In Proceedings of the 7th International Working Conference on Human-Centered Software Engineering.
Sophia Antipolis, France: Springer, 2018.

User Evaluations of Virtually Experiencing Mount Everest

Marta Larusdottir[1], David Thue and Hannes Högni Vilhjálmsson[1]

[1] Reykjavik University, Menntavegur 1, 101 Reykjavik, Iceland
{marta; davidthue; hannes}@ru.is

Abstract. In software development it is hard to know both whether the team has developed a product that fits the users' needs, and is easy to use. One way of gathering feedback from users on both these issues is to conduct formal user testing, which has been rated by IT professionals as one of the best methods for user involvement in software development. In this paper, we present a formal evaluation of a running prototype for a virtual reality experience that was scheduled to be launched 3 months later. We conducted formal user testing with five users, and recorded the problems that the users experienced while they used the VR prototype. We also collected data concerning each user's impressions of their experience immediately after it was complete. The results show that many serious problems were identified, and that the developers found several of them to be very useful. In some cases, the user testing was regarded as having been essential to discovering these problems.

Keywords: User testing, virtual reality, agile software development.

1. Introduction

In software development it is hard to know both whether the software development team has developed a product that fits the users' needs and is easy to use for the users. A recent study shows [22] that some developers who deliver software to users only obtain a vague idea of the usage of their system, mainly because they don't contact the users, and the users do not contact the developers. The users simply find ways to bypass any problems that they have while using the product, even though it delays their work or makes them frustrated. Both formal and informal methods have been defined for gathering feedback on the user experience from users during software development.

Agile software development has been the de facto standard for project management in software development for some time. Informal methods, such as short interviews or showing low-fi prototypes to users and discussing those, are used quite extensively in agile software development [19]. Still, formal user testing, with users solving predefined tasks while being observed, was rated as the best method for involving users in Agile projects [14]. The results from the same study showed that such testing is performed quite rarely, due to lack of time and money.

In this paper, we study a formal evaluation of a running prototype for *Everest VR* [26], a virtual reality experience which was scheduled to be launched three months later. We conducted formal user testing with five users, as suggested in the Google Design sprint process [17] and by other researchers and practitioners, for example Jakob Nielsen [24]. We recorded the problems that occurred while users participated in the VR experience and we collected data about the users' impressions of the experience immediately thereafter. We focused especially on how useful the results from the formal user testing were for further development of the product. This is rarely done in the literature. Specifically, we sought to answer the following research questions:

1. How many severe problems are found during formal user testing?
2. How useful are the identified problems for the further development of the system?

The contributions of this paper are twofold: We explain how VR software can be evaluated with 5 users in formal user testing by describing the process and the data collected. But perhaps the main contribution is that we collected data on the benefits of the results of the user testing from the actual developers and describe those results in the paper.

The remainder of this paper is organized as follows. We begin by describing some of the current literature that relates to our research questions. We then present the data gathering methods that we used in the study along with the results from the study itself. Finally, we discuss the results.

2. Background

In this section, we describe some of the current literature on designing virtual reality for users and on user evaluations.

2.1 Virtual Reality

Virtual Reality devices such as the HTC Vive (www.vive.com) allow users to observe and interact with a simulated environment as though they are physically situated within that environment. Specifically, by precisely tracking the 3D position and orientation of a display device mounted on the user's head, the user's perspective of the virtual world can be controlled using the muscles in their body, identically to how they control their perspective of the real world. For example, to obtain a better view of a nearby object on the ground of a virtual world, a user could physically move their body into a crouch and thereby move their virtual perspective closer to the object. Some VR devices (including the HTC Vive) also allow the precise tracking of hand-held input devices, which are often used to represent the user's hands inside the virtual world. These devices allow users to physically move their hands to interact with the virtual world, including manipulating virtual objects and performing gestures (e.g., pointing or waving).

An important limitation of the HTC Vive is that its ability to track the headset and hand controllers is limited to a predefined tracking volume, which has a recommended maximum base area of 3.5m x 3.5m; this volume limits the extent to which the user can use *only* their natural body movements to explore a virtual world. To overcome this limitation, many instances of VR software also implement a way for the user to traverse the world at scales larger than 3m x 3m (e.g., using a hand controller to point at and teleport to a target location).

Given the speed with which the latest generation of VR technology has been developed, it has been difficult for the designers of VR software to form and maintain a good intuition for how users will use this technology to interact with virtual worlds. This magnifies the importance of running frequent user tests during the development of VR software, to both account for the current lack of intuition and to start building stronger intuition for future projects.

