Tampereen yliopisto

Petrus Eskelinen

LEARNABILITY EVALUATION OF VIRTUAL REALITY APPLICATIONS

Master's thesis
Faculty of Information Technology and Communication Sciences
Inspector: Jari Varsaluoma
Inspector: Matti Gröhn
May 2023

ABSTRACT

Petrus Eskelinen: Learnability evaluation of virtual reality applications
M.Sc. Thesis
Tampere University
Master's Degree Programme in Information Technology
May 2022

Learnability is a part of the usability of systems. It relates to the ease with which users can learn to use a system on the first times of usage and how quickly users become efficient with the system over time. As a primary research goal, this work evaluates the learnability of a virtual reality application in a case study, using evaluation guidelines combined from relevant literature. As a secondary research objective, an outlook on the state of automation of usability and learnability testing without end-users is formed based on latest academic works. These learnability evaluation guidelines provide a good foundation for designing and running an evaluation for VR, but there are still gaps to be filled. Especially the qualitative analysis of the screen recordings of the VR headsets and the video of participants' behaviour needs more in-depth instructions.

As the case study, the Glue VR platform developed by Glue Collaboration was evaluated for learnability. The platform is a virtual office collaboration application, which allows users to have meetings in virtual reality spaces. A group of 10 participants with varying virtual reality expertise were involved in the study. The potential learnability problems were listed and presented to Glue Collaboration.

The research on automation of usability testing was found to be focusing on evaluating level design in 3D games. Similar tools could be applied in the Glue platform to find glitches in the environment. An AI agent with an affective model presented in one of the studies could possibly be used to evaluate VR applications in more depth in the future.

Keywords: Virtual reality, learnability evaluation

TABLE OF CONTENTS

1 INTRODUCTION

As VR (Virtual Reality) programs are often highly complex (and still quite novel) systems, software requirements specification and testing can prove to be arduous tasks (Correa Souza et al. 2018). Object simulation, collision dynamics and a wide range of interactions are some of the factors that explain the complexity. Glue collaboration is the developer of the Glue VR platform, a virtual office environment where meetings can be held. A specific challenge with Glue is to make the virtual collaboration sessions resemble physical ones as much as possible. This would lead to a reduced mental workload while learning to use the platform. Many features are designed to make transition to be smoother for users (especially inexperienced VR users). These include virtual counterparts of office appliances, such as whiteboards, sticky notes, and projector screens. A tablet to control actions in the virtual environment is programmed to behave like a physical tablet; users make a press gesture to activate buttons. Glue Collaboration wants to know whether they are making the right call with these design choices and would like a clear way to evaluate learnability of the product and specific features.

However, literature on learnability evaluation of VR applications is almost non-existent. In this paper, learnability-focused usability testing guidelines drawn from literature, including usability and learnability papers that focus on traditional systems, are presented. This will answer **RQ1**: **How can the learnability of a VR application be measured?** Moreover, Glue Collaboration is aware of the time-consuming nature of user testing and thus propose the second research question **RQ2: Can learnability evaluation of VR applications be automated without end-user input?**

Next in chapter 2, related work will be presented. Chapter 3 lays out the design and planning of a case study, the learnability evaluation of the Glue VR platform. In chapter 4, the evaluation results are presented and analysed. Finally in chapter 5, the results are discussed and conclusions are made.

2 RELATED WORK

In this chapter, previous research concerning usability and user experience will be introduced. Learnability definitions, evaluation methods and metrics and guidelines are presented as well.

Before discussing learnability, it is worthwhile to understand more of the process of human learning in the software context. A paper by Kao et al. (2021) summarizes the cognitive theory of multimedia learning, stating learning is based on three core cognitive principles. The *dual channel* principle implies that aural and visual information is processed by their own respective channels. Interestingly, other senses are not considered in the paper. The *limited capacity* principle states that these channels can only withhold a certain amount of information at a time. The psychologist George Miller does argue in a famous paper that humans can only process 7 plus or minus two pieces of information in their working memory (Miller, 1956). The third principle, namely *active processing* principle, argues that learning happens by comparing incoming information to existing knowledge by means of organization, filtering, and selection.

2.1 Defining Usability and User Experience

Usability is a concept revolving around the ease and efficiency of the use of products or services. User experience is a similar term, sometimes used interchangeably with usability.

