that are used do influence the way the evaluation is performed though. Since the adapted heuristics require the evaluator to take real world navigation and physical safety into account, it is very much advisable to actually perform the evaluation in a context sensitive manner. Problems relating to the real world navigation were also discovered when the Pinelle heuristics were used, but none of the given heuristics prompted this.

After performing the evaluations it also became apparent that some of the design guidelines introduced by Wetzel, Blum, Broll, and Oppermann (2011) should have a more prominent place. Although the legal safety of players is also important, a much more frequent problem is depth perception in AR games. As pointed out in the design guideline, occlusion rich areas should be avoided as it currently isn't really possible to take this into account when projecting virtual elements on top of the real world. This issue was originally grouped under "Provide unobstructed views that are appropriate for the users' current situation", but this appears to be a too generalized concept and points more in the direction of virtual elements occluding other virtual elements rather than the real world.

In order to improve the heuristics two things should be done. The description for the unobstructed views heuristic should include the occlusion of the real world by virtual elements and the legal safety heuristic should be replaced with one concerning depth perception.

Also further and more thorough validation of the heuristics should take place, as different sets of evaluators were used in different settings. Although this was necessary not to skew the results due to pre-knowledge about the games, this somewhat compromised the internal validity of the research. If the same experts were used the internal validity would also be compromised due to pre-knowledge, unless different games would have been used which would also have led to trouble because not every game has the same amount of usability issues. The only way to circumvent that would be to analyze either many games with the same experts or the same games with many experts. Both would unfortunately require more time than was available.

7.3 Heuristic Evaluation versus Diary

7.3.1 Quantitative comparison

As can be seen from figure 6.1 and table 6.4 the amount of issues found differs between the three methods, with the greatest difference being between the Pinelle heuristic evaluation and the diary study.

Hypothesis 4 *When evaluating an MMRG that has an elaborate interface that cannot be fully accessed within one play session, using a diary will yield a higher number of usability issues than a heuristic evaluation.*

As reported in section 6.2.5, the diary method managed to identify 13% more issues than the adapted heuristics. This is even with a very low participation rate (see figure 7.1) and one less participant than is advisable. If either at least one extra participant was found or the existing participants would have been more actively motivated it is not unlikely that many more issues were identified. However with the results presented here, no significant difference was identified.

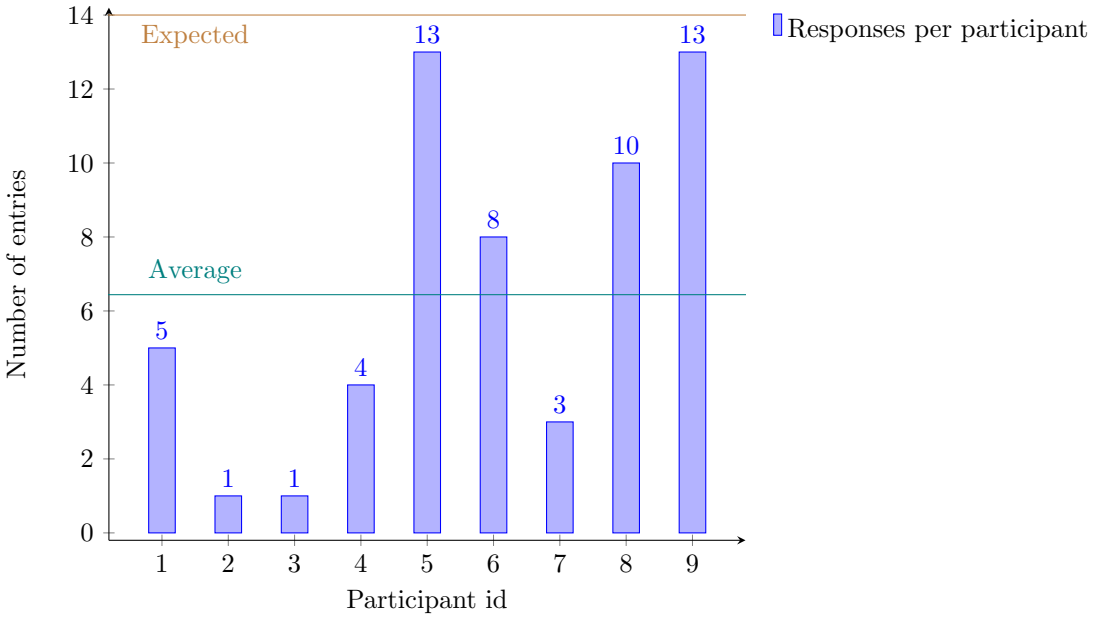


Figure 7.1: Bar chart showing the amount of diary entries per participant

When looking at what kind of issues both methods reported, it becomes clear that the diary method has not yielded anything with regards to the multiplayer element of the game whereas the adapted heuristics did. A possible explanation could be that the participants in the diary evaluation did not use this element of the game.

Also, the diary method has found many more issues where the mental model and knowledge of the users is concerned. These are heuristics concerning the used language, in game navigation and providing instructions. The adapted heuristics on the other hand were better at detecting issues concerning views that got obstructed or were irrelevant.

Although experts needn't be familiar with games, the ones participating in the adapted heuristics were. This means that they could have missed the fact that certain words or ways of navigating could be unclear to people who are not used to games. Experience with games could also account for the ability to notice the relevance of views more than the participants who filled out the diary.

Another explanation could simply be the difference in what the participants in the different sessions encountered. This is supported by the amount of unique issues per list, which is the largest part on each list.

Only looking at the heuristics that are unique or more relevant to the context as mentioned above, both methods perform about equal. Looking at the amount of issues found per participant, the diary method is clearly the worst method of the three. This was however to be expected, as explained in section 7.1. As the difference is not significantly different hypothesis 4 is rejected.

Hypothesis 5 *When evaluating an MMRG that has an elaborate interface that cannot be fully accessed within one play session, using a diary will yield more severe usability issues than a heuristic evaluation.*

Although the median severity is not equal it also is not significantly different and therefore this hypothesis can be rejected. As the same set of raters was used for this list of issues and the benchmark it is reasonable to assume that lack of difference is not there because of differences in opinion between sets of raters.

7.3.2 Qualitative comparison

Both heuristic evaluations were performed much quicker than the diary study. The latter took several days to set up, two weeks to run and then about a day to analyze the diary entries the participants had made. The heuristic evaluations on the other hand were completed within two hours.

The diary method took a lot longer to set up than the heuristic evaluations, as all the screens had to be identified and for each screen an input element was created. Also the diary was a completely custom build, which allows for a greater flexibility but also requires more time to construct. The biggest perceived advantage of a diary study done this way over a heuristic evaluation is that it's much easier to increase the number of participants without increasing the time that the experts need to spend on it a lot. Having one more expert perform a heuristic evaluation will require you to arrange that expert and have him spend up to a day analyzing your game. Adding one participant to your diary study will cost the expert that was already analyzing the results a bit more time, but never a day if the study was setup right.

It should also be taken into account that in an ideal situation this method would be used as part of the development process. At this point, an overview of every possible screen should be very easy to generate if not readily available whereas for this research an existing game had to be analyzed to create that list.

A benefit of using people that are actually likely to play the game is that you get better insight into what the target audience understands and as an added advantage you can use free form input fields for extra comments on other elements of the game.

7.4 Heuristic Evaluation versus Concurrent IDA

7.4.1 Quantitative comparison

Hypothesis 6 *When evaluating an MMRG that is slow paced and can be played by one player, using think aloud and instant data analysis as an evaluation technique will yield a higher number of usability issues than a heuristic evaluation.*

Instant data analysis has resulted in 38% more issues being found than using the adapted heuristics. It mainly did clearly better where issues concerning information on the game status, the used terminology and the ingame navigation. The heuristics were slightly better at discovering issues which related to relevance and obstruction of views.

Looking at the amount of issues per participant IDA did not perform as well as the adapted heuristics. An explanation could be that some participants just understood the interface for the most part. This of course then leads to few issues being identified based on the observations of that participant and a drop in the overall average.

A χ^2 test revealed that the difference in amount of issues identified was not significantly different between methods, even though difference was 38%. This means that both methods performed equally well, which leads to rejection of hypothesis 6.

Hypothesis 7 *When evaluating an MMRG that is slow paced and can be played by one player, using think aloud and instant data analysis as an evaluation technique will yield more severe usability issues than a heuristic evaluation.*

As shown the difference between the median severity of the heuristic evaluations and IDA is significantly different. Against expectations though, the heuristic evaluations found more severe issues. Therefore, this hypothesis can be rejected. Many of the issues with a rating of 0 can be considered usability issues, but stem from issues with the mission editor and therefore are not considered as usability issues of the game. Leaving them out of the analysis still shows a significant difference between the median severity of the heuristic methods and IDA, while reducing the difference in the raw amount of detected issues to 2. Based on literature this is exactly what was expected in a stationary context (Kjeldskov, Skov, & Stage, 2004), but now also seems to be supported for a mobile context.

The difference in the severity rating is somewhat surprising as literature (ibid) suggests that IDA should be more suitable to detect severe issues rather than cosmetic issues. The most likely explanation is that the results of the severity rating differ based on the way the rating was conducted and based on

the raters. A different set of raters was used to asses the severity for the heuristic evaluation results than for the IDA results.

7.4.2 Qualitative comparison

The heuristic evaluation required fewer people to perform than the instant data analysis session, 3 experts versus 3 experts and 6 participants respectively. This means it of course is easier to organize, as fewer people are needed and the experts in the heuristic evaluation don't necessarily need to evaluate the software at the same place and time. Using this method though, there is less room for discussion and experts may assume that defacto standards are understood by users while this may not be the case.

A valuable advantage of IDA over a heuristic evaluation is that the evaluator gets the opportunity to interact with people. This allows them to get a feeling for the actual users and see how they interact with the product. Sometimes this can be in unexpected ways, e.g. one participant expected in Tidy City that the images and riddles were general concepts, i.e. an ashtray meant any place to smoke and not that specific ashtray. This is something none of the usability experts came up with during the evaluation. Although one can not prove this, a plausible explanation would be that the experts understood that it's impossible to add every instance of a general concept within a certain region to the games, whereas this participant did not.

7.5 Heuristic Evaluation versus Retrospective IDA

Especially the retrospective part of the think aloud method seemed to be very challenging to participants, which is also clearly shown in the low number of issues that has been reported. When one compares the ratio of the number of issues found for RTA and the adapted heuristics (39:50 or 0.78), and the same for IDA and the heuristic evaluation of Tidy City (43:31, or 1.39) it becomes clear that there is major difference in performance between concurrent and retrospective think aloud for this context. This can of course also be due to the nature of the game, but that would have to be investigated in a follow up study.

Hypothesis 8 *When evaluating a fast paced MMRG using retrospective think aloud with instant data analysis as an evaluation technique will yield a higher number of usability issues than a heuristic evaluation.*

When comparing RTA to the adapted heuristics 11 fewer issues have been found. This difference of 22% does not look very promissing, which is confirmed by the non-significant difference as shown by a χ^2 test. Therefore, the hypothesis is rejected. Although a follow up study could be done to confirm this robustly with statistics, the results look so dismal and participants had such great difficulty recalling their thoughts that this may be wasted effort.

Hypothesis 9 *When evaluating a fast paced MMRG using retrospective think aloud with instant data analysis as an evaluation technique will yield more severe usability issues than a heuristic evaluation.*

Seeing how no significant difference in the median severity exists between the different heuristic evaluations and the RTA this hypothesis can also be rejected.

As with the concurrent IDA a different set of raters was used than for the adapted heuristics. This can explain the lack of difference just as likely as the presence there of.

7.5.1 Qualitative comparison

Similar to concurrent IDA, this method requires more people and may therefore be harder to organize. A drawback definitely seems to be the limited short term memory of people (Anderson, 2004). Some participants had great difficulty recalling what their thoughts were at one point or another in the game even with the availability of video and audio to support them.

Although being able to interact with the participants can be a valuable source of information, the really appalling results overall in this study shows that other methods are a better option. An evaluator could consider analyzing the recorded material too separate from the RTA protocol with the participant but this would require the evaluator to spend the whole time of the play session again on analyzing the video. As will be shown doing so is worth the effort, but also having to go through the RTA/IDA protocol adds a lot of time that may not be very useful.

7.6 HE versus Audio diary+Interaction Logs

Hypothesis 10 *If a game requires more players than there are evaluators available, using interaction logs and and an audio diary as an evaluation technique will yield a higher number of usability issues than a heuristic evaluation.*

For the raw amount of issues, the combination of interaction logs and audio diary has yielded a similar amount of issues (56) as the heuristic evaluation (50). There is a difference but since it's rather small, only 12%, it is not surprising that the χ^2 test revealed it was not significant.

Comparing the amount of issues per heuristic though, the heuristics only seem to outperform the audio diary and interaction log combination when it comes to providing help and instructions. The suggested method reports many more errors when it comes to how good the audio-visual representations are, clear terminology and the input mappings. As with the diary method this can be explained with the background knowledge the evaluators in both sessions have.

Looking at the amount of issues per participant this method again shows less success than the adapted heuristics by a large margin.

Mainly based on the results of the χ^2 test the hypothesis can be rejected.

Hypothesis 11 *If a game requires more players than there are evaluators available, using interaction logs and and an audio diary as an evaluation technique will yield more severe usability issues than a heuristic evaluation.*

Since the median severity is not significantly different when compared to the heuristic evaluations this hypothesis can be rejected.

The lack of difference could be explained by the severity and obviousness of the issues that are reported. Problems like a keyboard that does not appear as is normal or the complete lack of instructions on how to catch a portal are very likely to be found any way you perform the evaluation.

7.6.1 Qualitative comparison

What was evident from watching some of the participants in this condition was that they seemed much less inhibited than participants in both the concurrent and retrospective IDA session. A possible explanation for this could be that subconsciously the knowledge one would not be called out on ones actions allowed the participants to play more like they would normally. For evaluation purposes this is a very good thing, as the best case scenario would be that participants are observed in a completely normal use situation without their knowledge.[1] Even though the difference in amount of issues identified could very well be a random occurrence, the more natural behavior of participants allows for much better observations which makes this method a valuable tool for evaluating the usability of MMRGs.

7.7 Main hypothesis

In table 7.1 a quick recap of the hypotheses and if they were accepted or rejected is shown. It becomes clear that none of the suggested methods identify more severe issues than a heuristic evaluation. Looking at the raw amount of issues though most of the suggested methods identified a higher amount

[1] Of course this impossible, both for ethical and practical reasons

	Raw amount	Severity
HEA > HEP	Accepted	Rejected
IDA > HEA	Rejected	Rejected
RTA/IDA > HEA	Rejected	Rejected
Diary > HEA	Rejected	Rejected
AD/IL > HEA	Rejected	Rejected

Table 7.1: Overview of the hypotheses and their validity

of issues in this set of evaluations but the difference was never significant except for the adapted heuristics.

Hypothesis 1 *Depending on the style of the mobile mixed reality game, different usability evaluation methods will be more suitable*

Although one could point to the dismal results of RTA to support the claim that not every method is suited for every type of game, the only conclusion that can be drawn based on the statistical observations presented here is that the main hypothesis should be rejected.

7.8 Limitations

One of the biggest limitations in this research is the validity, as several different experts were used in different locations. Although currently results look favorably for the adapted heuristics, further validation is required to make sure the currently observed differences aren't due to variations in evaluators or locations.