2.2 User-Centered Evaluations

The goal of a user-centered evaluation activity is to gather feedback on the IT professional's work from the user's perspective [11]. The type of information gathered in user-centered evaluation has been evolving through the years. About 20 years ago, the major emphasis was on gathering information on usability problems, which are flaws in the interface that cause problems for users [23]. Parallel to this, the emphasis was also on measuring usability in a quantitative way by measuring effectiveness, efficiency, and satisfaction, as defined by the ISO 9241-11 standard [12]. During the last decade, the study of user experience has gained more attention, where more subjective factors regarding the users' perspectives using the software are measured [9]. Hence, user-centered evaluation can be used to gather information on usability problems, the three factors of usability (effectiveness, efficiency, and satisfaction), and the subjective factors of user experience. When evaluating virtual reality software, one of the factors that are of interest is the effect on the body of the user. A simulator sickness questionnaire was proposed by Kennedy et al. [16] to measure the various effects that using VR software can have on the user's body.

The evaluation activity needs to be planned, conducted, and the results need to be analyzed and reported [11]. Evaluations can be conducted to gather feedback regarding the context in which the software will be used, the requirements from the user's perspective, and on the user interface design. The feedback gathered in an evaluation identifies possible flaws or problems users have while using the software. The feedback can also include the experiences users have. These are described and IT professionals must decide what action to take in each particular case to improve the usability and the user experience of the software.

Some studies have been conducted on the benefits and drawbacks of conducting user centered evaluations. The results from a survey and interview study conducted in Denmark show that some type of user centred evaluation was conducted in almost 75% of the companies involved [2]. The study did not analyse what evaluation methods were used, and whether the evaluation included users or not. A similar study was done in Italy [1] and the results on the usage were similar; some evaluation was done in 72% of the companies involved. Internal evaluations were conducted in half of the companies involved in the study, but less than 20% conducted external evaluations by external consultants.

The major obstacles for doing user centred evaluation were examined in a study conducted in Denmark [2]. The two main obstacles found were resource demands, both in terms of time and money, and an obstacle called the “developer mind-set”. One example of an issue in this category, mentioned by respondents, is that developers find it hard to think like users. In relation to this, some informants described that the main focus of IT professionals was on the programming aspect - to write beautiful code - and not so much on participating in a usability evaluation. In a similar study in Italy [1] the most frequent obstacle mentioned was also resource demands. The most frequently mentioned advantage of usability evaluation was quality improvement, reported in almost half of the cases. In the study by Vredenburg et al. [28] the main benefit for doing usability evaluation was that the practitioners gain understanding in the context of use whereas the weaknesses mentioned were high cost and versatility. The benefits and weaknesses of doing informal expert reviews and formal heuristic evaluation were similar, the benefits being the low cost and speed, but the weakness being that users are not involved in the evaluation. In these three studies, the major evaluation obstable is the client’s budget resource.

Twenty years ago, it was common to study the benefits of conducting user-centered evaluations by counting how many usability problems were found by using a number of evaluation methods to evaluate the same software. Four studies comparing evaluation methods were published in the years 1990 to 1993 [13; 15; 7; 25]. In addition, a study by Cuomo and Bowen [6] is also discussed there. In these studies an aggregated list of all problems found during user observation is made and used to describe all usability problems that can be found in the software. Then the number of problems found by using another method is compared to the aggregated list. It is common to presume that problems found while observing users in user centred evaluation are true problems that users would have in real use. Problems found by using another method are compared with the list of problems found in the user observation to calculate the effectiveness of the method. These studies were focusing on that outcome of the evaluations, which is finding usability problems.

Other studies have covered how these outcomes can be described in an efficient way to help the IT professionals to decide what to do about the problems [27]. The most efficient way of describing the results of the evaluation to the IT professionals was, for example, studied by Hornbæk and Frøkjær [10]. Their results show that IT professionals assessed redesign proposals to have higher utility than usability problem lists.

One way of determining the effect of using usability evaluation methods is to look at the downstream utility, which is defined by Law [21] as: “*The extent the improved or deteriorated usability of a system can directly be attributed to fixes that are induced by the results of usability evaluations performed on the system.”* Here the effect of the usability evaluation is determined by how much it improves the actual usability of the software and not by how many problems are found. There are very few studies that report the downstream utility of using a particular user-centred evaluation method. Researchers do not agree on the scope of user centred evaluation. Cockton [5] argues that assessing the downstream utility is beyond the scope of pure evaluation methods.

In this paper we assess the downstream utility of conducting user evaluation with 5 users in a VR experience and report the findings.

3. Method

In this section we will explain the VR experience whose prototype we evaluated (*Everest VR*), how we gathered the data, the demographics of our participants, and the methods that we used to analyse the results.