The definitions according to standard ISO 9241, 2019:

Usability: “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.”

User Experience: “person’s perceptions and responses resulting from the use and/or anticipated use of a product, system or service.”

From the definition of usability, the aim of usability tests can be derived (Hertzum, 2020). The tests must provide a measurement of three qualities: 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”) and satisfaction (“freedom from discomfort and positive attitudes towards the use of the product”) (ISO 9241, 2019).

By definition, user experience is how the use situation (and anticipation of use) feels from the user's subjective perspective. The experience includes "All the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors and accomplishments that occur before, during and after use" (ISO 9241, 2019).

2.2 Defining Learnability

Learnability is an important part of usability, claimed by some to be the key usability component (Grossman et al. 2009). As cited by Lee & Sah (2020), learnability is referred to how long it takes (Shneiderman & Plaisant, 2005) or how easy it is for typical users (Nielsen, 1993) to learn a system and to reach a certain level of competency (Lesgold, Ivill-Friel, & Bonar, 1989). Although these definitions form a general idea of learnability, it's unclear what exactly is that level of competency and how do we take learnability into account after that point, in the journey from "reasonable" to "expert" performance. The terms "initial learnability" and "extended learnability" correspond to these parts of the user experience (Grossman et al. 2009).

Grossman et al. (2009) also cite a few definitions that take extended learning into account: "Ease at which new users can begin effective interaction and achieve maximal performance" (Dix et al. 2003), and "Initial user performance based on self-instruction" and "[allowing] experienced users to select an alternate model that involved fewer screens or keystrokes" (Butler 1985) are such statements. A structure was laid out that summarizes the different learnability definitions (Figure 1).

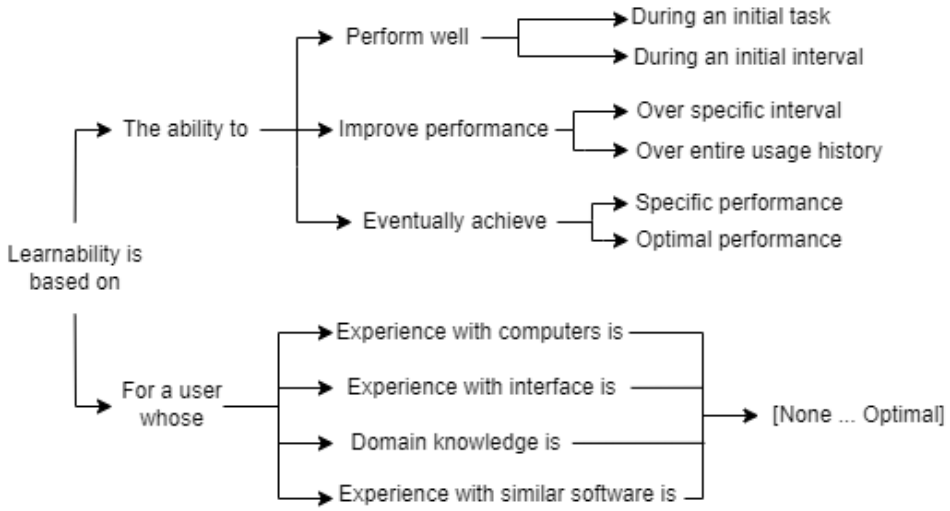

Figure 1. *Taxonomy of Learnability definitions (Redrawn based on Grossman et al. 2009)*

With the aid of the taxonomy in figure 1, researchers and developers can narrow down the focus of a learnability study. For example, we may want to find out whether a user whose experience with virtual reality systems is non-existent can perform well during an initial interval of 2 hours. Driving the selection of the focus could be user feedback or observed problems in usage of the system.

2.3 Learnability Metrics

Much like the definitions of learnability, there seems to be no consensus on the metrics of measuring learnability. Ntoa et al. (2021) suggest, in addition to an expert-based cognitive walkthrough, to measure learnability with following numbers: Interaction errors over time, input errors over time and help requests over time. In an NNgroup article by Joyce (2019), time on task is considered the most relevant metric, followed by the number of errors. Grossman et al. (2009) collected learnability metrics found in research from varying fields of studies, shown in Table 1.