For the other methods a similar case can be made. Although they generally identified a higher number of issues, none of the differences were statistically significant. This can be caused by many factors like the location, the participants or the evaluators. Current comparisons are only based on a single observation and it would show due diligence if the experiments were repeated in such a fashion that more solid statistical evidence could be found. As pointed out before though, this would fall far outside the scope of this thesis.

The framework introduced in this thesis has been dubbed ARGUEMENT, where AR refers to Augmented Reality but the framework is meant for MR games. This in contradiction with the point that has been made at the start, explaining how AR and MR are not interchangeable. Although I am aware of this, I couldn't come up with a nice, plausible acronym using MR. And as Shakespeare put it so poignantly in Romeo & Juliet:

> "What's in a name? That which we call a rose
> By any other name would smell as sweet."

All the smartphones used in this evaluation ran Android, but only a few different models and Android versions were used. From investigating the currently available games for Android-phones, it has become clear that both the version of Android and the model of the phone can have a huge impact on how well a game functions. The best example of this is Third Eye, which was usable on one type of phone and Android combination but on every other device it was tested it had severe problems. Quite often to the degree that the game would crash or stop responding.

To avoid contamination of the results, all devices were tested beforehand to assure that they had similar performance and that the games actually worked on them. Nonetheless technical difficulties were experienced on multiple occasions for reasons unknown.

As only smartphones running Android have been used, no investigation into what is currently available for the other operating systems has been made. Although it's possible that on other operating systems some methods could be applied differently, mainly capturing interaction logs, one can assume that the methods as used here are independent of the type of smartphone in use.

MMRGs are not only available on smartphones, but also on mobile hand held game consoles. These were unavailable at the time of writing so they have not been used. Although it seems plausible that the

results of this are generalizable to this context, it can not be said for certain as currently not all of these devices come equipped with GPS and mobile internet capabilities[2] and the device user interface differs greatly from that of smartphones. This is something that has to be investigated further.

Finally, this research has only looked at truly mobile mixed reality games. That is games in which participants are required to change their geographical location in order to play the game. The counterpart of these games are faux mobile games, which are also played on a smarpthone but do not require the player to change their geographical location. An often seen type of game in this category is one where the player stands in a spot and has to look around using a magic lens to shoot down virtual enemies. It seems plausible that ARGUEMENT can also be used for these games when the fifth question, about physical activity, is taken somewhat more liberally. Spinning around to shoot down enemies can in this case be explained as physical activity as the player would have to concurrently think aloud, aim and keep their balance. Though plausible this would need to be looked into further to actually verify it.

7.9 Revisiting ARGUEMENT

Even though the main hypothesis is rejected, the raw number of issues identified by the different methods was generally higher though not significantly so. This means it is still interesting to look into improving and further validating ARGUEMENT. One improvement will be suggested here, while further validation is left as an open question for future work.

The only method that performed clearly worse than the adapted heuristics was RTA and therefore it should be removed from ARGUEMENT. Audio-diaries combined with interaction logs have shown to be a useful method and so should remain as a suggestion in ARGUEMENT and the same can be said for the diary study. As interaction with actual players can allow the evaluators also to gather data on the playability of games and gain better insight into the way actual players interact with the game, IDA should also still be considered. Even though the insight into playability was not the goal, it can be a nice added bonus when applied in production setting rather than an academic setting.

Both the diary method and the interaction logs + audio diary method will be left in, as these have proven to be quite useful. The former mainly due to the easy scalability and the latter mainly due to the more natural behavior of the participants.

Finally it has been shown that the adapted heuristics are a good tool to evaluate the usability of MMRGs and will therefore be added to each suggestion.

These improvements can be seen in the final version of ARGUEMENT in figure 7.2

7.10 Conclusion

This thesis started out by exploring the world of augmented and mixed reality games and discovering that evaluating their usability was still unproven ground. To increase the knowledge available on this subject first a broad market study was done to determine what the state of the art was that the general public can access. In order to make this a manageable amount of work the spectrum was reduced to truly mobile mixed reality games running on Android smartphones. Truly mobile means that changing ones location is an integral part of the game (Wetzel, Blum, Broll, & Oppermann, 2011).

The next step was to perform a literature study to determine which methods would potentially be suitable to evaluate these games in a context sensitive manner. Based on the information learned from the market study and literature study about usability evaluation methods the main hypothesis was formulated as follows: *Depending on the style of the mobile mixed reality game, different usability evaluation methods will be more suitable than a heuristic evaluation.*

This resulted in a first version of the framework which suggested different methods for different situations. These suggestions were:

1. If games normally span over multiple play sessions use a combination of interaction logs and diaries to gather data.
2. If you need more players for one session than there are evaluators available to monitor them, use a combination of interaction logs and spoken diaries.

[2]PS Vita can do this, Nintendo 3DS can not e.g. `http://us.playstation.com/psvita/tech-specs/` and `http://www.nintendo.com/3ds/features/specs`

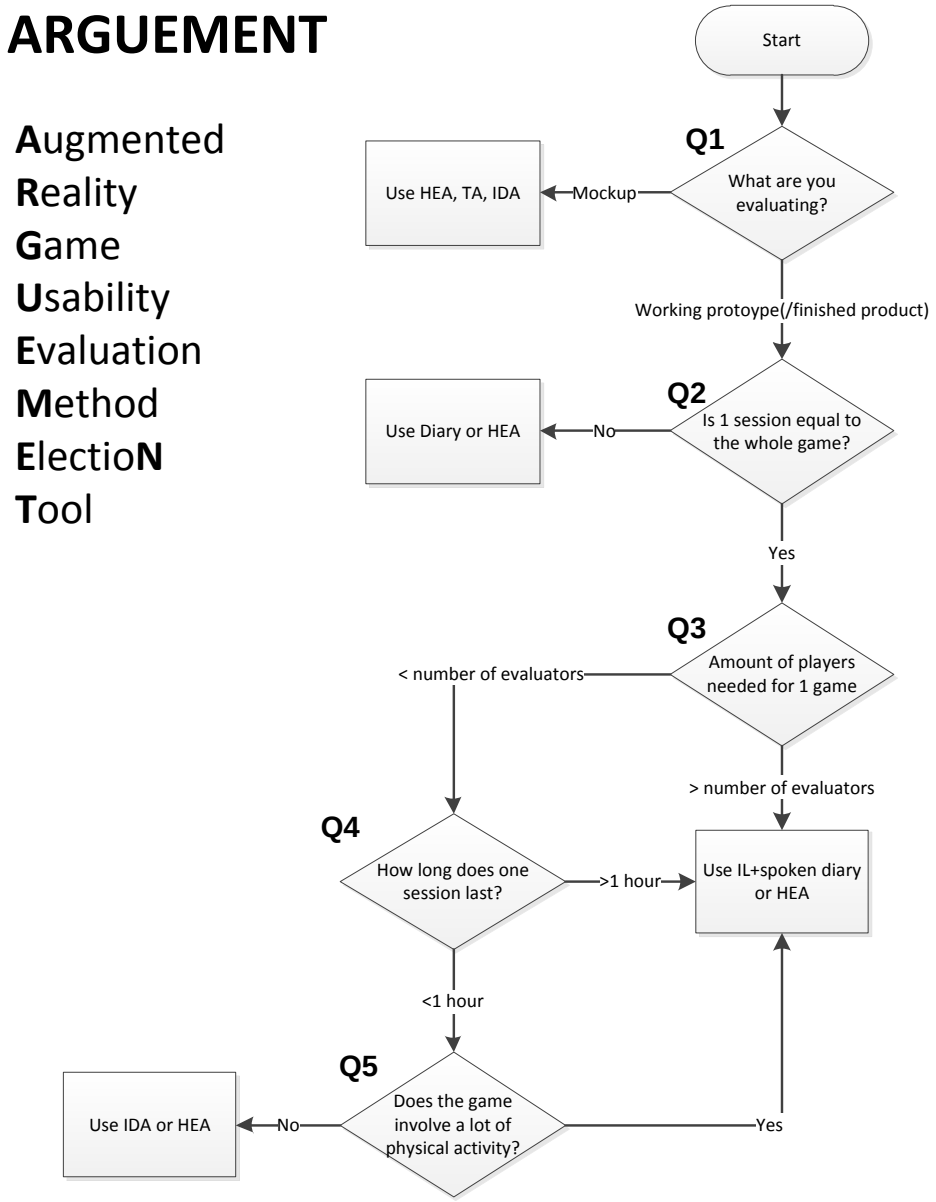


Figure 7.2: Final version of ARGUEMENT, with improvements resulting from the evaluations.

3. If one game lasts for a long time, over an hour, use interaction logs and spoken diaries.
4. If a game takes less than an hour and players do not have to run a lot, use Instant Data Analysis
5. If a game takes less than an hour and players have to run a lot, use a combination of Retrospective Think Aloud and Instant Data Analysis.

These suggestions were transformed in to hypotheses concerning the raw amount of issues detected by the evaluation methods and the median severity of the identified issues by each method. Generally speaking the hypotheses postulated that the suggested methods would find a greater number of issues or more severe issues than a conventional heuristic evaluation would.

As performing a robust quantitative study to validate the hypotheses would require over 1000 participants and more than 200 usability experts it was decided to compare the results using χ^2 test and look

at the results qualitatively.

For each suggestion a suitable game had to be identified to test the suggestion with. For the diary study Parallel Kingdom was used, as this is an extensive game. For both the interaction logs + spoken diary and retrospective think aloud with IDA Portal Hunt was used. This way a lot of time could be saved, fewer evaluations were needed and fewer games needed to be analyzed in general thus saving time and effort without compromising the results of the experiment. Lastly, IDA was tested using the game Tidy City.

When evaluating the existing heuristics that could apply to MMRGs, it quickly became apparent that there were a plethora of relevant sets. Determining which one to use proved a hard job, as there was no definition of usability for MMRGs. Although a definition of usability for video games in general was available it did not take into account the idiosyncrasies that came with the mobile and mixed reality elements. In order to have a suitable definition for this context, the definition of usability for regular video games by Pinelle et al. (2008a) was adapted to the following: *Game usability is the degree to which a player is able to learn, control, understand and safely play a game in the environment it was designed for.*

Looking at all the available heuristics with this definition in mind it became apparent that all of them contained important aspects but none of them covered all the bases to meet the definition. To alleviate this, an adapted list of heuristics was created that merged all the available lists and introduced several new heuristics to better meet the needs for the mobile context. To validate this new list of heuristics an evaluation was performed with both this adapted list and the most suitable list of game heuristics which were the heuristics by Pinelle et al. (2008a). The results were compared the same way as the other hypotheses, i.e. based on the raw amount of issues and the median severity rating of the issues. The evaluations with the adapted heuristics discovered more issues than the heuristics by Pinelle, but their severity ratings did not differ. Going forward with this knowledge, the adapted heuristics were used as a benchmark for the raw amount of issues and both sets of heuristics were used as a benchmark for the severity rating.

After performing an evaluation with each method the results were compared to those of the heuristic evaluation. None of the suggested methods found more severe results than a heuristic evaluation, but for these evaluation sessions they did in most cases identify more issues than the heuristic evaluation. Only the retrospective think aloud with IDA performed clearly worse than the heuristic evaluation. Using a χ^2 though revealed that none of the differences in the discovered amount of issues between the adapted heuristics and the suggested methods was significant. Based on qualitative merits of the suggested methods though, a revised version of ARGUEMENT is introduced in 7.2 along with the suggestion for further validation with other games and/or evaluators.

Acknowledgements

Large parts of this work are based on work performed in the TOTEM project (`http://www.totem-games.org`) which is supported by the Programme Inter Carnot Fraunhofer from BMBF and ANR.

Although my own name is in big letters on the cover and also my supervisors, Herre van Oostendorp, Christof van Nimwegen and Richard Wetzel, are prominently mentioned there this thesis didn't just come to be based on my own elbow grease and their guidance. Many other people have been there for me in all kinds of ways to make sure this was a great success.

In the beginning Hans Voorbij was a great source of information on how to approach going abroad and where interesting stuff might be going on for me. After I managed to secure a place at Fraunhofer FIT it was time to get a room near Sankt Augustin. This unforunately didn't turn out to be as easy as I had hoped with people not showing up for appointments and suddenly only offering two year leases. Having two wonderful parents who are crazy enough to drive to Bonn and back several times sure made that a lot easier, just like moving me there and back when I finally found a room.

Though Sankt Augustin may not be on the other side of the world, it still does help if there's tons of people around you that make you feel at home. Luckily I had just those people both as colleagues and at home. Whenever Richard would be away conquering the world, Leif Oppermann and Lisa Blum would take over his role and try to make sure I didn't burn down the place.

Sharing a room at the institute with Vivek was quite an experience in the most positive of ways, though I should've put up a swearing tin to make some extra money.

Of course I also have to thank all the usability experts that have been willing to help me out with evaluating the games. Audrius Jurgelionis, Paul Vreugdenhil, Ferdy van Varik, Ines van Drie, Jeroen Hulman, Koen Ekelschot and Esther van der Drift thank you all very much for your valuable input that has allowed me to draw up this document.

Part of living in Germany is of course also improving your language skills, I think it's still safe to say that my grammar is "Käse" but I have definitely learned many new words and expressions not in the least thanks to Hagen Buchholtz, who was also invaluable when it came to all things concerning the 3D-printer.

Finally I want to thank my grandfather for always showing interest and being able to ask relevant questions and Dr. Ir. O. Gielkens and Dr. D.E. Gielkens-Keller for the mental support.

References

Adams, E. (2009). *Fundamentals of Game Design* (Second ed.). Berkeley, CA: New Riders.

Alexander, L. (2010). *Nielsen: Current Gen Console Penetration Reaches 41 Percent.* Retrieved from `http://www.gamasutra.com/view/news/26692/Nielsen_Current_Gen_Console_Penetration_Reaches_41_Percent.php` (Retrieved April 18 2011)

Allen, M. (2002, January). A case study of the usability testing of the University of South Floridas virtual library interface design. *Online Information Review*, *26*(1), 40–53. doi: 10.1108/14684520210418374

All the worlds a game. (2011). Retrieved from `http://www.economist.com/node/21541164` (Retrieved April 8, 2012)

Alonso-Rios, D., Luis-Vazquez, I., Mosqueira-Rey, E., Moret-Bonillo, V., & del Rio, B. B. (2009, October). An HTML analyzer for the study of web usability. *2009 IEEE International Conference on Systems, Man and Cybernetics*(October), 1224–1229. doi: 10.1109/ICSMC.2009.5345901

Amaya, G., Davis, J. P., Gunn, D. V., Harrison, C., Pagulayan, R. J., Phillips, B., & Wixon, D. (2008). Games User Research (GUR): Our Experience with and Evolution of Four Methods. In K. Isbister & N. Schaffer (Eds.), *Game usability: advice from the experts for advancing the player experience* (1st ed., pp. 35–65). Burlington, MA: Morgan Kaufmann.

Anderson, J. R. (2004). *Cognitive Psychology.* Worth Publishers Inc.,U.S.