3.1 Everest VR

Everest VR [14] is an interactive experience in Virtual Reality that simulates some parts of the (real-world) experience of climbing Mount Everest. The experience consists of a sequence of scenes. Some scenes can only be observed (e.g., a helicopter ride with narration near the beginning of the experience; Figure 1), while other scenes require active participation from the user before the experience will proceed (e.g., crossing a crevasse by walking along a ladder that bridges the two sides; Figure 2).

Fig. 1: A scene from the helicopter ride.

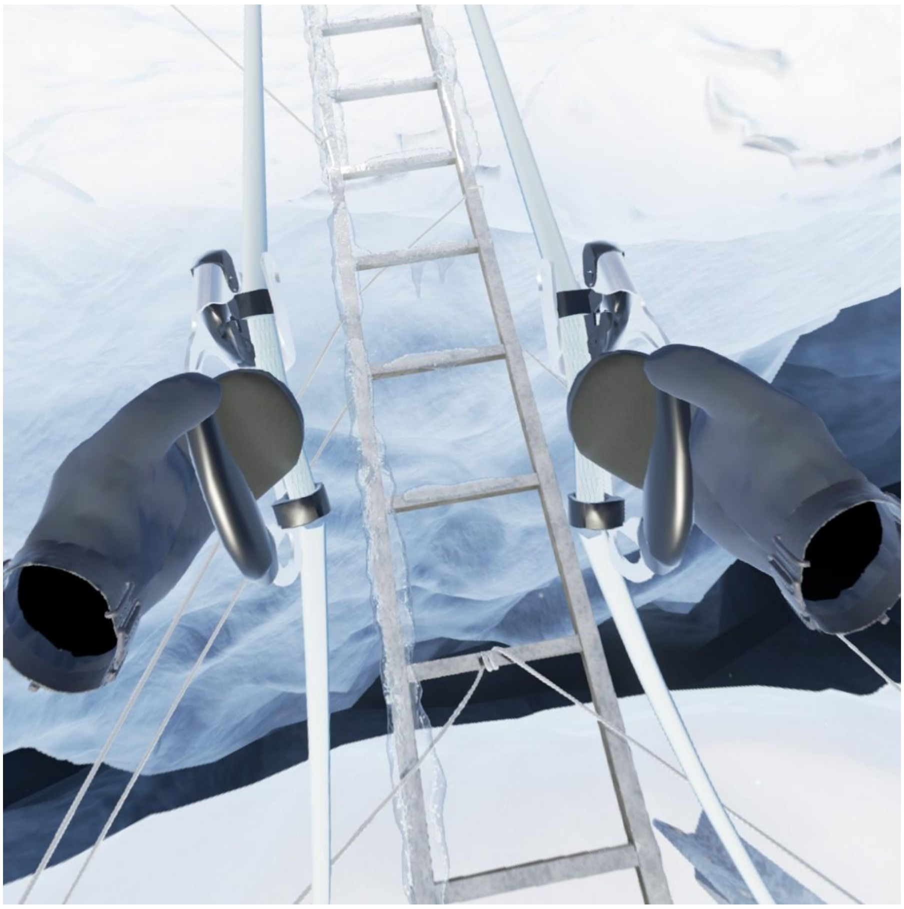


Fig. 2: A scene in which the user crosses a chasm using a ladder.

In every scene, the user can freely control their perspective in the virtual world by physically moving their head. In the scenes that require activity, the user's hands are represented by virtual mittens whose positions match the positions of the HTC Vive's hand-held controllers. Each scene that can only be observed causes an automatic transition to the next scene as soon as the former is complete. Each scene that requires interaction ends when the user performs a specific gesture in a particular context (e.g., waving in the direction of a virtual character); the next scene then begins automatically. Within an active scene, the user can use their virtual mittens to grasp various virtual objects, including ropes across a chasm and the rungs of a ladder (the user "climbs" the ladder by reaching for and grasping higher rungs to raise their virtual body; the user's feet are not tracked by the HTC Vive).

While the visuals of Everest VR are intended to be realistic (Sólfar Studios, 2016), some parts of the experience's simulation are intentionally unrealistic. For example, if the user stops grasping all of a ladder's rungs partway through a climb, their virtual body will not fall. Instead, the user's virtual, vertical position will remain unchanged, and the user will perceive that they are standing beside the ladder on a surface that is invisible in the virtual world. This decision to break from realism was made to support a more important objective of the overall experience, which was to provide a pleasurable fantasy of climbing Mount Everest.

3.2 Data Gathering

We gathered data through user testing, which was structured in 6 sections: a) an introduction to the testing, b) filling in a pre-questionnaire, c) experiencing the VR prototype, d) filling in a post-questionnaire, e) debriefing by watching a video of what happened during the experience, and f) thanking the user. Following the Google Design process [4], we planned and conducted 5 user testing sessions, each with one user. The user testing was conducted in a lecture room at Reykjavik University. We separated the VR play area from the experimenters by moving tables to define the area, as can be seen in Figure 3. Figure 4 shows one of our users experiencing the video game inside the playing area while the conductor of the evaluation observes.