Table 1. *Table 1 Learnability metrics (Grossman et al. 2009)*

Task Metrics: Metrics based on task performance T1. Percentage of users who complete a task optimally. T2. Percentage of users who complete a task without any help. T3. Ability to complete task optimally after certain time frame. T4. Decrease in task errors made over certain time interval. T5. Time until user completes a certain task successfully. T6. Time until user completes a set of tasks within a time frame. T7. Quality of work performed during a task, as scored by judges.
Command Metrics: Metrics based on command usage C1. Success rate of commands after being trained. C2. Increase in commands used over certain time interval. C3. Increase in complexity of commands over time interval. C4. Percent of commands known to user. C5. Percent of commands used by user.
Mental Metrics: Metrics based on cognitive processes M1. Decrease in average think times over certain time interval. M2. Alpha vs. beta waves in EEG patterns during usage. M3. Change in chunk size over time. M4. Mental Model questionnaire pre-test and post test results.

Subjective Metrics: Metrics based on user feedback S1. Number of learnability related user comments. S2. Learnability questionnaire responses. S3. Likert statements.
Documentation Metrics: Metrics based on documentation usage D1. Decrease in help commands used over certain time interval. D2. Time taken to review documentation until starting a task. D3. Time to complete a task after reviewing documentation.
Usability Metrics: Metrics based on change in usability U1. Comparing “quality of use” over time. U2. Comparing “usability” for novice and expert users.
Rule Metrics: Metrics based on specific rules R1. Number of rules required to describe the system.

A learnability attributes model has been developed, specifying multiple subcomponents of learnability under six main components: Interface understandability, feedback suitability, predictability, task match, system guidance appropriateness and operational momentum. In a case study, the model was tested by means of subjective Likert scale questionnaire, using both self-developed and widely accepted questionnaire prompts. (Rafique et al. 2012)

A study by Kaminska et al. (2022) states that the most reliable objective metrics are derived from the user’s biomedical signals: motion tracking, eye tracking, heart rate monitoring, EEG brain signals and speech analysis are deemed most suitable for usability testing for VR applications. In a case study performed by Kaminska et. al (2022), these types of data were collected and analysed by AI in an attempt to automize the analysis phase of usability studies. Upon the AI usability problem recognition accuracy of 84%, they state: “While 84.23% maximum recognition rate is not high enough to consider it a valid and proven automatic usability testing method, it is definitely enough to grant further exploration”. On the other hand, in the study it is explained that the simplest way to obtain objective data is by observing user behavior, accompanied with recordings. This method has a problem of being biased. This may be due to the varying expertise of the evaluators in domain knowledge and/or usability evaluation.

Kaminska et al. (2022) define user-subjective metrics to be the perceived feelings of users taking part in an experience. They argue that subjective metrics aren’t valuable when testing early-stage prototypes and are rather used to evaluate finished products.

The data is typically acquired by questionnaires and interviews. For example, another study by Kao et al. (2021) used a subscale of PENS (Player Experience of Need Satisfaction) scale to measure controls learnability in a VR game. The subscale included three statements about controls learnability and participants were to select an answer on a 7-point Likert scale. Although subjective metrics are easy to gather, a drawback of subjective metrics with VR usability tests is that users may associate a novel VR experience with the application itself.

2.4 Desinging VR for learnability

The paper by Kao et al. (2021) discusses the implications of the cognitive theory of multimedia learning for designing VR applications. Four best practices are laid out in the theory. First, using words with graphics is better than using either of these alone. However, the graphics should relate closely to the object of learning, and not be merely decorative. Aligning words close to the graphics is another important step, for example showing informational controller tooltips close to the representation of the controllers in the virtual world. The third principle states that unnecessary material hinders learning – anything that is irrelevant to the object of learning can be distracting and 'seduces' attention elsewhere, preventing the formation of mental models. For example, if the VR controllers are represented in the virtual world as 3D models, a separate diagram explaining their function is unnecessary – the information can be rather shown right next to the 3D models of the controllers. Lastly, using visual cues is recommended, directing attention to an object of learning. Continuing with the controller example, adding a flashing animation to the thumbstick to indicate that interacting with it allows movement.

The study concludes after empirical testing, that using text-based tooltips along with spatial graphics (the controller representation), is more effective in controller learnability than using just text or text and diagrams (Kao et al. 2021).