Ardito, C., Lanzilotti, R., Buono, P., & Piccinno, A. (2006, May). A tool to support usability inspection. In *Proceedings of the working conference on advanced visual interfaces - avi '06* (pp. 278–281). New York, New York, USA: ACM Press. doi: 10.1145/1133265.1133322

Asunka, S., Chae, H. S., Hughes, B., & Natriello, G. (2009, January). Understanding Academic Information Seeking Habits through Analysis of Web Server Log Files: The Case of the Teachers College Library Website. *The Journal of Academic Librarianship*, *35*(1), 33–45. doi: 10.1016/j.acalib.2008.10.019

Bartolic, E., Basso, M., Schefft, B., & Glauser, T. (1999, June). Effects of experimentally-induced emotional states on frontal lobe cognitive task performance. *Neuropsychologia*, *37*(6), 677–683. doi: 10.1016/S0028-3932(98)00123-7

Bell, S., McDiarmid, A., & Irvine, J. (2011). *Nodobo Capture: Mobile Data Recording for Analysing User Interactions in Context.* (Submitted to Mobile HCI 2011)

Bias, R. G. (1991). Interface-Walkthroughs: efficient collaborative testing. *IEEE Software*, *8*(5), 94–95. doi: 10.1109/52.84220

Bias, R. G. (1994, June). Usability inspection methods. In J. Nielsen & R. Mack (Eds.), (pp. 63–76). New York: John Wiley & Sons.

Bier, E. A., Stone, M. C., Pier, K., Buxton, W., & DeRose, T. D. (1993, September). Toolglass and magic lenses. In *Proceedings of the 20th annual conference on computer graphics and interactive techniques - siggraph '93* (pp. 73–80). New York, New York, USA: ACM Press. doi: 10.1145/166117.166126

Björk, S., & Holopainen, J. (2004). *Patterns in Game Design (Charles River Media Game Development)* (1st ed. ed.). Hingham, MA.: Charles River Media.

Bolchini, D., & Garzotto, F. (2007). Quality of web usability evaluation methods: an empirical study on MILE+. In *Web information systems engineeringwise 2007 workshops* (pp. 481–492). Springer.

Brewster, S. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing*, *6*(3), 188–205.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 189–194). London: Taylor & Francis.

Brown, B. A. T., Sellen, A. J., & O'Hara, K. P. (2000, April). A diary study of information capture in working life. In *Proceedings of the sigchi conference on human factors in computing systems - chi '00* (pp. 438–445). New York, New York, USA: ACM Press. doi: 10.1145/332040.332472

Bulman, J., Crabtree, B., Gower, A., & Oldroyd, A. (2006). Mixed reality applications in urban environments. In A. Steventon & S. Wright (Eds.), *Intelligent spaces* (pp. 109–124). Springer London.

Burke, J., McNeill, M., Charles, D., Morrow, P., Crosbie, J., & McDonough, S. (2010, March). Augmented Reality Games for Upper-Limb Stroke Rehabilitation. In *2010 second international conference on games and virtual worlds for serious applications* (pp. 75–78). IEEE. doi: 10.1109/VS-GAMES.2010.21

Caoili, E. (2008). *A History of Gaming Platforms: Mattel Intellivision.* Retrieved from `http://www.gamasutra.com/php-bin/news_index.php?story=18518` (Retrieved Februari 21, 2011)

Card, S., Moran, T., & Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale N.J.: L. Erlbaum Associates.

Carter, S., & Mankoff, J. (2005, April). When participants do the capturing. In *Proceedings of the sigchi conference on human factors in computing systems - chi '05* (pp. 899–908). New York, New York, USA: ACM Press. doi: 10.1145/1054972.1055098

Cheong, Y., Jhala, A., Bae, B., & Young, R. (2008). Automatically generating summary visualizations from game logs. In C. Darken & M. Mateas (Eds.), *Proceedings of the fourth artificial intelligence and interactive digital entertainment conference in stanford, california, usa.* The AAAI Press.

Chin, J. P., Diehl, V. A., & Norman, L. K. (1988, May). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the sigchi conference on human factors in computing systems - chi '88* (pp. 213–218). New York, New York, USA: ACM Press. doi: 10.1145/57167.57203

CIA. (2012). *Country Comparison :: GDP (purchasing power parity).* Retrieved from `https://www.cia.gov/library/publications/the-world-factbook/rankorder/2001rank.html` (Accessed April 8, 2012)

Coursaris, C., & Kim, D. (2011). A Meta-Analytical Review of Empirical Mobile Usability Studies. *Journal of Usability Studies*, *6*(3), 117–171.

Cummings, H. M., & Vandewater, E. a. (2007, July). Relation of adolescent video game play to time spent in other activities. *Archives of pediatrics & adolescent medicine*, *161*(7), 684–9. doi: 10.1001/archpedi.161.7.684

Czerwinski, M., Horvitz, E., & Wilhite, S. (2004, April). A diary study of task switching and interruptions. In *Proceedings of the 2004 conference on human factors in computing systems - chi '04* (pp. 175–182). New York, New York, USA: ACM Press. doi: 10.1145/985692.985715

Davidsson, O., Peitz, J., & Björk, S. (2004). *Game design patterns for mobile games* (Tech. Rep.). Finland: Nokia Research Center. Retrieved from `http://procyon.lunarpages.com/~gamed3/docs/Game_Design_Patterns_for_Mobile_Games.pdf`

Dennis, A., Wixom, B. H., & Tegarden, D. (2004). *Systems Analysis and Design with UML Version 2.0: An Object-Oriented Approach* (Int. 2nd ed.). Wiley.

Desurvire, H., Caplan, M., & Toth, J. A. (2004, April). Using heuristics to evaluate the playability of games. In *Extended abstracts of the 2004 conference on human factors and computing systems - chi '04* (pp. 1509–1512). New York, New York, USA: ACM Press. doi: 10.1145/985921.986102

Desurvire, H., Kondziela, J., & Atwood, M. E. (1992). What is gained and lost when using methods other than empirical testing. In *Posters and short talks of the 1992 sigchi conference on human factors in computing systems - chi '92* (p. 125). New York, New York, USA: ACM Press. doi: 10.1145/1125021.1125115

Ducheneaut, N., & Moore, R. (2004). The social side of gaming: a study of interaction patterns in a massively multiplayer online game. In *Proceedings of the 2004 acm conference on computer supported cooperative work* (pp. 360–369). ACM.

Duh, H., Tan, G., & Chen, V. (2006). Usability evaluation for mobile device: a comparison of laboratory and field tests. In *Proceedings of the 8th conference on human-computer interaction with mobile devices and services* (pp. 181–186). ACM.

En, L., & Lan, S. (2010). Social gaminganalysing Human Computer Interaction using a video-diary method. In *Computer engineering and technology (iccet), 2010 2nd international conference on* (Vol. 3, pp. V3–509). IEEE. doi: 10.1109/ICCET.2010.5485833

Ericsson, K., & Simon, H. (1985). *Protocol analysis : verbal reports as data* (1st paperb ed.). Cambridge Mass.: The MIT Press.

Fabricatore, C., Nussbaum, M., & Rosas, R. (2002, December). Playability in Action Videogames: A Qualitative Design Model. *Human-Computer Interaction*, *17*(4), 311–368. doi: 10.1207/S15327051HCI1704_1

Facer, K., Joiner, R., Stanton, D., Reidz, J., Hullz, R., & Kirk, D. (2004). Savannah : mobile gaming

and learning ? *Journal of Computer Assisted Learning*(1980), 399–409.

Federoff, M. (2002). *Heuristics and usability guidelines for the creation and evaluation of fun in video games*. Unpublished master's thesis. Retrieved from `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.8294&rep=rep1&type=pdf` doi: 10.1.1.89.8294

Fernandez, A., Insfran, E., & Abrahão, S. (2011, August). Usability evaluation methods for the web: A systematic mapping study. *Information and Software Technology*, *53*(8), 789–817. doi: 10.1016/j.infsof.2011.02.007

Field, A. (2009). *Discovering Statistics Using SPSS* (Third Edition ed.). London: Sage Publications Ltd.

Froehlich, J. (2009). *the MyExperience blog: MyExperience Version 0.9.1 Released.* Retrieved from `http://myexperiencetool.blogspot.com/2009/08/myexperience-version-091-released.html` (Retrieved July 18, 2011)

Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., & Landay, J. A. (2007, June). MyExperience. In *Proceedings of the 5th international conference on mobile systems, applications and services - mobisys '07* (pp. 57–70). New York, New York, USA: ACM Press. doi: 10.1145/1247660.1247670

Frøkjær, E., & Hornbæk, K. (2002). Metaphors of human thinking in HCI: Habit, stream of thought, awareness, utterance, and knowing. In F. Vetere, L. Johnston, & R. Kushinsky (Eds.), *Proceedings of hf2002 human factors conference/ozchi 2002, melbourne, australia* (pp. 25–27). Swinburne University of Technology.

Frøkjær, E., & Hornbæk, K. (2008, January). Metaphors of human thinking for usability inspection and design. *ACM Transactions on Computer-Human Interaction*, *14*(4), 1–33. doi: 10.1145/1314683.1314688

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. M. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional.

Ghaphery, J. (2005, January). Too quick? Log analysis of Quick Links from an academic library website. *OCLC Systems & Services*, *21*(3), 148–155. doi: 10.1108/10650750510612353

Gielkens, C. (2011, jul). Mobile Mixed Reality Games. *Vakidioot*. Retrieved from `http://www.a-eskwadraat.nl/vakid/vakid1011-6.pdf`

Gielkens, C., & Wetzel, R. (2011, 11). ARe mobile mixed reality games pervasive ? Evaluating mobile mixed reality games in the market for pervasiveness. In (p. 4). Presented at Mobile Gaming workshop at Advances in Computer Entertainment 2011, Lisbon, Portugal.

Gray, W., & Salzman, M. (1998, September). Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Human-Computer Interaction*, *13*(3), 203–261. doi: 10.1207/s15327051hci1303_2

Green, W. (n.d.). *Big Game Hunter*. Retrieved from `http://www.time.com/time/specials/2007/article/0,28804,1815747_1815707_1815665,00.html`

Haak, M. V. D., De Jong, M., & Schellens, P. J. (2003, September). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339–351. doi: 10.1080/0044929031000

Henrysson, A., & Ollila, M. (2004). UMAR: Ubiquitous mobile augmented reality. In *Proceedings of the 3rd international conference on mobile and ubiquitous multimedia* (pp. 41–45). ACM.

Herbst, I., Braun, A., McCall, R., & Broll, W. (2008). TimeWarp: interactive time travel with a mobile mixed reality game. In *Proceedings of the 10th international conference on human computer interaction with mobile devices and services* (pp. 235–244). ACM.

Herman, L. (2001). *Phoenix: The Fall & Rise of Videogames*. Rolenta Press.

Hollingsed, T., & Novick, D. G. (2007, October). Usability inspection methods after 15 years of research and practice. In *Proceedings of the 25th annual acm international conference on design of communication - sigdoc '07* (pp. 249–255). New York, New York, USA: ACM Press. doi: 10.1145/1297144.1297200

Hoonhout, H. C. (2008). Let the Game Tester Do the Talking: Think Aloud and Interviewing to Learn About the Game Experience. In K. Isbister & N. Schaffer (Eds.), *Game usability: advice from the experts for advancing the player experience* (1st ed., pp. 65–79). Burlington, MA: Morgan Kaufmann.

Hornbæk, K., & Frøkjær, E. (2004a, October). Two psychology-based usability inspection techniques studied in a diary experiment. In *Proceedings of the third nordic conference on human-computer interaction - nordichi '04* (pp. 3–12). New York, New York, USA: ACM Press. doi: 10.1145/1028014.1028016

Hornbæk, K., & Frøkjær, E. (2004b, September). Usability inspection by metaphors of human thinking

compared to heuristic evaluation. *International Journal of Human-Computer Interaction*, *17*(3), 357–374. doi: 10.1207/s15327590ijhc1703_4

Hornbæk, K., Hoegh, R., Pedersen, M., & Stage, J. (2007). Use Case Evaluation (UCE): A Method for Early Usability Evaluation in Software Development. *Human-Computer Interaction–INTERACT 2007*, *4662*(4662), 578–591.

Hvannberg, E., Law, E., & Larusdottir, M. (2007, March). Heuristic evaluation: Comparing ways of finding and reporting usability problems. *Interacting with Computers*, *19*(2), 225–240. doi: 10.1016/j.intcom.2006.10.001

Isbister, K., & Schaffer, N. (2008). Interview about Prototyping and Usability with Jenova Chen. In K. Isbister & N. Schaffer (Eds.), *Game usability: advice from the experts for advancing the player experience* (1st ed., pp. 305–309). Burlington, MA: Morgan Kaufmann.

ISO/IEC. (1998). *9241-11 Ergonomic Requirements for Office Work with Visual Display Terminals (Vdts) – Part 11: Guidance on Usability* (Tech. Rep.). International Organization for Standardization.

Jamali, H. R., Nicholas, D., & Huntington, P. (2005, January). The use and users of scholarly e-journals: a review of log analysis studies. In *Aslib proceedings* (Vol. 57, pp. 554–571). doi: 10.1108/00012530510634271

Järvinen, A., Heliö, S., & Mäyrä, F. (2002). Communication and community in digital entertainment services. *University of Tampere: Hypermedia Laboratory Net Series*, *2*, 9–43. Retrieved from `http://www.arts.rpi.edu/public_html/ruiz/public_html/EGDSpring09/readings/CreatingcommunityUniv.ofTampere.pdf`

Jeng, J. (2005). What is usability in the context of the digital library and how can it be measured? *Information Technology and Libraries*, *24*(2), 46–56.

Jensen, K. L. (2009). Recon: capturing mobile and ubiquitous interaction in real contexts. In *Proceedings of the 11th international conference on human-computer interaction with mobile devices and services* (pp. 76:1–76:2). New York, NY, USA: ACM. doi: 10.1145/1613858.1613948

Ji, Y. G., Park, J. H., Lee, C., & Yun, M. H. (2006, July). A Usability Checklist for the Usability Evaluation of Mobile Phone User Interface. *International Journal of Human-Computer Interaction*, *20*(3), 207–231. doi: 10.1207/s15327590ijhc2003_3

Johnson, J., Roberts, T., Verplank, W., Smith, D., Irby, C., Beard, M., & Mackey, K. (1989, September). The Xerox Star: a retrospective. *IEEE Computer*, *22*(9), 11–26. doi: 10.1109/2.35211

Jørgensen, A. H. (2004). Marrying hci/usability and computer games: a preliminary look. In *Proceedings of the third nordic conference on human-computer interaction* (pp. 393–396). New York, NY, USA: ACM. doi: 10.1145/1028014.1028078

Kahn, M. J., & Prail, A. (1994, June). Formal usability inspections. In J. Nielsen & R. Mack (Eds.), *Usability inspection methods* (pp. 141–171). John Wiley & Sons.

Kallio, T., & Kaikkonen, A. (2005). Usability testing of mobile applications: A comparison between laboratory and field testing. *Journal of Usability Studies*, *1*(1), 4–16.

Kim, H., Kim, J., Lee, Y., Chae, M., & Choi, Y. (2002, January). An empirical study of the use contexts and usability problems in mobile Internet. In *Proceedings of the 35th annual hawaii international conference on system sciences* (pp. 1767–1776). IEEE Comput. Soc. doi: 10.1109/HICSS.2002.994090

Kirakowski, J. (1996). The Software Usability Measurement Inventory: Background and usage. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 169–178). London: Taylor & Francis.

Kjeldskov, J., Skov, M. B., Als, B. S., & Høegh, R. T. (2004). Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In S. Brewster & M. Dunlop (Eds.), *Proceedings of the 6th international mobile hci 2004 conference* (pp. 529–535). Springer-Verlag.