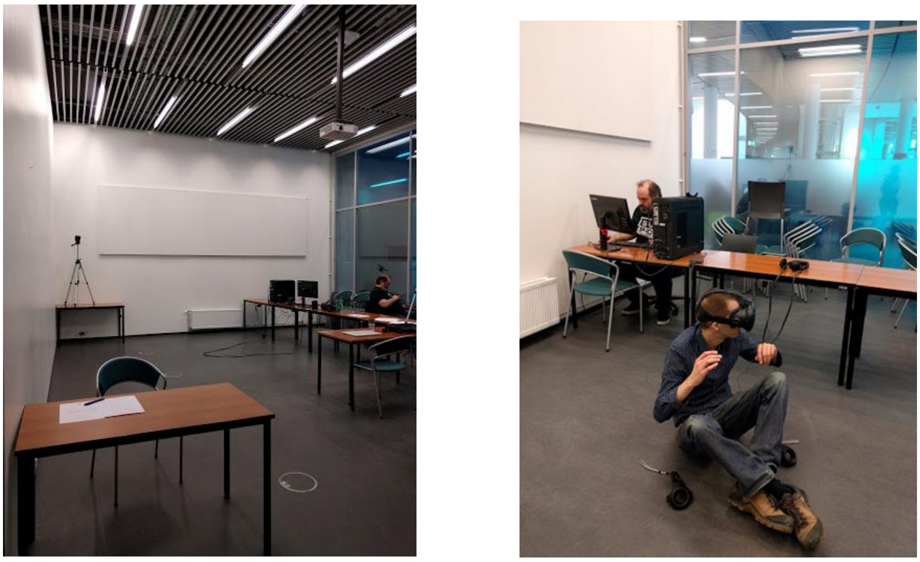
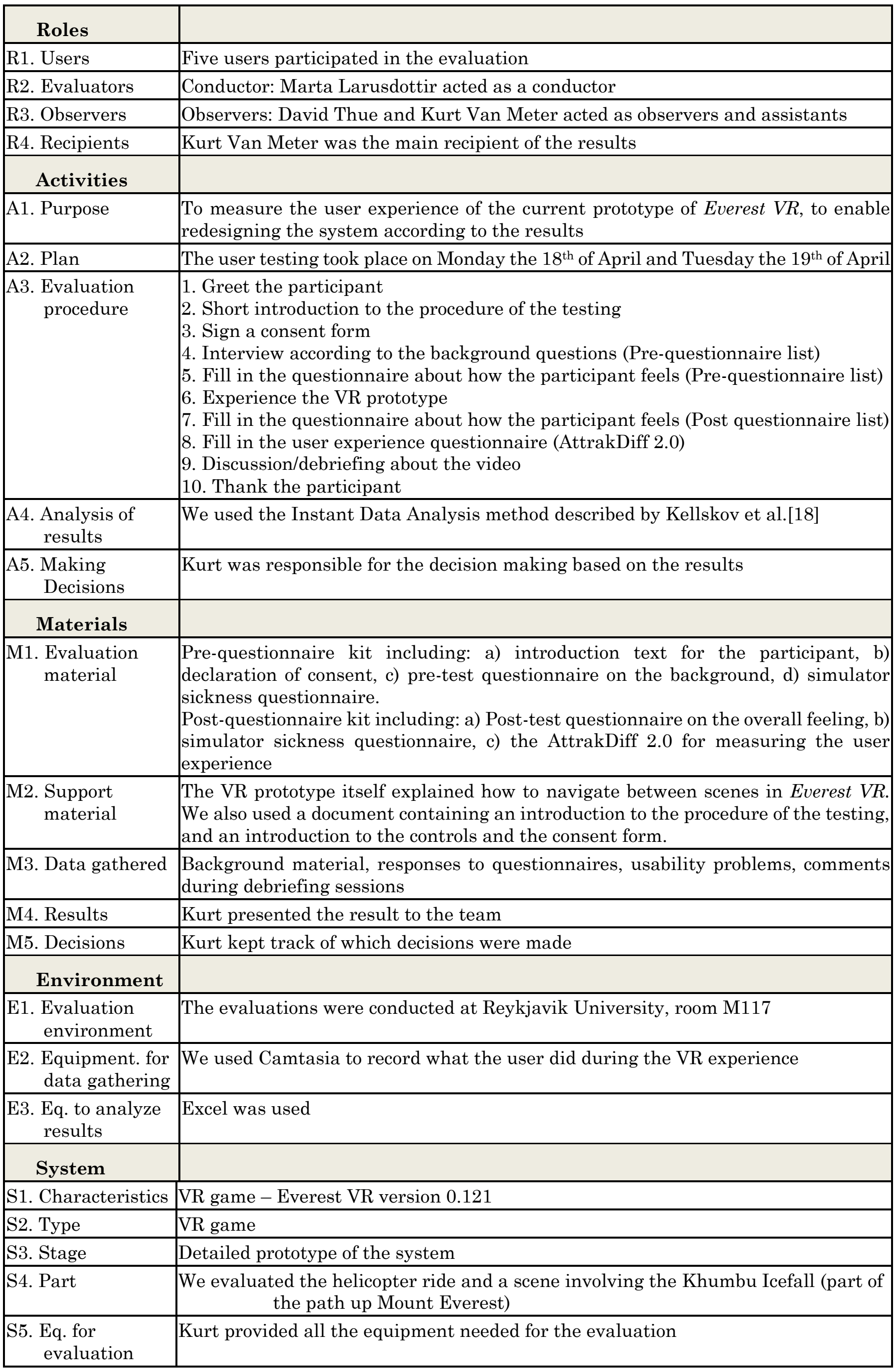