Kjeldskov, J., Skov, M. B., & Stage, J. (2004, October). Instant data analysis: conducting usability evaluations in a day. In *Proceedings of the third nordic conference on human-computer interaction - nordichi '04* (pp. 233–240). New York, New York, USA: ACM Press. doi: 10.1145/1028014.1028050

Kjeldskov, J., & Stage, J. (2004, May). New techniques for usability evaluation of mobile systems. *International Journal of Human-Computer Studies*, *60*(5-6), 599–620. doi: 10.1016/j.ijhcs.2003.11.001

Korhonen, H., & Koivisto, E. M. I. (2006, September). Playability heuristics for mobile games. In *Proceedings of the 8th conference on human-computer interaction with mobile devices and services - mobilehci '06* (pp. 9–16). New York, New York, USA: ACM Press. doi: 10.1145/1152215.1152218

Kücklich, J. (2004). *Play and playability as key concepts in new media studies.* Dublin: Dublin City

University.

Kurniawan, S. (2008, December). Older people and mobile phones: A multi-method investigation. *International Journal of Human-Computer Studies*, *66*(12), 889–901. doi: 10.1016/j.ijhcs.2008.03.002

Laitinen, S. (2006). Do usability expert evaluation and test provide novel and useful data for game development. *Journal of Usability Studies*, *1*(2), 64–75.

Laitinen, S. (2008). Usability and Playability Expert Evaluation. In K. Isbister & N. Schaffer (Eds.), *Game usability: advice from the experts for advancing the player experience* (1st ed., pp. 91–112). Burlington, MA: Morgan Kaufmann.

Lankoski, P., Heliö, S., Nummela, J., Lahti, J., Mäyrä, F., & Ermi, L. (2004). A case study in pervasive game design: the songs of north. In *Proceedings of the third nordic conference on human-computer interaction* (pp. 413–416). ACM.

Lewis, C., Polson, P., Wharton, C., & Rieman, J. (1990, March). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the sigchi conference on human factors in computing systems empowering people - chi '90* (pp. 235–242). New York, New York, USA: ACM Press. doi: 10.1145/97243.97279

Lewis, J. (1992). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Human factors and ergonomics society annual meeting proceedings* (Vol. 36, pp. 1259–1263). Human Factors and Ergonomics Society.

Matera, M., Costabile, M., Garzotto, F., & Paolini, P. (2002). SUE inspection: an effective method for systematic usability evaluation of hypermedia. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *32*(1), 93–103. doi: 10.1109/3468.995532

Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems E series D*, *77*(12), 1321–1321.

Montola, M. (2009). Games and Pervasive Games. In M. Montola, J. Stenros, & A. Waern (Eds.), *Pervasive games: Theory and design* (pp. 7–23). Morgan Kaufmann.

Montola, M., Stenros, J., & Waern, A. (2009). *Pervasive Games: Theory and Design*. Morgan Kaufmann.

Mulder, I., Ter Hofte, G., & Kort, J. (2005). SocioXensor: Measuring user behaviour and user eXperience in conteXt with mobile devices. In *Proceedings of measuring behavior* (Vol. 2005, pp. 355–358).

Nacke, L. (2009, May). From playability to a hierarchical game usability model. In *Proceedings of the 2009 conference on future play on @ gdc canada - futureplay '09* (p. 11). New York, New York, USA: ACM Press. doi: 10.1145/1639601.1639609

Naur, P. (1995). *Knowing and the mystique of logic and rules*. Dordrecht, The Netherlands: Kluwer Academic.

Nielsen, C. (1998). Testing in the Field. In *Computer human interaction, 1998. proceedings. 3rd asia pacific* (pp. 285–290). IEEE.

Nielsen, C. M., Overgaard, M., Pedersen, M., Stage, J., & Stenild, S. (2006, October). It's worth the hassle! In *Proceedings of the 4th nordic conference on human-computer interaction changing roles - nordichi '06* (pp. 272–280). New York, New York, USA: ACM Press. doi: 10.1145/1182475.1182504

Nielsen, J. (1990). *10 Heuristics for User Interface Design*. `http://www.useit.com/papers/heuristic/heuristic_list.html`. (Retrieved March 17, 2011)

Nielsen, J. (1992, June). Finding usability problems through heuristic evaluation. In *Proceedings of the sigchi conference on human factors in computing systems - chi '92* (pp. 373–380). New York, New York, USA: ACM Press. doi: 10.1145/142750.142834

Nielsen, J. (1994a). Heuristic Evaluation. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (1st ed., pp. 25–62). John Wiley & Sons.

Nielsen, J. (1994b, April). Usability inspection methods. In *Conference companion on human factors in computing systems - chi '94* (pp. 413–414). New York, New York, USA: ACM Press. doi: 10.1145/259963.260531

Nielsen, J. (1995). *Severity Ratings for Usability Problems*. Retrieved from `http://www.useit.com/papers/heuristic/severityrating.html`

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the sigchi conference on human factors in computing systems - chi '93* (pp. 206–213). New York, New York, USA: ACM Press. doi: 10.1145/169059.169166

Nilsen, T., Linton, S., & Looser, J. (2004). Motivations for augmented reality gaming. *Proceedings of FUSE*, *4*, 86–93.

Nokia. (n.d.). *Snake game - Mobile revolution - Story of Nokia - Company - About Nokia*. `http://www.nokia.com/about-nokia/company/story-of-nokia/mobile-revolution/snake-game`. (Re-

trieved Februari 21, 2011)

OECD. (2009). *OECD factbook 2009: economic, environmental and social statistics.* Organisation for Economic Co-operation and Development.

Okada, H., & Asahi, T. (1999). GUITESTER: A log-based usability testing tool for Graphical User Interfaces. *IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E SERIES D*, *82*(6), 1030–1041.

Palen, L., & Salzman, M. (2002, November). Voice-mail diary studies for naturalistic data capture under mobile conditions. In *Proceedings of the 2002 acm conference on computer supported cooperative work - cscw '02* (pp. 87–95). New York, New York, USA: ACM Press. doi: 10.1145/587078.587092

Palen, L., Salzman, M., & Youngs, E. (2001, July). Discovery and Integration of Mobile Communications in Everyday Life. *Personal and Ubiquitous Computing*, *5*(2), 109–122. doi: 10.1007/s007790170014

Pinelle, D., Wong, N., & Stach, T. (2008a). Heuristic evaluation for games. In *Proceeding of the twenty-sixth annual chi conference on human factors in computing systems - chi '08* (pp. 1453–1462). New York, New York, USA: ACM Press. doi: 10.1145/1357054.1357282

Pinelle, D., Wong, N., & Stach, T. (2008b, November). Using genres to customize usability evaluations of video games. In *Proceedings of the 2008 conference on future play research, play, share - future play '08* (pp. 129–136). New York, New York, USA: ACM Press. doi: 10.1145/1496984.1497006

Pinelle, D., Wong, N., Stach, T., & Gutwin, C. (2009, May). Usability heuristics for networked multiplayer games. In *Proceedings of the acm 2009 international conference on supporting group work - group '09* (pp. 169–178). New York, New York, USA: ACM Press. doi: 10.1145/1531674.1531700

Po, S., Howard, S., Vetere, F., & Skov, M. (2004). Heuristic evaluation and mobile usability: Bridging the realism gap. *Mobile Human-Computer InteractionMobileHCI 2004*, *3160*, 591–592. doi: 10.1007/b100594

Polson, P., Lewis, C., Rieman, J., & Wharton, C. (1992, May). Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies*, *36*(5), 741–773. doi: 10.1016/0020-7373(92)90039-N

Raento, M., Oulasvirta, A., Petit, R., & Toivonen, H. (2005, April). ContextPhone: A Prototyping Platform for Context-Aware Mobile Applications. *IEEE Pervasive Computing*, *4*(2), 51–59. doi: 10.1109/MPRV.2005.29

Rashid, O., Bamford, W., Coulton, P., Edwards, R., & Scheible, J. (2006). PAC-LAN: mixed-reality gaming with RFID-enabled mobile phones. *Computers in Entertainment (CIE)*, *4*(4), 1–17.

Salen, K., & Zimmerman, E. (2003). *Rules of Play: Game Design Fundamentals.* Cambridge, MA: MIT Press.

Sampanes, A., Snyder, M., Rampoldi-Hnilo, L., & White, B. (2011). Photo DiariesA Peek into a Mobile Workers Life. *Design, User Experience, and Usability. Theory, Methods, Tools and Practice*(6769), 640–647.

Sauro, J., & Kindlund, E. (2005, April). A method to standardize usability metrics into a single score. In *Proceedings of the sigchi conference on human factors in computing systems - chi '05* (pp. 401–409). New York, New York, USA: ACM Press. doi: 10.1145/1054972.1055028

Sawyer, P., Flanders, A., & Wixon, D. (1996, April). Making a difference—the impact of inspections. In *Proceedings of the sigchi conference on human factors in computing systems common ground - chi '96* (pp. 376–382). New York, New York, USA: ACM Press. doi: 10.1145/238386.238579

Schmettow, M. (2005). Towards a pattern based usability inspection method for industrial practitioners. In *Workshop on integrating software engineering and usability engineering in verbindung mit der international conference on human-computer interaction, rom.*

Schmettow, M., & Niebuhr, S. (2007, September). A pattern-based usability inspection method: first empirical performance measures and future issues. In *Bcs-hci '07 proceedings of the 21st british hci group annual conference on people and computers: Hci...but not as we know it* (pp. 99–102).

Schutte, N., & Schuettpelz, E. (2001). Emotional intelligence and task performance. *Imagination, Cognition and Personality*, *20*(4), 347–354.

Seibert, P. S., & Ellis, H. C. (1991, September). Irrelevant thoughts, emotional mood states, and cognitive task performance. *Memory & Cognition*, *19*(5), 507–513. doi: 10.3758/BF03199574

Sharp, H., Rogers, Y., & Preece, J. (2007). *Interaction Design: Beyond Human-Computer Interaction.* Wiley.

Shepard, C., Rahmati, A., Tossell, C., Zhong, L., & Kortum, P. (2011, January). LiveLab. *ACM SIGMETRICS Performance Evaluation Review*, *38*(3), 15. doi: 10.1145/1925019.1925023

Soubeyrand, C. (2000). *The Royal Game of Ur.* Retrieved from `http://www.gamecabinet.com/history/Ur.html` (Retrieved April 17, 2011)

Spencer, R. (2000, April). The streamlined cognitive walkthrough method, working around social constraints encountered in a software development company. In *Proceedings of the sigchi conference on human factors in computing systems - chi '00* (pp. 353–359). New York, New York, USA: ACM Press. doi: 10.1145/332040.332456

Spiteri, L., Tarulli, L., & Graybeal, A. (2010, November). The public library catalogue as a social space: Transaction log analysis of user interaction with social discovery systems. In *Proceedings of the american society for information science and technology* (Vol. 47, pp. 1–2). doi: 10.1002/meet.14504701307

Tan, C., & Soh, D. (2010). Augmented Reality Games: A Review. In *Proceedings of gameon-arabia, eurosis.*

Thawonmas, R., & Iizuka, K. (2008). Visualization of Online-Game Players Based on Their Action Behaviors. *International Journal of Computer Games Technology*, 1–9. doi: 10.1155/2008/906931

Thomas, B., Close, B., Donoghue, J., Squires, J., Bondi, P. D., & Piekarski, W. (2002, February). First Person Indoor/Outdoor Augmented Reality Application: ARQuake. *Personal and Ubiquitous Computing*, *6*(1), 75–86. doi: 10.1007/s007790200007

Thompson, K., Rozanski, E., & Haake, A. (2004). Here, there, anywhere: remote usability testing that works. In *Proceedings of the 5th conference on information technology education* (pp. 132–137). ACM.

Tomitsch, M., Singh, N., & Javadian, G. (2010, November). Using diaries for evaluating interactive products. In *Proceedings of the 22nd conference of the computer-human interaction special interest group of australia on computer-human interaction - ozchi '10* (p. 204). New York, New York, USA: ACM Press. doi: 10.1145/1952222.1952266

Triacca, L., Inversini, A., & Bolchini, D. (2005). Evaluating web usability with mile+. In *Proceedings of the seventh ieee international symposium on web site evolution* (pp. 22–29). Washington, DC, USA: IEEE Computer Society. doi: 10.1109/WSE.2005.6

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., & Bergel, M. (2002). An empirical comparison of lab and remote usability testing of Web sites. In *Usability professionals association conference.* Retrieved from `http://home.comcast.net/~tomtullis/publications/RemoteVsLab.pdf`

Van Velsen, L., Van Der Geest, T., Klaassen, R., & Steehouder, M. (2008, September). User-centered evaluation of adaptive and adaptable systems: a literature review. *The Knowledge Engineering Review*, *23*(03), 261–281. doi: 10.1017/S0269888908001379

Varsaluoma, J. (2009). Scenarios in the Heuristic Evaluation of Mobile Devices: Emphasizing the Context of Use. *Human Centered Design*, 332–341.

Viggers, C. (2011). *Usability Testing: Face the Fear and Learn to Love It.* Retrieved from `http://www.gamasutra.com/view/feature/6431/usability_testing_face_the_fear_.php` (Accessed July 18, 2011)

Vogiazou, Y., Raijmakers, B., Geelhoed, E., Reid, J., & Eisenstadt, M. (2006, June). Design for emergence: experiments with a mixed reality urban playground game. *Personal and Ubiquitous Computing*, *11*(1), 45–58. doi: 10.1007/s00779-006-0068-5

Waterson, S., Landay, J. A., & Matthews, T. (2002). *In the lab and out in the wild.* New York, New York, USA: ACM Press. doi: 10.1145/506443.506602

Wetzel, R., Blum, L., Broll, W., & Oppermann, L. (2011). Designing mobile augmented reality games. In B. Furht (Ed.), *Handbook of augmented reality* (1st ed., pp. 513–529). Springer.

Wetzel, R., Blum, L., Feng, F., Oppermann, L., & Straeubig, M. (2011). Tidy City: A Location-based Game for City Exploration Based on User-created Content. In *Proceedings of mensch & computer 2011* (pp. 487–498). Chemnitz, Germany.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994, June). The cognitive walkthrough method: a practitioner's guide. In J. Nielsen & R. Mack (Eds.), *Usability inspection methods* (pp. 105–140). New York: John Wiley & Sons.

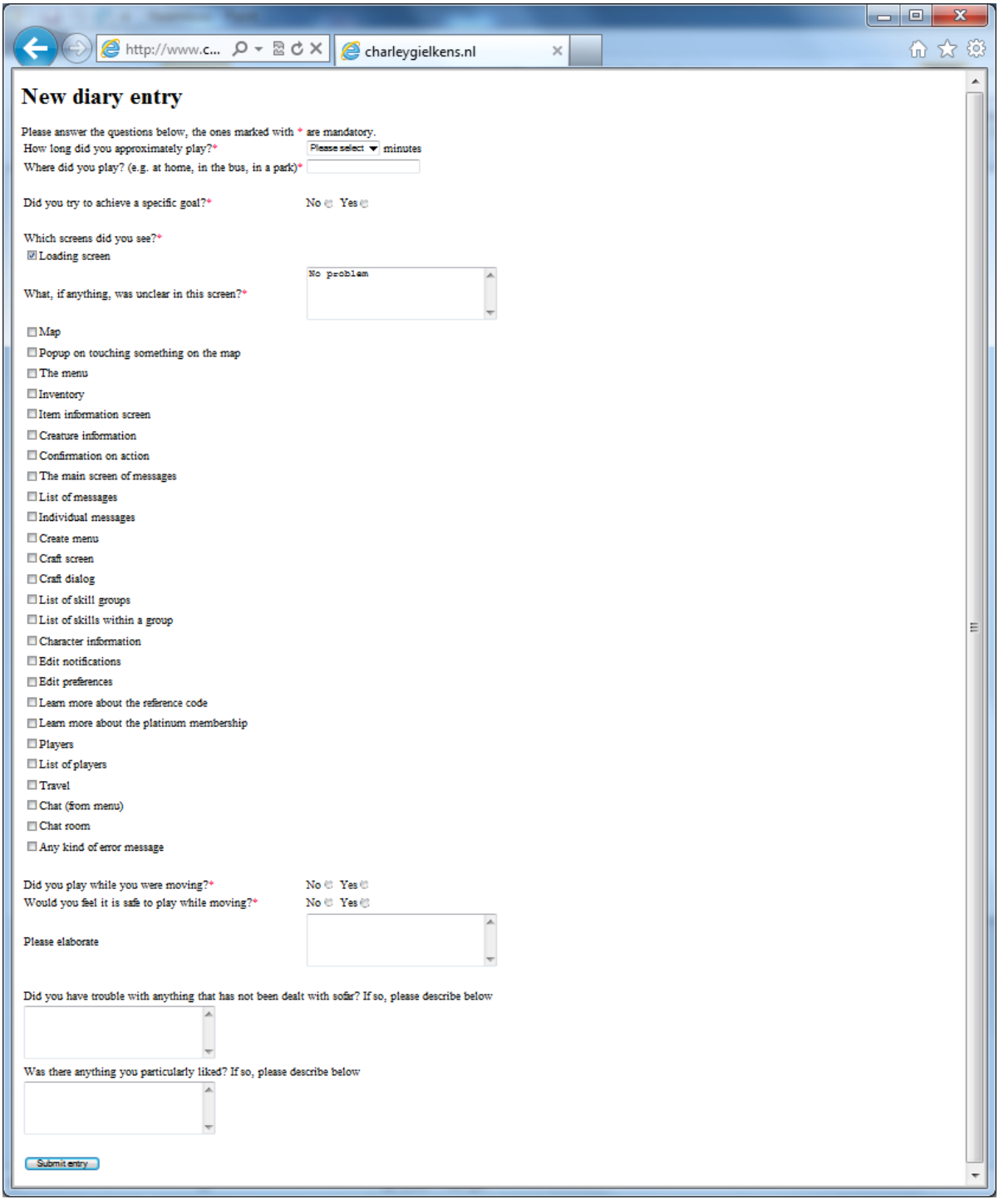
Wixon, D., Jones, S., Tse, L., & Casaday, G. (1994, June). Inspections and design reviews: framework, history and reflection. In J. Nielsen & R. Mack (Eds.), *Usability inspection methods* (pp. 77–103). New York: John Wiley & Sons.

Zhang, D., & Adipat, B. (2005). Challenges, methodologies, and issues in the usability testing of mobile applications. *International Journal of Human-Computer Interaction*, *18*(3), 293–308. doi: 10.1207/s15327590ijhc1803

Zhang, Z., Basili, V., & Shneiderman, B. (1999). Perspective-based Usability Inspection: An Empirical Validation of Efficacy. *Empirical Software Engineering*, *4*(1), 43–69–69. doi: 10.1023/A:1009803214692

Appendix A

Diary screenshot



Appendix B

Full list of games

B.1 Games used in the study

B.1.1 Mister X Mobile

Mister X is an MR version of the board game with the same name[1], in which a team of up to five detectives have to catch a "spy". Both the detectives and the spy get a map display on which the current location of the detectives is always visible, but only the last known location of the spy is shown. This location is updated in set, large intervals so that it is not too easy for the detectives to find the spy.

On the map all the players can see a circle, within which the action is to take place. When Mister X leaves this circle his position is shown to the detectives. Also shown on the map are coins that can be picked up and give all the players extra items, on top of the ones they are given at the start of the game. Both detectives and Mister X have items that can help them or hinder the other team. The detectives can e.g. let the spy's device emit a loud, high pitched noise that can help them locate him when they are in the vicinity. The spy on the other hand can for example extend the time between two updates of his location.

If the detectives have entered their telephone numbers they can call each other directly from the game.

Location coupling: weak
Social interaction required: some for the detectives (optional)
Persistence of the game world: one game

Figure B.1: Screenshot of Mister X Mobile. Source: Android Market

B.1.2 Parallel Kingdom

Parallel Kingdom is a MMORPG with the game world overlaid on top of a map of the real world. MMO (Massively Multiplayer Online) refers to the fact that many players share the same persistent game world via the internet simultaneously, rather than a limited amount of players in a non-persistent world. It is not used to differentiate purely between online and offline games in this context.

You can move in the game world either by moving in real life or by making use of game mechanics. By changing your location though, you can travel to places where you may not be able to go (easily) using the game mechanics.

Players can claim land by planting flags and building houses they can travel to via in game mechanics, as well as join cities. Although it is possible to communicate with other players, cooperation is not possible.

Location coupling: weak
Social interaction required: some
Persistence of the game world: persistent

Figure B.2: Screenshot of Parallel Kingdom. Source: Android Market

[1] In some countries the game is known as Scotland Yard

B.1.3 PhotoFari

Figure B.3: Screenshot of PhotoFari. Source: Android Market

In PhotoFari, players have to take pictures that match a pattern for at least 70%. The patterns are individually presented in a list from which the player can select one he wants to try to match. The selected pattern is shown as an overlay, using a magic lens display (Bier et al., 1993). When a pattern is successfully matched, a new pattern is unlocked.
Location coupling: none
Social interaction required: none
Persistence of the game world: persistent

B.1.4 Portal Hunt

Figure B.4: Someone playing Portal Hunt.

On preset locations portals have opened and your goal in Portal Hunt is to catch them. You can see the portal through a magic lens, but also on a map. Portals can show three types of behavior. They can either stand still, move constantly about or jump between locations. There are also three types of portals, that each require a different tactic to be caught. Some portals can be caught by a single player, others require everybody from the same team to surround the portal and the third type requires that at least two people of the same team surround the portal.

Players can compete with each other by joining different teams.
Location coupling: Weak
Social interaction required: Some
Persistence of the game world: One game

B.1.5 Seek 'n Spell

Figure B.5: Screenshot of Seek 'n Spell. Source: Android Market

This game is an MR version of the classic game Rummikub (with letters). Players walk around in a predefined radius, collecting letters to form words in a set amount of time. The letters that can still be collected are shown on a map and the letters that you have already collected are shown below it. Depending on the length of the words players form, they get points. The winner of the game is the person with the most points.
Location coupling: none
Social interaction required: none
Persistence of the game world: one game

B.1.6 Spectrek

Figure B.6: Screenshot of SpecTrek. Source: Android Market

In SpecTrek ghosts are roaming the neighborhood and you have to catch them. You do this by walking around in the real world and using the map to navigate towards ghosts. When you are near enoug, you can use the devices camera as a magic lens to see them and use items to catch them. Because this all has to be accomplished within a time limit, which the user can set, you often are required to run to make it. The time limit also determines the amount of ghosts and how far they are spread out.

By catching ghosts, you gather experience points which you can use to increase skills like how far away you can see or catch ghosts using the magic lens.
Location coupling: weak
Social interaction required: none
Persistence of the game world: player progress is saved, has no direct influence on the world though

B.1.7 Tidy City

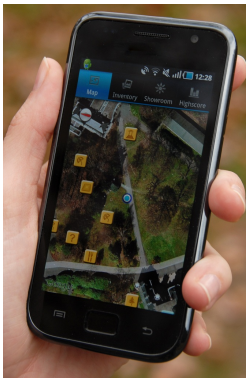


Figure B.7: Screenshot of Tidy City. Source: `http://goo.gl/t9NAT`

Tidy City is a slow paced game that focusses on solving riddles concerning locations. Once a mission, a related set of riddles, is selected the riddles are displayed on a map and the area of the game is bounded by an orange line. When the player is close enough to a riddle, they can view the riddle text and picture and place the riddle in their inventory. If they are too far removed from the riddle, they can just see the title and a hint to go closer.

To solve a riddle the player has to read the riddle text and look at the image, to figure out where the image was taken. When he has figured this out, he will have to go there and click the solve button. If you are in the right place, you will gain points. Should you be in the wrong place, you will lose points.

There is a competitive element to the game, because per mission a highscore list available so that players can compare their results. Players can cooperate by using the same device.

Location coupling: Weak
Social interaction required: none (can be played alone, though playing together with multiple players is possible too)
Persistence of the game world: persistent

B.2 Other games

B.2.1 Android Hunt

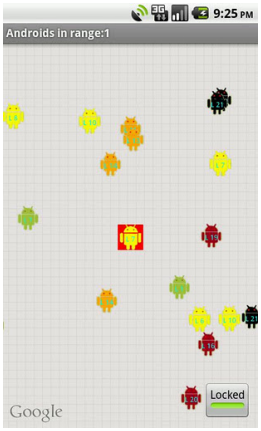


Figure B.8: Screenshot of Android Hunt. Source: Android Market

Android hunt lets players fight with other players that are in the vicinity in MMORPG style. Successful attacks allow the player to gather rewards and improve their skills.

It is also possible to play the game in single player mode. The difference with the multiplayer mode, is that now the other players that are shown on the map are virtual.

Location coupling: weak
Social interaction required: some
Persistence of the game world: persistent

B.2.2 AR Bots

Figure B.9: Screenshot of ARBots. Source: `http://www.totem-games.org`

In AR Bots (url: `http://goo.gl/32XW4`) evil robots from another dimension are invading the world and you have to stop them. You can do this by combatting them with your own, good, robots. All the game elements are displayed through a magic lens display.

Location coupling: Weak
Social interaction required: None
Persistence of the game world: One game

B.2.3 ConquAR

Figure B.10: Screenshot of ConquAR. Source: `http://www.cnqar.com`

As with Mister X, this is an AR/MR version of a well known board game: Risk. Users have to conquer cities and places by traveling there and attacking the local armies using the same mechanics as the board game. Players can see the game world, that is overlaid on the real world, via a magic lens, a map and a list.

Location coupling: weak
Social interaction required: none
Persistence of the game world: persistent

Figure B.11: Screenshot of DJ Rivals. Source: Android Market

B.2.4 DJ Rivals

This is an MMORPG in which the player takes on the role of DJ that has to defeat the commercial music industry drones as well as other players. You do this by entering virtual establishments which carry names of real world businesses in your surroundings. Here you can challenge drones to DJ'ing battles.

Battles consist of you performing certain "musical moves" by interacting to the beat of the music. Based on how well you do this your attack does more or less damage. Challenging players works in much the same way, except the top players you have beat can only be found in certain types of businesses.

Succesfully beating a drone or player will be rewarded with items, money and experience points that all can be used to improve your characters statistics and attacks.

Location coupling: weak
Social interaction required: Optional
Persistence of the game world: persistent

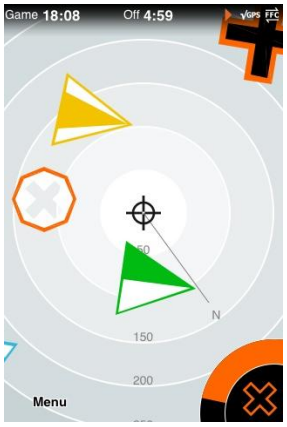


Figure B.12: Screenshot of Fastfoot. Source: `www.androidpit.com`

B.2.5 FastFoot GPS Jump'n'Run

Fastfoot is for the largest part similar to Mister X. The main differences are that you do not have a street map and capturing the runner is mediated by the phone. X, i.e. the one fleeing, is caught when they are within 30 meters of a runner, i.e. the ones doing the capturing. Also, the headstart of X is 6 minutes, rather than the default 2 in Mister X.

Location coupling: None
Social interaction required: A lot
Persistence of the game world: One game

Figure B.13: Screenshot of GPS Earth Defense. Source: Android Market

B.2.6 GPS earth defense

The earth is being invaded by aliens and you have to stop them, alone or in a team. The game is probably an MMORPG, but due to a lack of other players I am not able to definitively prove this. Aliens seem to appear randomly when the game is started and the player has to kill them. If you succeed you get experience and items that can help you improve your skills. The game elements are shown on top of a map and the only way for the player to move in game, is to move in the real world.

According to the description of the game in the Android Market, working together becomes necessary at some point because the aliens become to strong.

Location coupling: None
Social interaction required: Optional
Persistence of the game world: Persistent

Figure B.14: Screenshot of GPS Invaders. Source: Android Market

Figure B.15: Screenshot of Home Invasion. Source: Android Market

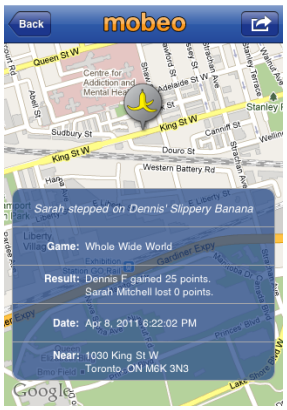


Figure B.16: Screenshot of Mobeo. Source: `www.appdork.com`

B.2.7 GPS invaders

This game is an MR take on the classic arcade game Spave Invaders. As with the classic game, the player sees invading alien ships coming closer and closer to their ship on the map and the goal is to destroy them before they reach you. The MR part in this game is that instead of moving the ship by pushing buttons, the player actually has to run left, right, back and front to move their ship.
Location coupling: None
Social interaction required: None
Persistence of the game world: Persistent

B.2.8 Home Invasion

This game lets the user invade the neighborhood in which he currently is, in the guise of a flying saucer that has to collect dots that are spread out along the streets in the neighborhood. Hindering you are tanks that also follow the roads and will shoot at you if they are within range.
Location coupling: weak
Social interaction required: none
Persistence of the game world: none

B.2.9 Mobeo

In the game Mobeo, players join or create a game world that is overlaid on top of the real world using a map. In this world, they attack the other players by throwing weaponry like flaming cabbages at the location where they believe the other player is. Alternatively, they can also lay traps like banana peels where they know the others often pass by.

Successfully hitting another player gives the players points they can use to access better weapons.

As of October 2011, the game is no longer supported.
Location coupling: Weak
Social interaction required: Some
Persistence of the game world: Persistent

Figure B.17: Screenshot of Nuclear. Source: Android Market

B.2.10 Nuclear

In nuclear the player has to locate and diffuse a nuclear bomb. The game interface consists of a Geiger counter to determine the proximity of the bomb and an electromagnetic pulse to disable the bomb.
Location coupling: weak
Social interaction required: none
Persistence of the game world: one game

B.2.11 Outbreak, zombie apocalypse

Instead of plainly escaping zombies, the player either has to kill as many zombies as possible or infect as many healthy people as possible in this game. Both healthy people and zombies are played by people.

Combatting the players on the other team is done by travelling through the world and when you are within reach of a player on the other team you can select them and attack them. There are no other game mechanics for moving around.
Location coupling: weak
Social interaction required: some
Persistence of the game world: persistent

Figure B.18: Screenshot of Outbreak, Zombie Apocalypse. Source: Android Market

B.2.12 Phone bomber

Just like Parallel Kingdom, this game is a MMORPG which means that all the players inhabit a persistent game world. The goal of the game is to blow up the other players by placing virtual bombs somewhere you expect one or more players will be when the timer on the bomb reaches 0. If you succeed you gain experience points which will allow you to buy better equipment. It is also possible to try and defuse bombs, though this has the risk of setting them off. If you succeed you also gain experience points. The bombs are displayed either in a list view or on a map.