Fig. 3 and 4: The user testing area (Left) and a user participating in the VR experience (Right).

The execution of the user evaluations is shown in Figure 5, by using an early version of the RAMES framework [11].



Roles	
R1. Users	Five users participated in the evaluation
R2. Evaluators	Conductor: Marta Larusdottir acted as a conductor
R3. Observers	Observers: David Thue and Kurt Van Meter acted as observers and assistants
R4. Recipients	Kurt Van Meter was the main recipient of the results
Activities	
A1. Purpose	To measure the user experience of the current prototype of *Everest VR*, to enable redesigning the system according to the results
A2. Plan	The user testing took place on Monday the 18th of April and Tuesday the 19th of April
A3. Evaluation procedure	1. Greet the participant 2. Short introduction to the procedure of the testing 3. Sign a consent form 4. Interview according to the background questions (Pre-questionnaire list) 5. Fill in the questionnaire about how the participant feels (Pre-questionnaire list) 6. Experience the VR prototype 7. Fill in the questionnaire about how the participant feels (Post questionnaire list) 8. Fill in the user experience questionnaire (AttrakDiff 2.0) 9. Discussion/debriefing about the video 10. Thank the participant
A4. Analysis of results	We used the Instant Data Analysis method described by Kellskov et al.[18]
A5. Making Decisions	Kurt was responsible for the decision making based on the results
Materials	
M1. Evaluation material	Pre-questionnaire kit including: a) introduction text for the participant, b) declaration of consent, c) pre-test questionnaire on the background, d) simulator sickness questionnaire. Post-questionnaire kit including: a) Post-test questionnaire on the overall feeling, b) simulator sickness questionnaire, c) the AttrakDiff 2.0 for measuring the user experience
M2. Support material	The VR prototype itself explained how to navigate between scenes in *Everest VR*. We also used a document containing an introduction to the procedure of the testing, and an introduction to the controls and the consent form.
M3. Data gathered	Background material, responses to questionnaires, usability problems, comments during debriefing sessions
M4. Results	Kurt presented the result to the team
M5. Decisions	Kurt kept track of which decisions were made
Environment	
E1. Evaluation environment	The evaluations were conducted at Reykjavik University, room M117
E2. Equipment. for data gathering	We used Camtasia to record what the user did during the VR experience
E3. Eq. to analyze results	Excel was used
System	
S1. Characteristics	VR game – Everest VR version 0.121
S2. Type	VR game
S3. Stage	Detailed prototype of the system
S4. Part	We evaluated the helicopter ride and a scene involving the Khumbu Icefall (part of the path up Mount Everest)
S5. Eq. for evaluation	Kurt provided all the equipment needed for the evaluation

Fig. 5. Execution of the user testing explained using RAMES

During the user testing, the participants filled in two questionnaires: one before the VR experience, and another afterwards. The pre-questionnaire covered background questions and the simulator sickness questionnaire proposed by Kennedy et al. [10]. The post-questionnaire contained one question on the overall feeling of the participant, the simulator sickness questionnaire, and the AttrakDiff 2.0 questionnaire by Hassenzahl [12]. We estimate that the entire process took approximately 60 person hours, including preparation, conducting the user testing, and analysing the data.