Other than blowing people up and placing contracts on certain locations, there is no way to actually communicate with other players.
Location coupling: Weak
Social interaction required: Some
Persistence of the game world: Persistent

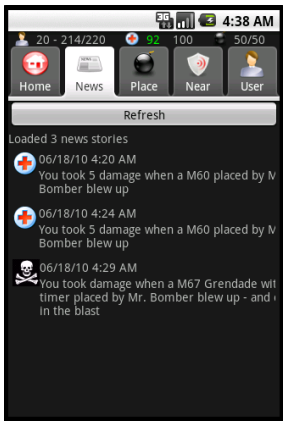


Figure B.19: Screenshot of Phone Bomber. Source: Android Market

Figure B.20: Screenshot of scavengar. Source: `www.scavengarhunt.com`

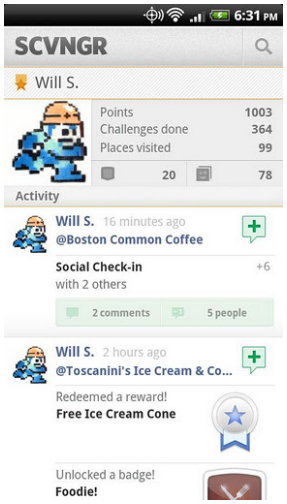


Figure B.21: Screenshot of SCVNGR. Source: Android Market

Figure B.22: Screenshot of ThirdEye. Source: `http://www.viewdle.com`

B.2.13 ScavengAR hunt

This is a digital take on the real life pass time scavenger hunt. Instead of collecting tangible objects, players look for digital objects and collect them. In doing so, they gain points and rise on the leader board. Players can see the game world, that is overlaid on the real world, via a magic lens, a map and a list.
Location coupling: weak
Social interaction required: None in game.
Persistence of the game world: The leader board is persistent, but the persistence of the items is unknown.

B.2.14 SCVNGR

SVNGR is a scavenger hunt game in which players complete tasks at certain locations to qualify for both virtual and real rewards. These tasks can be set by normal people, but also by companies. The latter is most often the case where real world rewards are involved.
Location coupling: Weak
Social interaction required: Unclear
Persistence of the game world: Persistent

B.2.15 ThirdEye

In ThirdEye the player is either a vampire or a slayer who has to build an army to defeat the other side. Sides are determined by using facial recognition. Building an army takes places by scanning other peoples faces and then taking action if they are not on your side. Not long after testing the game, it was removed from the market.
Location coupling: none
Social interaction required: a lot
Persistence of the game world: persistent

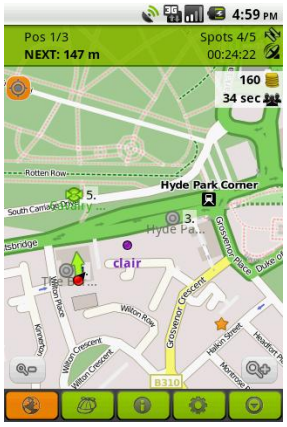


Figure B.23: Screenshot of Tourality. Source: Android Market

B.2.16 Tourality/YouCatch

Tourality is platform and YouCatch is one of the games that runs on it. YouCatch is games that focusses purely on letting the players run. The game can be played alone, with multiple players against each other or in teams which compete against each other. In the single player mode, the goal is to reach goals quicker than you have done before or within a certain time limit. When playing with multiple players, the goal is to reach the designated spots before the other players do so. If you succeed in doing so, you are rewarded virtual gold that you can use to buy powerups.

Both the points that need to be reached and the other players, can be seen on a map.

Location coupling: Weak
Social interaction required: Some
Persistence of the game world: One game

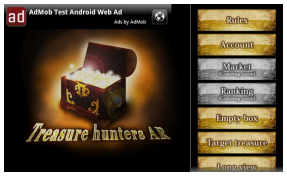


Figure B.24: Screenshot of Treasure Hunters AR. Source: Android Market

B.2.17 Treasure hunters AR

Players bury a virtual treasure chest somewhere and the longer it stays burried the more valuable it becomes. It is also possible to look for treasures that have been buried by other people using a map and when near enough a magic lens interface.

When you find a treasure you can dig it up and get the gold that is inside. The goal of the game is to acquire as much gold as possible.

Location coupling: weak
Social interaction required: none
Persistence of the game world: persistent

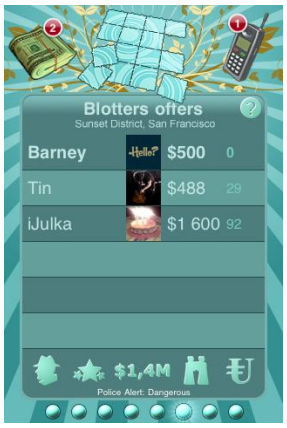


Figure B.25: Screenshot of Underworld. Source: Android Market

B.2.18 Underworld

In Underworld, the players poses as a drug dealer. The goal of the game is to get as much money as possible by buying drugs at a low price from labs and other players, and selling them for a profit to junkies or other players. Both junkies and labs are game entities that are somewhat randomly distributed over the tiles of a map.

This game again is an MMORPG, as the same gameworld is shared by all players that can improve their skills by succesfully completing transactions.

Location coupling: None
Social interaction required: Optional
Persistence of the game world: Persistent

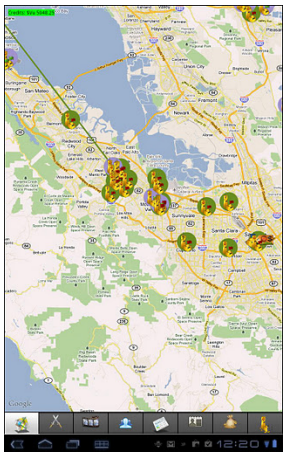


Figure B.26: Screenshot of VuHunt. Source: Android Market

B.2.19 VuHunt

VuHunt allows the player to build their own castles and conquer castles by others. Attacking the castles is done by selecting them on a map and then (trying) to complete a challenge. A challenge can be answering a question, making a picture or movieclip that contains a certain object or checking in at the actual place.

Succesfully conquering a castle is rewarded by ownership of the castle, an in-game monetary reward and daily revenue of the castle in in-game money. This money can be used to change the challenge for a castle you own and for attacking other castles. The further away a castle is, the more expensive it is to attack it. So although it is not actually necessary for every castle to physically move there to attack it, doing so can give you a benefit.

Social interaction is not required, apart from attacking the castles that are owned by other players.
Location coupling: Weak
Social interaction required: Some
Persistence of the game world: Persistent

Figure B.27: Screenshot of Mister X. Source: Android Market

B.2.20 Woomba Mania

In Woomba Mania the player has to catch virtual, fluffy creatures that have escaped from a circus. There are bad creatures too and between dusk and dawn all the creatures turn into ghosts which either turn into good or bad creatures when caught. The good creatures give you points, the bad ones cost points. Players can see the game world, that is overlaid on the real world, via a magic lens, a map and a list.
Location coupling: Weak
Social interaction required: None
Persistence of the game world: Persistent

Figure B.28: Screenshot of Zombie, Run! Source: Android Market

B.2.21 Zombie, run!

In zombie run the player has to get from his current location to a new location of his choosing without getting caught by (virtual) zombies. The game gives the player a map on which the current location of the zombies is displayed, so that the player can navigate around them.
Location coupling: weak
Social interaction required: only in multiplayer
Persistence of the game world: one game

Appendix C

List of MMRG heuristics

C.1 General usability heuristics for mobile mixed reality games

C.1.1 Audio-visual representation supports the game and are easy to interpret

Since MMRGs are meant to be played outside, it is important to keep this in mind when designing the interface, both the graphical and auditive parts. In bright sunlight a higher contrast is needed because the user may else be unable to see what is on the display. Similarly, when the game is played in an area with a lot of background noise the user may not hear auditive feedback.

C.1.2 Provide unobstructed views that are appropriate for the users' current situation

Depending on the game context and what actions need to be taken certain views are better suited than others. E.g. for capturing game characters that are floating around the world an augmented reality approach may be better suited, while a map might be better suited for tracking down a hidden treasure.

Whatever view the user is using, the limited display size should be taken into account when the interface is designed. The interface elements should not unnecessarily block the view of the world.

C.1.3 Device UI and game UI are used for their own purposes

Dedicated device interface elements, like the home and back button, should not get different functionality in a game. The same way, game UI elements should not affect the device.

C.1.4 Provide users with information on game status

It should be clear to the user in what state the game is.

C.1.5 The player understands the terminology

In order to avoid confusion, the game should use terminology with which its players are familiar. This need not be terminology that is used in real life ("hitpoints", "mana", "juice from the Shrub Of Awesomeness"), but it does have to make sense within the game context.

C.1.6 Navigation is consistent, logical and minimalist

Because of the limited screen space that is available, the number and size of menu's should be kept to a minimum. When they are used though, their layout should be consistent and logicals as should their content be.

C.1.7 The game gives immediate and consistent feedback on the players actions

If the user performs an action, the game should immediately react to this by giving feedback about the users action even if the initiated game action is not completed instantly. When this happens the user knows that his input has been received and is being processed.

C.1.8 Provide intuitive input mappings that are easy to manage and have an appropriate level of sensitivity and responsiveness

If an ad hoc standard for an input scheme for this type of game has developed it is best to adhere to this. If this is not the case, it may prove beneficial to look at related games and genres and see how they handle user input. This way it is easier for the user to learn the input scheme. Furthermore, the level of sensitivity and responsiveness of the controls should feel natural when compared to the actions they are meant to symbolize.

C.1.9 The player cannot make irreversible errors

Making an irreversible error can hinder the player at a later stage of the game and should thus be avoided. Exceptions could be made for games that are over after one session, as this could be part of the game. E.g. making a move that leaves your king exposed in chess is undesirable, but part of the game.

C.1.10 The player does not have to memorize things unnecessarily

Because the working memory of people is limited, the game should not rely on people memorizing a lot of information. Especially with MMRGs this is important, as the highly dynamic environment in which they take place produces a lot of distractions.

C.1.11 The game contains instructions and help, so that the user does not need to look things up

MMRGs are meant to be played while outside and on the move. It is ok if the game needs a bit of explaining, but this information should be easily accessible from within the game. This can for example be done using tutorial levels or tooltips during the game. Users should not need to access an online or, worse still, paper manual.

C.1.12 The player can turn the game easily off and on, and save games in different states either by choice or by temporarily losing connectivity.

Characteristic for mobile games is that they are played in short sessions, it should thus be easy for the user to save the current game state and resume at a later time of their choice. Part of MMRGs is that they often rely on data or GPS connections, which may be interrupted for various reasons. This should at the very least not cause problems and allow the user to continue when the relevant connection is restored.

C.1.13 Real world navigation takes into account the type of game and is logical

Navigation through the real world should also be logical and take into account both the size of the game world and the type of game. In a game that focuses on running, it may be acceptable to send the player across the game world time and again but in a puzzle game this can hinder people that would otherwise be able to play it.

C.1.14 Display a short warning message about physical safety

If the game can put the player in dangerous situations, like crossing busy roads or running into objects, display a short and entertaining warning. The length should not exceed three lines on an average screen, because people will else just ignore it.

C.1.15 Take into account the characteristics of the environment for which the game is designed

If the game is meant to be played in the streets, design it so that the user does not need to look at a map constantly while navigating the real world at the same time at high speed. On the other hand, in an open space like a park this may not be problem.

C.1.16 Safeguard the players legal safety

If the area for which the game is designed has certain laws or regulations that can get the player into trouble, this should be taken into account to the extent that players are not forced to break them. Make sure for example that items do not become inaccessible because they are in places which are not freely accessible or that players have to get from one place to another in a time that's only possible when breaking the speedlimit.

C.2 Usability heuristics for multiplayer games

A game is considered multiplayer when players can interact with each other in some form via the game, but not when there is just a highscore list to compare results. Still, not every heuristic may apply to every multiplayer game. Some games are clearly multiplayer, but have no session management or matchmaking going on for example.

C.2.1 Simple session management

When a game is based on sessions, rather than a persistent world, users should be able to easily find or create game sessions that are appropriate to them.

C.2.2 Flexible matchmaking

A system should be available that allows the user to find other users that relevant for them. For MMRGs this can mean friends or players that are in the vicinity.

C.2.3 Appropriate communication tools

Especially in games where the players are required to cooperate but are not within earshot of each other, it is important to provide the possibility to communicate. How this is done depends on what is appropriate for the game, for some games this could be via spoken communication (e.g. calling or voice over IP) and for others it might be written (e.g. SMS or in game chat).

C.2.4 Support cooperation

When a game requires players to cooperate, this should also be supported via the interface. Relevant actions should both be available and easily accessible. What the relevant actions are, depends on the game.

C.2.5 Provide meaningful information about players

Relevant information should be provided about the players in a game. In some games it may be very useful to know where everyone is during the game or what their statistics are.

C.2.6 Identifiable avatars

Often games rely on the use of avatars to convey locational information and sometimes other information associated with players as well. For this to be a good method of conveying the information, it is important that the avatars that are used can be identified as belonging to a certain player.

C.2.7 Support social interactions

Especially in a persistent game world it is desirable for the players to be able to not just interact usefully, but also socially. This can increase their involvement in the game.

For non persistent games, social interaction can be supported by providing a platform on which the results can be shared and discussed.

C.2.8 Manage bad behaviour

If people show unwanted behaviour, like abusive language or cheating, it should be technically possible to sanction the offending player. Social ways of handling this are also a viable option, e.g. support users in spotting and reporting unwanted behavior.

Appendix D

3D model of a holder

Figure D.1: 3D model that can be inspected in greater detail

Appendix E

Results – Full lists of usability issues

E.1 Parallel Kingdoms

E.1.1 Heuristic Evaluation - Adapted

Issue	Heuristic	Exp 1	Exp 2	Exp 3	Avg
No music, it would make the game more exciting.	1.1	1	2	1	1
Next to a notification beep (which is used for multiple events), there isn't any sound in the game at all.	1.1	1	2	1	1
It is not possible to restart in the game.	1.12	1	0	3	1
The items in the inventory list are very close together, and hard to press when standing or walking.	1.6	0	1	3	1
Cannot log out to have multiple players on one device	2.1	2	1	1	1
Game opens with a screen, where you can create a new character, called 'About your character'. It is not shown why I need a character, or what I can do with it after I have created the character	0	0	1	2	1
Auditive feedback would improve the immersion and communicate relevant information	1.1	2	3	1	2
Password just needs to be entered once, but isn't displayed. On a mobile keyboard, it is very easy to make typing errors. These can't be identified as the input is masked (*****), presenting problems at the next login.	1.9	0	2	3	2
In the 'About your character'-screen: when you've rotated the device to enter your mail address, it isn't shown what needs to be entered in the next form field. The device first needs to be unrotated.	1.9	2	1	2	2
When the game is moving to the real world, the telephone shows you a location. However, it does not show a arrow or a locator.	1.13	2	2	2	2
It is not immediately evident that the "pages" counter is a button, it is only explained in the tutorial.	1.6	1	3	2	2
Sometimes the game reacted slowly, .e.g. when character walks	1.7	1	3	3	2
No GPS status indicator, even though GPS state influences what actions you can take	1.4	1	3	2	2
After dying, it isn't clear how long the user needs to wait before respawning.	1.4	2	3	2	2
Not very easy to find how to communicate.	2.3	2	2	2	2
You can easily be attacked while communicating, because your keyboard takes up most of your screen.	1.2	2	3	2	2
When receiving messages, you do not get information about the player that sent it.	2.5	2	2	2	2
The contrast of the game items and the background is also very low.	1.1	3	4	2	3
Can't zoom in or out which means it is difficult to get an overview of the area.	1.8	3	4	3	3
The instructions are sometime above the items you have to touch. Therefore you can't see the items	1.2	3	3	3	3
On the map with the satelite background, it isn't clear which items can be clicked.	1.1	3	4	2	3
Got an error about GPS location after creating a character. After fixing the problem, I need to enter all the character information again, because it wasn't saved.	1.10	3	3	3	3
Losing connections requires a restart of the application	1.12	4	3	3	3
Switching connection modes between mobile and wifi causes data sending errors which require the game to be restarted.	1.12	4	3	2	3

When the connection is lost, it is still possible to walk around. Only it isn't effective: the position is reset to the last synchronized position when the server connection is restored.	1.4	3	3	4	3
Game map isn't visible anymore after too many messages/pop-ups.	1.2	2	3	4	3
Error message about location during loading is placed in an unclear and unreadable location	1.2	3	2	3	3
It isn't clear who the player is attacking, when the player is surrounded by multiple enemies.	1.7	3	3	3	3
When you choose "try again" when moving to a new location, the map is not visible because you are in the menu.	1.2	2	3	4	3
Enemies can stand on top of each other, showing only one enemy to the user when there are actually multiple enemies.	1.2	4	2	3	3
Walking your dog costs feathers but you can not see how many feathers you actually have in the screen where you opt to travel	1.10	3	3	3	3
Error message "Unable to find your location" is displayed when the GPS is turned off. The error message does not aid in turning on GPS. Neither does it link to the phone's setting page where GPS can be turned on. It links to another page, which contains even more text. The alternatives presented could easily be checked by the program itself.	1.11	1	4	3	3
Avatars are not instantly identifiable, only after clicking does it become apparent who they represent	2.6	4	3	2	3
The cities have too many items and building to find your own buildings.	2.6	2	4	3	3
The skill points have explanations in a lighter colour and are not visible when the sun is shining on your screen.	1.1	4	3	4	4
If multiple enemies of the same type are on top of eachother, a list appears but it you can not distinguish which enemy is which in the list even if they have different characteristics.	1.2	4	3	4	4

Table E.1: Usability issues found in Parallel Kingdom by means of a heuristic evaluation using the adapted heuristics and their severity rating

E.1.2 Heuristic Evaluation - Pinelle

Issue	Heuristic	Exp 1	Exp 2	Exp 3	Avg
Characters all have the same avatar, making it difficult to recognize them	2.6	3	3	3	3
The tutorial speaks about a dotted circle while it isn't dotted	1.1	1	0	1	1
Error messages aren't always very meaningful (e.g. "A temporary error occurred"	1.5	4	3	4	4
Messages sometimes disappear too quickly	1.3	3	3	4	3
Error messages don't always fit in the area making it hard to read them.	1.2	4	2	1	2
Tapping on an error (to display more information) results in a fatal crash of the app.	1.7	4	4	3	4
Application asks for the wrong information: age is asked instead of date of birth. When asked for date of birth, people can't state that they are 9001 years old (too old for the game).	1.9	1	1	1	1
It's obligatory to give in an e-mail	0	2	3	1	2
If something goes wrong during the signing up you have to do the process all over again	1.10	3	3	2	3
The system placed me in France	1.13	3	2	2	2
The icon of the monk is not clear	1.1	3	2	2	2
The goal of the game is not clear	1.11	4	3	4	4
The tutorial doesn't tell how to gain skills	1.11	3	3	3	3
There's always a chat going on in the top of your screen	2.3	1	1	2	1
Notifications take up a large part of your screen	1.2	1	2	3	2
You cannot zoom in or out on the map	1.8	4	1	2	2
It's not made clear when you can build something	1.4	3	2	3	3
You cannot see how strong other characters are before you attack them	1.2	4	3	3	3
The messages only are usefull when you do what they state. They don't give feedback on your own actions	1.7	3	3	3	3
What does the icon in the right-uppercorner do	1.1	3	2	3	3
Errors are displayed for a short amount of time. Especially when the error contains important information like an e-mail adress it is frustrating when the message disappears before fully read the text.	1.4	3	3	4	3
Only one player can create an account on a mobile phone	2.1	3	1	1	2
There is a way of communicating with others, however you don't know where these persons are and it is unlikely to get an answer on a question	2.5	3	2	1	2
It was unclear where I was able to walk	1.1	4	3	3	3

After about half an hour of playing the character didn't do anything anymore (no fighting, walking the dog, picking berries or buying stuff)	1.6	4	4	3	4
Two icons on top of the screen are not visible.	1.2	2	3	4	3
The game didn't show the current internetconnection, it was not clear when the game had no connection anymore.	1.3	4	2	3	3
When you are looking around, it's hard to find your character back	1.6	3	4	3	3
The chat was not meaningful to me	2.3	3	2	1	2

Table E.2: Usability issues found in Parallel Kingdom by means of a heuristic evaluation using the heuristics by Pinelle et al. (2008a) and Pinelle et al. (2009) and their severity rating

E.1.3 Diary study

Issues	Heuristic	Exp 1	Exp 2	Exp 3	Avg
Long loading times	0	2	3	2	2
Concepts were not understood unclear language	1.5	4	3	2	3
Participants expected the possibility to zoom	1.8	2	2	1	2
The view is sometimes obstructed by messages or hints	1.2	4	3	3	3
Certain messages are unclear and annoying	1.5	3	2	3	3
Information overload at the beginning	1.6	4	3	4	4
Not everything is explained in the tutorial	1.11	3	2	2	2
The meaning of information about armor and weapons (+X/y%) are unclear	1.5	3	2	1	2
No clear difference between read and unread messages	1.1	2	4	1	2
Descriptions in the create menu can be clearer, not always clear what the different items are that you need to build something	1.5	1	4	2	2
Playing the game while moving is not a good idea, as it requires too much attention	1.13	1	1	2	1
The game does not handle losing connections very well	1.12	4	3	2	3
There are no (clear) goals	1.11	0	3	1	1
The game crashed without any message	1.4	4	4	2	3
The partly transparent colour overlay to show to whom an area belongs was unclear	1.1	2	2	2	2
One participant lost her avatar and couldnt find it again for some time	1.6	2	3	3	3
Lightning symbol in the top corner was unclear	1.1	2	1	1	1
The user is presented with options he is not allowed to perform	1.6	4	2	2	3
Touching the right interface element is hard	1.8	3	3	3	3
Too few items on the screen, which makes it hard to locate the one you are looking for	1.2	2	3	2	2
Selecting a different language, in the item info screen, is very hard because the buttons are place too close together	1.2	2	4	3	3
Slow reaction times without intermediate feedback	1.7	3	4	4	4
Some messages had to be marked as read manually	1.6	3	3	1	2
Not clear that you can click on a badge	1.6	3	3	2	3
After clicking on a badge there is no feedback and after a while something new appears	1.7	3	3	3	3
Not clear what a hat is	1.11	3	1	1	2
Changes in the preference screen present the user with a blank screen while loading	1.7	1	3	3	2
Search function for players was not clear	1.1	1	2	2	2
The meaning of travel in the menu is unclear	1.5	3	3	2	3
Preview does not provide a good enough preview for the participant to actually determine if it is correct	1.2	3	2	1	2
Some information may become inaccessible after clicking it away	1.10	4	3	4	4
Logging out by pressing the home button on the device is unexpected behavior	1.12	3	3	4	3
When registering the password only has to be entered once, allowing for easy mistakes	1.9	1	4	3	3
Buttons at the bottom of the screen are too small	1.6	2	2	4	3
The function of food was unclear	1.5	1	1	2	1
There is no way to snap back to your character when you have scrolled around	1.6	3	4	2	3
You cant revisit random flags youve built in no-mansland	1.6	0	3	1	1
It is not clear for one participant where he could buy food for gold, rather than real world money	1.11	1	1	1	1
The meaning of Request Passage is unclear	1.5	2	2	1	2
Not clear that you could only travel after finishing the tutorial	1.4	2	1	3	2
One participant was told that a dungeon was nearby, but was given no further information on how to find it	1.6	1	1	2	1

Table E.3: Usability issues identified in Parallel Kingdom by means of a diary study

E.2 Portal Hunt

E.2.1 Heuristic Evaluation - Adapted

Issue	Heuristic	Exp 1	Exp 2	Exp 3	Avg
When loading a file after inputting user data, user data is removed	1.10	2	2	3	2
Spellingmistakes in the help menu	0	4	1	2	2
Devices get rather hot	0	3	1	1	2
No option to store login data	1.10	2	3	1	2
Statusbar of the phone disappears, making it impossible to determine connection status	1.3	2	2	3	2
Settings and help button do not function in camera mode	1.6	3	2	2	2
Grey background is not a clear indication of the loading status	1.7	2	1	4	2
options wheel does not give options, but back and exit	1.8	3	1	2	2
Software does not compensate for measurement inaccuracy in gyroscope (enough)	1.1	1	2	3	2
Can not change server	2.1	2	3	2	2
once chosen, you can not change teams	2.2	2	3	2	2
Display a short message about physical safety	1.14	2	2	3	2
When exit is clicked, password/username is not stored	1.10	1	2	2	2
Back button does not do what was expected	1.6	3	2	1	2
Screen switches off	1.3	2	1	4	2
No feedback when you catch a portal	1.7	3	2	2	2
Unclear how many points you get for a portal	1.11	4	1	2	2
No avatars to identify other players	2.6/2.5	2	2	1	2
Not possible to manage bad behavior	2.8	3	1	1	2
Green/yellow portals differ too little in contrast to be distinguishable by colour deficient people	1.1	3	4	2	3
Ambiguous text in the help menu	1.5	3	3	2	3
You have to load the file every time again	1.12	4	3	3	3
It does not remember what the last loaded file is	1.12	3	2	4	3
When the phone is rotated the view is rotated, but not full screen	1.2	3	3	2	3
Portals do not rotate when the view is rotated	1.2	3	4	3	3
Two login buttons without an explanation	1.1	4	3	3	3
Error when logging into game without an activated account does not give appropriate information	1.7/1.11	4	3	3	3
options wheel is too small to easily tap	1.8	3	2	3	3
pinch to zoom does not function	1.8	3	2	3	3
You are unable to communicate with other players	2.3/2.4	3	3	2	3
A German keyboard is presented even though the phone is set to English	1.8	4	4	1	3
Map breaks after zooming in or out	1.2	3	4	2	3
Add text cue to sphere to make it clear on the map how they can be caught	1.1	2	3	3	3
Help button gives feedback but does not actually work	1.11	4	2	3	3
Hard to navigate because you can't see which way you're facing	1.13	3	2	4	3
Enormous battery drain (30 minutes playtime on full battery; galaxy s)	-	2	4	2	3
No cooperation support	2.4	3	2	3	3
Unclear which input field is selected	1.1	4	4	4	4
Text in leader tab is unreadable	1.1	4	4	4	4
It appears GPS places you in the middle of the playing field, even you are not actually there.	1.2	4	4	4	4
Keyboard does not extend/retract like normal	1,8	4	4	4	4
You need a certain filemanager, but this is not explained	1.11	4	4	3	4
camera is not always accessible, and when not no error is given just a black or white screen	1.4	4	3	4	4
No error when the filemanager is not available	1.7	4	4	3	4
Mission select text does not always appear in mission select	1.1	4	4	4	4
The phone completely crashed without an explanation	1.9	4	4	4	4
Connection seems to randomly fail without an explanation	1.12	4	3	4	4
Portal Hunt but not Osmo crashed without any notice, resulting in just a black screen	1.4	4	4	4	4
The game/phone froze for no apparent reason	1.4	4	4	4	4
It's not clear how to catch a portal	1.8	4	3	4	4

Table E.4: Usability issues found in Portal Hunt by means of a heuristic evaluation using the adapted heuristics and their severity rating

E.2.2 Heuristic Evaluation - Pinelle

Issue	Adapted	Exp 1	Exp 2	Exp 3	Avg
”Select mission” did not always appear	1.1	1	1	3	2
No GPS status indicator (it was switched off and no hint at why it didn’t work)	1.4	2	2	3	2
Part of login data got deleted after exiting from cogwheel menu	1.10	2	2	1	2
Sometimes the camera blinks	1.2	2	1	2	2
The map was not adapting to my location	1.2	4	4	4	4
It takes some time before a button reacts (e.g. keyboard appeared only after demanding it 5 times)	1.8	3	3	4	3
There was no help button	1.11	2	4	3	3
During the game there was no help	1.11	1	2	2	2
When starting the game it is not clear what to do, and even if the game is working at all	1.11	4	3	4	4
It is not clear which file to open when you have to select a file	1.11	3	4	3	3
It is not possible to hit portals while you are in camera view	1.8	3	2	3	3
There is no communication tool to communicate with people of your own team	2.3	2	1	2	2
The map didn’t show the locations of the competitors/teammates	2.5/2.6	2	3	3	3
The portals moved away before the game responded to the catch I tried to make	1.7	1	3	3	2
There is no game status in the camera/map view which makes it hard to determine a strategy while playing the game	1.4	1	2	2	2
It was not clear what influenced how many points you got per portal	1.11	3	2	3	3
The game crashed and after that it wasn’t possible to get screens of the camera or the map	1.4	4	4	4	4
Keyboard didn’t appear when pressing in the login input fields.	1.8	4	4	2	3
I didn’t get a notification when i scored points	1.7	2	3	3	3
Screen blacked out (for battery protection) during the game	1.3	3	2	4	3
The system did not guide you when choosing a game type or group to play with	1.11	3	2	4	3
The map wouldn’t load at the right location	1.2	2	3	4	3
If you pressed the ’settings’ button you could only get back to the menu or exit the game, but you couldn’t go back to the application.	1.6	2	2	4	3
The portals weren’t on the exact same location on the map and the AR	1.13	3	4	4	4
Portals could block the view of the world which could eventually make you bump into things	1.15	1	3	3	2
The system wouldn’t adapt (to a wide screen) if you turned your phone horizontally	1.2	3	4	3	3
The options for teams were not clear to me	2.1	3	2	4	3
There was not much interaction with the other players during the game	2.4	2	2	2	2
It is not clear when the game ends?	1.11	3	3	2	3
Flexible matchmaking: there is no way to determine how experienced players in a game are until you join it.	2.2	2	1	2	2
Provide instructions, training, and help: there is a little bit of help, but it doesn’t explain how to catch a portal. (i.e. it says you can catch it, but not how)	1.11	3	3	4	3
Information on game status: the player’s status is provided in a separate ”tab”, but this may lead to micro management.	1.4			2	2
There is no training level for new players in which the basics are explained.	1.11	3	2	2	2
Input mappings aren’t intuitive (at least in the login screens)	1.8	3	4	3	3
There is no support for coordination amongst team members; the only way to accomplish this is by shouting	2.4			2	2
The clicking on the screen (in order to catch a portal) was not very accurate	1.8	3	3	3	3
The language in the help section was incorrect (”catched” instead of ”caught”)	0	0	1	2	1
The portals could appear within solid objects on the playfield.	1.15	2	1	3	2
It is unclear which server has to be chosen when logging in.	1.1	3	3	3	3

Table E.5: Usability issues found in Portal Hunt by means of the heuristics by Pinelle et al. (2008a) and Pinelle et al. (2009) and their severity rating