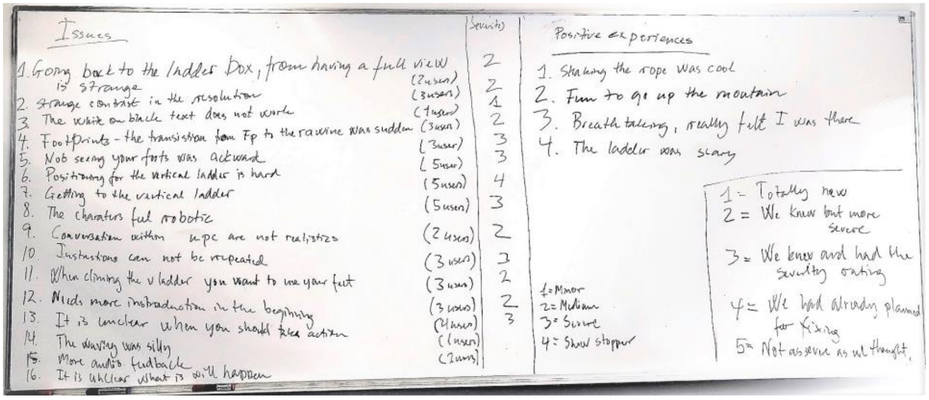
3.3 The Participants

We had 4 males participating in the user testing and one female. Their age was from 25 to 54 and they had all experienced virtual reality before (some only once or twice). Three said they had played a lot of video games, one said some, and one said that he/she had played none. We also asked about their experience in hiking and mountain climbing: one had extensive experience, three had some experience, and one had no experience. Four of the participants had heard about Everest VR before.

We asked if the participants ever experienced fear of heights or vertigo and some had experienced some, one said he/she had not, and one had quite serious fear of heights, but no vertigo. We also asked them to fill in a simulator sickness questionnaire to be able to see the difference in how they felt before and after experiencing the VR prototype.

3.4 The Data Analysis

To analyse both our observations of the users experiencing the *Everest VR* prototype as well as what they told us during the debriefing, the conductor and the two observers met the day after the user testing was complete and performed an instant data analysis as proposed by Kellskov et al. [13]. Some of the results of our instant data analysis session can be seen in Figure 6. In total, we recorded 30 user problems and four positive experiences.



Issues

Severity

1. Going back to the ladder box, from having a full view is strange (2users) 2
2. strange contrast in the resolution (3users) 2
3. The white on black text does not work (1user) 1
4. Footprints - the transistion from Fp to the ravine was sudden (3users) 2
5. Not seeing your feets was ackward (3user) 3
6. Positioning for the vertical ladder is hard (5user) 3
7. Getting to the vertical ladder (5users) 4
8. The charaters feel robotic (5users) 3
9. Conversation with npc are not realistics (2users) 2
10. Instrustions can not be repeated (3users) 3
11. When climing the v ladder you want to use your feet (3users) 2
12. Needs more instroduction in the beginning (3users) 2
13. It is unclear when you should take action (4users) 3
14. The waving was silly (1user)
15. More audio feedback (2users)
16. It is unclear what is will happen

Positive experiences

1. Shaking the rope was cool
2. Fun to go up the moutain
3. Breath taking, really felt I was there
4. The ladder was scary

1=Minor
2=Medium
3=Severe
4=Show stopper

1 = Totally new
2 = We knew but more severe
3 = We knew and had the severity rating
4 = We had already planned for fixing
5 = Not as severe as we thought.

Fig. 6. The results of the instant data analysis session.

We brainstormed on the outcome of the user testing on a whiteboard by writing a list of user problems. For each problem we wrote the number of users having that problem and agreed on a severity for that problem. We categorized the severity as follows: 4 = Showstopper, 3 = Severe problem, 2 = Moderately severe problem, and 1 = Minor problem.

3.5 Analyzing the Impact of the User Problem List

Two months after we completed the user testing sessions, one of the observers of the user testing (who was also responsible for Quality Assurance (QA) at Solfar Studios) analyzed the impact of the user problems for further development of the system. He categorized each problem in these categories:

1 = Addressed
2 = No action
3 = No action, new tech needed
4 = No action, good for future design of the system
5 = Action not decided yet

He also remarked on whether any of the problems were particularly useful; these included problems that the team did not know about before the testing began, or problems whose severity they had estimated incorrectly.

4. Results

In this section, we describe the results of the user testing. First we describe the results from the questionnaires, and then we describe the results from the user problem list and the categorization of the impact of the user problems.

4.1 The User Problem List

The user problem list contained 30 problems. In Table 1, the number of problems in each severity category can be seen.

Table 1: Number of User Problems in Each Severity Category

Severity Category	Number of problems	Average number of users
Showstopper	1 problem	5 users
Very severe problem	5 problems	3,6 users
Moderately severe problem	15 problems	2 users
Minor problem	9 problems	1 user
Total:	30 problems	

As shown in Table 1, 21 of the found problems were severe (either moderately, very, or a showstopper). The showstopper had the description: “Getting to the vertical ladder by waving confused everyone, and the icon is almost invisible – this is particularly bad because you feel trapped in a corner (and are in the corner of the space)”. An example of a very severe problem is: “Instructions cannot be repeated (people missed, didn‘t hear or were too overwhelmed by the visuals to take in verbal instructions)”. Additionally, an example of a moderately severe problem is: “Want more audio feedback (immersion – environmental audio)”.