E.2.3 Retrospective Think Aloud/Instant Data Analysis evaluation

Issue	Severity	Heuristic

One participant assumed the keyboard enter key would result in using the correct login without an obvious reason for the difference	0	0
There is a grammar error in the help text	1	0
One participant tried rotating the map, but couldnt.	1	1.2
Zooming in or out did not alter the size of the portals on the map, leading one participant to believe that if he zoomed out enough he would be able to catch them all without moving.	1	1.1
Some participants were unable to close the keyboard	1	1.8
Camera was not always understood as meaning Augmented Reality view	1	1.5
The login button does not always respond	2	1.8
The list of missions did not elicit input from every participant	2	1.8
The cogwheel button is associated with options, not back or exit	2	1.8
The portals flicker heavily in the AR view	2	1.1
Pinch-to-zoom is expected, but not available	2	1.8
The differently colored portals are understood as giving different amounts of points or being of different difficulty, not as requiring different tactics to be caught	2	1.1
One participant reported that the red and yellow descriptions are similar, but was amazed that he could catch the red one when he was alone in a team	2	1.5
Not clear that different spheres could give different amounts of points	2	1.11
Not clear how close you have to be to the portal to catch it, i.e. in it or touching it	2	1.13
Some participants took a while to figure out they were the crosshair	2	2.6
Back in cogwheel was interpreted as close this menu by one participant, not as go back to the mission menu	2	1.6
Multiple participants reported that they thought there were many more portals in AR view than in the map view	2	1.4
Several participants did not catch on to the fact that certain portals were moving/jumping	2	1.4
Keyboard does not automatically appear when clicking an input field	3	1.8
There are two login buttons without any clue as to which is the correct one	3	1.1
Loading can take very long without giving the user feedback on if stuff is actually still happening	3	1.4
It is not clear that each team has only access to one session	3	2.1
GPS updates only very slowly, making it hard for the player to determine their position	3	1.13
The meaning of the grey area around the tile on which the user is, is unclear. Some participants interpreted it as that it was not part of the game.	3	1.1
There is no feedback when a player actually catches a sphere	3	1.7
All participants reported problems with depth perception	3	1.1
One participant almost walked into a tree	3	1.15
On occasion loading took so long that participants restarted the game, and then were able to quickly log in	3	1.4
The participants that used help, did not find everything they needed in it	3	1.11
On starting Osmo4, the user is presented with a black screen without any instructions	4	1.11
It is unclear where you are in relation to the playing field	4	1.1/1.13
The help button does not always work	4	1.7
It is not clear how to catch the portals. Participants expected walking through them.	4	1.8
The game froze when pressing certain buttons (zoom, help, cogwheel)	4	1.4
The cogwheel button does not always respond	4	1.7
When in the Choose mission screen, load file doesnt work	4	1.7
The game crashed without any notice	4	1.4
Close button in help did not always respond	4	1.8

Table E.6: Usability issues found in Portal Hunt by means of retrospective think aloud and instant data analyss and their severity rating

E.2.4 Audio diary & Interaction Logs evaluation

Issue	Heuristic	Exp 1	Exp 2	Exp 3	Avg
The game is called Portal Hunt but the score sheet tells you that youre catching spheres.	1.5	1	1	1	1
Choose mission list made one participant first try tapping the grey area	1.1	2	0	1	1
Grammatical error in the help text	0	1	1	1	1
Screen rotates but not full screen	1.2	3	2	1	2
Participants expect the keyboard to disappear on its own	1.8	4	2	1	2
The unable to connect message is interpreted as potentially having entered the wrong password.	1.5	1	3	3	2
Close in help doesnt always respond	1.8	4	1	2	2
Participant expected to be able to close the cogwheel menu tapping outside it, but couldnt	1.6	4	1	2	2
The map is perceived as more trustworthy, which also is strange to participants as they are used to trusting their eyes rather than a map.	1.1	3	2	2	2

Not directly clear for every participant if the spheres on the map represent the players or the portals	2.6	4	2	1	2
To some participants it was not evident what the camera tab is for at first	1.5	3	2	2	2
One participant actually walked into a tree	1.15	4	2	1	2
One participant expected to have to interact with loading screen, due to long loading times	1.4	3	2	2	2
One device showed only one login button	1.6	4	2	1	2
Keyboard does not pop up	1.8	4	4	1	3
Keyboard does not always respond on login screen, most notably the backspace	1.8	4	3	3	3
When the connection to the server fails, you get a message that says Unable to connect to server!. This however does not make clear if that is because the server is broken or you have no connection	1.5	3	3	3	3
Loading takes a long time, without it being clear for the participants if its doing something	1.4	3	2	3	3
Help doesnt always respond	1.7	4	2	2	3
Cogwheel doesnt always work	1.7	4	2	2	3
Text in leader screen isnt completely readable	1.2	3	2	3	3
The game froze/tabs didnt respond for no apparent reason	1.4	4	4	2	3
Not clear that surrounding a sphere means with people from the same team	1.5	4	3	2	3
Exit in cogwheel does not always respond	1.8	4	2	3	3
Participant tried tapping portals they were quite some way away from them	1.13	2	4	4	3
Player location updates very slowly, sometimes causing the player to run past a portal	1.13	3	4	3	3
It is very difficult to gauge the depth, due to bad positioning	1.1	3	3	3	3
From a certain distance, all portals are the same size	1.1	4	2	3	3
No feedback on success or failure when capturing portals	1.7	3	4	3	3
There seems to be a difference in the distance the portals are removed from you between the camera and the map.	1.1	4	2	2	3
File browser does not always respond	1.7	4	3	2	3
Last file randomly does or does not load	1.12	4	2	2	3
Participants did not easily identify the cross as their avatar	2.6	4	2	3	3
The cross changes color, this was not noticed by participants. Hence, no meaning was derived	1.1	4	3	2	3
Participants interpreted strongly flickering portals as moving, rather than as GPS inaccuracy	1.1	3	3	3	3
One participant had trouble navigating, because they couldn't determine their orientation	1.13	3	4	3	3
Keyboard sometimes covers input fields for the login screen	1.2	4	3	2	3
When login screen is (partially) covered, it is not possible to scroll	1.8	4	3	1	3
Starting osmo sometimes freezes the whole phone	1.12	4	4	2	3
Clicking a mission in the mission select menu did not always work	1.8	4	3	2	3
Camera did not work on one phone	1.2	4	2	3	3
The map randomly disappeared for one participant and later reappeared	1.1	4	3	2	3
At one point, one participant could neither use the map nor the camera to play the game	1.1	4	3	3	3
Not clear that teams are bound to locations	2.1	4	4	2	3
Choose mission text appears just like the black bar in which the mission text and description should be, but there is no text in it	1.2	4	2	2	3
Tapping the empty bar at first does nothing, but after several taps and waiting several seconds the game does start	1.8	4	4	2	3
Participants expect that they can enter their own names and/or register via the game, but can't	1.6	4	4	3	4
Not clear how to catch portals	1.8	4	4	4	4
"Portal" suggests walking through it, but you actually need to tap it in map view	1.5	4	4	4	4
Items outside the map tile are interpreted as not being part of the game	1.1	4	4	3	4
On starting Osmo4 players are, sometimes, presented with a black screen without any further instructions or feedback.	1.11	4	4	3	4
Not clear which server to use for logging in	1.1	4	3	4	4
Participants restarted the game due to long loading times	1.4	4	4	3	4
Choose mission text did not appear on every device	1.1	4	4	3	4
The game randomly crashed without any feedback as to what happened	1.4	4	4	3	4
Help does not actually explain how to catch a portal	1.11	4	4	4	4

Table E.7: Usability issues found in Portal Hunt by means of analyzing the audio diary and interaction logs and their severity rating

E.3 Tidy City

E.3.1 Heuristic evaluation - Adapted

Issue	Heuristic	Exp 1	Exp 2	Exp 3	Avg
When you make an account you do not login with the account automatically	1.10	2	1	1	1
Your navigation sense is hindered by omitting a path	1.13	1	2	0	1
Display a short message about physical safety	1.14	2	1	1	1
Allow forbidden zones inside the zone	1.15	1	2	1	1
The browse missions button can also be interpreted as a loading bar	1.1	1	1	4	2
Text of the mission icons runs over other mission icons	1.2	2	2	3	2
When I download a mission the player is returned to the selection screen, not to the game itself.	1.6	1	2	2	2
The map does not rotate to match my view	1.8	2	2	1	2
It is not clear when you are close enough to pick up a riddle	1.13	1	3	3	2
The meaning of the circle around the dot representing you is not clear.	1.1	2	1	4	2
The hint text about what to do is too long	1.11	1	3	3	2
When the showroom is empty, it is not necessarily clear what it is for	1.11	1	2	3	2
Loading of the map (satellite) takes long. The button to switch to map view is not an evident solution.	1.1	2	2	3	2
The map could/should be cached alongside the riddles when they are downloaded	1.12	2	2	2	2
An overview of the pictures would be clearer in the inventory	1.10	1	3	3	2
Items are not automatically picked up when you open them	1.6	0	2	3	2
The auditive feedback when solving a riddle is very minimalistic, compared to the other cases	1.1	2	2	2	2
The cursors hides behind riddles	1.2	3	3	1	2
The compass hides behind riddles	1.2	3	1	1	2
Allow the level designer to give a popup upon starting the mission	1.11	2	2	1	2
Make it more explicit (red lines) that outside the yellow line is off limits for the game.	1.15/1.16	1	3	2	2
You do not get feedback after creating an account	1.7	3	2	3	3
The maps does not automatically fly to where you are now	1.2	3	4	3	3
Not clear what the solve button does	1.5	1	3	4	3
After you lock your screen, the missions disappear from the mission overview screen but return after pressing the back button	1.2	4	2	3	3
If loading is not instantaneous, it is not clear the game is loading. The screen looks a lot like it is hanging/frozen.	1.4	3	3	4	3
Goal of the game is not immediately clear	1.11	1	4	4	3
The map does not follow you when you move	1.2	3	4	3	3
The expected action of the backbutton in the inventory is to go back to the map, not to the previous app.	1.8	3	2	3	3
Auditive feedback does not work for the hearing impaired or in noisy environments (maybe improve with vibrate feedback)	1.1	3	3	2	3
When creating account with a name that already exists, you get an error message "Errors: you have the following errors" but no actual errors	1.7/bug	4	4	3	4

Table E.8: Usability issues found in Tidy City by means of a heuristic evaluation using the adapted heuristics and their severity rating

E.3.2 Heuristic Evaluation - Pinelle

Issue	Heuristic	Exp 1	Exp 2	Exp 3	Avg
After registering, you're not automatically logged in	1.10	1	2	2	2
GPS is not always 100% accurate	0	4	3	4	4
Checkpoints can be not accessible due to constructions or weather conditions	1.15	2	1	1	1
It's not clear how many points you get for a question	1.11	3	2	3	3
If you can't find a checkpoint, you can't finish the game	1.15	3	3	2	3
It is not possible to leave the e-mail entry empty	0	2	1	1	1
If you browse through the inventory and browse to the last assignment, you cannot browse further to the first assignment	1.6	2	2	1	2
The browse button looks like a loading bar	1.1	3	3	2	3
The icons for places aren't that clear	1.1	2	2	3	2
It is not possible to play in teams instead of against each other	2.4	2	1	2	2

You can't see what items your competitors have or haven't found	2.5	1	1	3	2
In the beginning there is no explanation what to do	1.11	3	3	3	3
It is not possible to customize audio settings for the game. This can be helpful for competitors	1.3	3	2	1	2
You can't get extra information on a riddle	1.11	3	1	4	3
The 'car' icon is not intuitive for map view	1.1	3	2	2	2
The names of the buildings that are shown on the map are very small and don't get larger when using the zoom function	1.1	3	1	1	2
The information about score is in another tab as the map	1.10	3	1	2	2
It can become difficult to remember which title belongs to which question/location	1.10	2	2	3	2
The "inventory" tab is initially empty, and displays no hint of what it will be used for.	1.11	3	3	3	3
The register button is at the bottom of the login form, which makes it look like you can register by filling in that form.	1.1	3	2	2	2
Questions are in a popup while they take up all of the screen estate.	1.2	1	1	3	2

Table E.9: Usability issues found in Tidy City by means of a heuristic evaluation using the heuristics by Pinelle et al. (2008a) and Pinelle et al. (2009) and their severity rating

E.3.3 Think Aloud evaluation/Instant Data Analysis

Issue	Severity	Heuristic
The icons for changing between map and satellite view are not intuitive	1	1.1
Meaning of the blue circle is not always interpreted correctly. (One participant assumed it meant they were in the right spot)	2	1.1
Participants referred to the riddles by their icons Ill pick up the tree or Lets go to that sunscreen/umbrella, but they did not report this as having any meaning.	1	1.1
One participant wanted to go and look for a solution outside the boundary, i.e. they didnt interpret the yellow line as a boundary	1	1.15/1.16
One participant saw the browse missions button as a loading bar	1	1.1
Slow loading map means you can not always navigate very well.	3	1.10
Inventory shows only names, which means you have to remember the pictures that belong to it.	2	1.10
Generally participants found out what to do by trial and error, rather than by reading the help file.	2	1.11
Scoring system is unclear, i.e. what determines how many points you get	2	1.11
The images were used as general concepts, not specific spots. (e.g. ash tray meant any smoking space)	1	1.11
One participant wondered if there was a limit to the size of your inventory		1.11
Distance between riddle and solution are unclear	1	1.13
Seeing position was hard at times, as the blue spot disappeared under the riddle icons	3	1.2
Titles of missions can overlap, making them unreadable	1	1.2
Messages on pickup and solving correctly are gone quite quickly	1	1.4
Unclear how close one needs to be to pick up a riddle	1	1.13
Unclear how close one needs to be to solve a riddle	2	1.4
Some participants expected the riddles to pop up automatically when you are near their pick up point (not drop off point).	1	1.4
Slow loading map prompted one participant to restart the mission, as they thought it was broken rather than loading		1.4
Difficulty: a number doesnt say that much, because you do not know the limits. Also if you do know the limits, you have to know the way they are sorted.	2	1.5
Difficulty: not clear if there is any relation between difficulty and the distance between riddle position and solution position.	2	1.5
The meaning of showroom was not clear to most participants, saying stuff like Oh showroom, its empty no idea what its for	2	1.5
Solve is often not understood as Im at the right place, but more like I want to try and solve this now or I give up, show me the answer or Please give me a hint.	3	1.5
Unclear what kiosk mode means in settings	1	1.5
Meaning of cancel is unclear when having pressed solved in the wrong place, does this mean that I can avoid losing points?	1	1.5
Text was not really considered part of the game. Focus was solely on image.	2	1.5
Not clear what download means when you select a new mission		1.5
Several people had trouble determining in which direction they were going, because they did not notice the compass.	3	1.6
When you register the name in the login screen is not updated	2	1.6
When you enter a wrong password, the username disappears	2	1.6
Having a login button under the menu is weird, because you are already logged in	1	1.6
One user said they did not expect to go back to play or pick mission screen after actually selecting a mission.	1	1.6

One participant tried to zoom in on the image (/would have liked bigger pictures)	1	1.6
Unclear how to change your mission	1	1.6
No confirmation if the registration was successful	1	1.7
Missions disappear from/do not appear on the map	2	1.2
There were missions with riddles that had no text or just the standard from the editor notes, which should not be possible when set to playable	0	
Participants wondered if there was time limit	0	
Meaning of icons at the bottom is not always clear	0	
Missions are only available in one language, which makes them inaccessible to people who do not understand it	0	1.11
Participants would have liked hints as to how far away they are when they were in trouble	0	1.11
A deleted game was still visible, but inaccessible from the phone. (not reproducible)	0	
Accidentally returned to pick or play screen	0	

Table E.10: Usability issues found in Tidy City by means of Think Aloud/Instant data analysis and their severity rating