4.2 The Impact from the User Problems List

The impact of the user problem list was estimated by asking one of the observers, (who was also responsible for Quality Assurance (QA) at Solfar Studios) to report what decisions had been made regarding each of the user problems, two months after the user testing. In Table 2, the results of this categorization are described.

Table 2. Number of User Problems in Each Severity Category

Severity Category (number of problems)	**Impact**	**Marked as useful**
Showstopper (1)	Addressed	Very useful
Very severe problem (5)	4 addressed, 1 future design	1 very useful, 1 useful
Medium severe problem (15)	10 addressed, 2 no action 1 new tech needed, 1 not decided, 1 future design	4 useful
Minor problem (9)	4 addressed, 3 no action, 2 new tech needed	

The QA person especially marked 3 problems that were categorized as a showstopper or very severe problems right after the evaluation, as very useful or useful two months after the user testing. For the showstopper he remarked: “This testing was key in pointing out the importance of that.” Additionally, for one of the very severe problems he remarked: “Huge impact on this from testing and we continue to reposition to find the best layout”.

Out of the 15 problems marked as medium severe problems right after the evaluation, he marked 4 as useful two months after the evaluation. He also remarked for one of the medium severe problems: “Being addressed, and was useful to have fresh eyes to underline the importance of this”.

5. Discussions

Public adoption of Virtual Reality devices has been slow [15]. As a result, a large percentage of the potential participants of software evaluations will have never experienced any environments using Virtual Reality technology. When evaluating the use of VR devices that perform positional tracking (such as the HTC Vive), some amount of initial unfamiliarity seems likely to persist even for participants who have experienced VR environments before, as the most widespread VR devices (e.g., the Gear VR and Google Cardboard) lack any positional tracking; this lack substantially limits the user's experience of a virtual world. Having participants who are unfamiliar with (aspects of) VR technology represents a serious challenge to any evaluation of a VR experience, as each user's reaction to the technology itself will likely be confounded with their reaction to the experience that one hopes to evaluate. To evaluate a VR experience independently from its supporting VR technology, one must consider only participants who have sufficient prior familiarity with that particular technology. Unfortunately, the percentage of the population that meets this criteria is likely to be very low, making it difficult to obtain a sufficient sample size to support reliable generalizations. A potential alternative could be to attempt to control for prior VR familiarity by applying statistical analysis techniques, but doing so would require a reliable, population-general model of how prior VR familiarity affects the metrics along which a target experience is to be evaluated. To the best of our knowledge, such a model does not yet exist. Until the population's familiarity with VR devices increases, it seems likely that all studies of VR experiences will suffer from the confounding effects of each user's unfamiliarity with the VR technology being used.

User involvement during agile software development in the software industry has been found to be both informal and explicit [16]. Feedback from users is gathered in an informal way, and not through formal user testing [2]. Still formal user testing has been given the highest rating of methods used by practitioners for involving users in the software industry [3]. The main reason for not conducting formal user testing is that it is time consuming in relation to the benefits the developers receive from the results of the evaluations [2]. Little has been done to try to estimate the value for software developers of gathering feedback from users through user testing. The study we have described shows that, even with just a few participants, useful information was indeed found in the formal testing of a novel VR experience.

References

1. Ardito, C., Buono, P., Caivano, D., Costabile, M. F., Lanzilotti, R., Bruun and A., Stage, J.: Usability evaluation: A survey of software development organizations. *In: Proceedings of international conference on software engineering and knowledge engineering (SEKE 2011)* Miami, FL, USA. 282-287 (2011).
2. Bak, J. O., Nguyen, K., Risgaard, P. and Stage, J.: Obstacles to usability evaluation in practice: A survey of software development organizations, *In: Proceedings of NordiCHI 2008 conference*, Lund, Sweden. ACM Press (2008).
3. Bradshaw, T. VR industry faces reality check on sales growth, In Financial Times (2017): Online https://www.ft.com/content/f7e231ee-fc84-11e6-96f8-3700c5664d30, last accessed 2018/05/06
4. Cajander, Å., Larusdottir, M., & Gulliksen, J. Existing but not explicit-The user perspective in scrum projects in practice. In *IFIP Conference on Human-Computer Interaction* (pp. 762-779). Springer, Berlin, Heidelberg, (2013).
5. Cockton, G.: I can't get no iteration. *Interfaces,* 63, (2005).

6. Cuomo, D.L. and Bowen, C.D.: Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting with Computers,* 6(1), 86-108, (1994).
7. Desurvire, H. W., Kondziela, J. M. and Atwood, M. E.: What is gained and lost when using evaluation methods other than empirical testing. *In*: *People and Computers VII.* Cambridge, UK: Cambridge University Press, 173-201, (1992).
8. Hassenzahl, M., Burmester, M., & Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In Mensch & Computer 2003 (pp. 187-196). Vieweg+ Teubner Verlag (2003).
9. Hassenzahl, M. and Tractinsky, N., User experience-a research agenda. Behaviour & Information Technology, 25(2), 91-97, (2006).
10. Hornbæk, K. and Frökjær, E.: Comparing usability problems and redesign proposals as input to practical systems development. *In: Proceedings of the CHI 2005 conference*, Portland, Oregon, USA. ACM Press (2005).
11. ISO 9241-210: Ergonomics of human-system interaction -- Part 210: Human-centred design process for interactive systems. Geneva, Switzerland: International Organisation for Standardization, (2010).
12. ISO 9241-11: Ergonomic requirements for office work with visual display terminals. Geneva, Switzerland: International Organisation for Standardization, (1998).
13. Jeffries, R., Miller, J. R., Wharton, C. and Uyeda, K.: User interface evaluation in the real world: A comparison of four techniques. *In: Proceedings of CHI '91 conference*, New Orleans, Louisiana, USA, (1991).
14. Jia, Y., Larusdottir, M.K., and Cajander, A., The usage of usability techniques in Scrum projects. In: M. Winckler, P. Forbrig, R. Bernhaupt, eds. Proceedings of the HCSE 2012. LNCS, vol. 7623. Heidelberg: Springer, 331–341, (2012).
15. Karat, C.M., Campbell, R. and Fiegel, T.: Comparison of empirical testing and walkthrough methods in user interface evaluation. *In: Proceedings of the CHI '92 conference*, Monterey, CA, USA. ACM Press, (1992).
16. Kennedy, R. S., Lane, N. E., Berbaum, K. S., Lilienthal, M. G., Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness, The International Journal of Aviation Psychology, 3:3, 203-220, (2009), DOI: 10.1207/s15327108ijap0303_3
17. Knapp, J., Zeratsky, J., Kowitz, B., Sprint: How to solve big problems and test new ideas in just five days. Simon and Schuster, (2016).
18. Kjeldskov, J., Skov, M. B., & Stage, J.: Instant data analysis: conducting usability evaluations in a day. In Proceedings of the third Nordic conference on Human-computer interaction pp. 233-240, (2004).
19. Larusdottir, M. K., Cajander, Å., Gulliksen, J.: Informal feedback rather than performance measurements – user-centred evaluation in Scrum projects, Behaviour & Information Technology, 33:11, 1118-1135, (2013) DOI: 10.1080/0144929X.2013.857430
20. Larusdottir, M. K., Gulliksen, J, Hallberg, N.: The RAMES Framework for Planning and Documenting User-Centred Evaluation. In publication by the journal: Behaviour and Information Technology, (2018).
21. Law, E.L.-C.: Evaluating the downstream utility of user tests and examining the developer effect: A case study. *International Journal of Human-Computer Interaction,* 21(2), 147-172, (2006).
22. Law, E. L., Larusdottir, M. K.: Whose Experience Do We Care About? Analysis of the Fitness of Scrum and Kanban to User Experience, International Journal of Human–Computer Interaction, 31:9, 584-602, (2015) DOI: 10.1080/10447318.2015.1065693
23. Lazar, J., Feng, J. H. and Hochheiser, H.: Research methods in human-computer interaction. John Wiley & Sons Inc., (2009).

24. Nielsen, J.: How many test users in a usability study. Internet: https:// https://www.nngroup.com/articles/how-many-test-users/ (2012).
25. Nielsen, J. and Phillips, V.L.: Estimating the relative usability of two interfaces: Heuristic, formal, and empirical methods compared. *In: Proceedings of the INTERCHI '93 conference,* Amsterdam, Netherlands, ACM Press, (1993).
26. Sólfar Studios. Everest VR. (2016) Online: http://www.solfar.com/everest-vr/, last accessed: 2018/05/06.
27. Thorgeirsson, T. and Larusdottir, M.K.: Case study: Are CUP attributes useful to developers? *In: Proceedings for the COST-294 workshop: Downstream utility: The good, the bad and the utterly useless usability feedback*, Toulouse, France. 50-54, (2007).
28. Vredenburg, K., Mao, J.-Y., Smith, P.W. and Carey, T.: A survey of user-centered design practice. *In: Proceedings of the CHI 2002 conference*, Minneapolis, Minnesota, USA. ACM Press, (2002).