2.5 Evaluation Methodologies

Usability (as well as learnability) studies can be categorized as summative, formative or predictive. Examples of these methods from both VR and non-VR studies are laid out below.

2.5.1 Formative methods

Formative studies are described to be iterative, done during the development of the product. Such a study tries to find the most significant usability issues and common errors, as well as features loved by users. The idea is to make recommendations and repeat the study later. (Tullis & Albert, 2008)

A widely used formative method is the heuristic evaluation. Hillmann (2021) proposes that the well-known usability heuristics for UI design by Nielsen (1994) can be utilized in the VR context, claiming using them as a blueprint to evaluate VR interface design and usability is a good starting point. Summarized, they are:

- 1: Visibility of system status (keep the user informed)
- 2: Match between the system and the real world (speak the user's language)
- 3: User control and freedom (help the user avoid unwanted situations)
- 4: Consistency and standards (follow conventions)
- 5: Error prevention (present users with confirmation options)
- 6: Recognition rather than recall (minimize the user's memory load)
- 7: Flexibility and efficiency of use (cater to beginners and advanced users)
- 8: Aesthetic and minimalist design (eliminate irrelevant information)
- 9: Help users recognize, diagnose, and recover from errors (communicate problems and solutions clearly)
- 10: Help and documentation (make it easy to search for answers)

Think-aloud protocol is another one of the most common types of formative usability testing and was originally used to evaluate initial learnability (Grossman et al. 2009). It involves directing the test participants to verbalize their thoughts, plans, strategies and issues during the use of the product, as well as while studying possible system documentation. Another formative method is the "coaching" or "question-asking" protocol. During the testing, and expert sits next to a participant, who is instructed to ask about anything that comes into their mind while using the product. One study that used a similar protocol tracked the types of questions asked by and help given to participants as a metric. The coaching protocol is thought to measure initial learnability. (Grossman et al. 2009)

Grossman et al. 2009 proposed a Question-Suggestion protocol, a similar approach to the “coaching” protocol. Instead of just answering users’ questions, the coach may offer suggestions if inefficient use of the software is noticed. In table 2, protocol instructions for both the coach and participant are outlined.

Table 2. *Question-Suggestion protocol instructions (Grossman et al. 2009)*

Question-suggestion Protocol - Instructions to Participant: 1. Ask relatively specific, procedural questions. 2. Try to answer your own questions first, but do not engage in extensive problem solving. 3. Focus on getting the task done, as you would in the real world.
Question-suggestion Protocol - Instructions to Moderator: 1. Reply with specific, procedural questions. 2. Do not tutor or explain at length. 3. Maintain focus on the task, even when providing suggestions. 4. Provide suggestions if you notice inefficient usage behaviours and if the suggestions would likely be welcomed and beneficial to the user.

When thinking about using such a protocol for learnability evaluation, it would first seem counterproductive, because some metrics (such as time on task) would be skewed, as the instructor gives hints to improve performance. Though, when compared to the think-aloud protocol, the Question-Suggestion protocol seemed to find more learnability issues (Grossman et al. 2009). By allowing the instructor to provide suggestions, users advance in expertise faster, revealing a larger scope of learnability problems while going into the area of extended learnability.

When preparing to test the Question-Suggestion protocol in a case-study, Grossman et al. (2009) argued that the arising learnability issues should be recorded by an experimenter separate from the instructor (coach), to allow them to fully focus on the participant’s actions.

2.5.2 Summative methods

The goal of summative studies is to determine whether the product reaches its usability

targets, in other words to assess the overall usability (Grossman et al. 2009). A summative test is unrelated to product-improvement efforts and its aiming for example to determine which of two existing products is preferable (Hertzum, 2020).

An example of a summative method is quantitative analysis of users' performance during the task and time spent learning documentation. As cited by Grossman et al. (2009) study by Davis and Wiedenbeck gave participants a fixed timeframe to complete a task, and the resulting products were scored by judges. (Grossman et al. 2009)

2.5.3 Predictive methods

Predictive methods are usability testing methods that do not require users. An older one and manual one, the GOMS method, will be discussed in addition to more recent methods using simulation and automation.

GOMS, a framework to evaluate and predict learnability was developed in 1983. It is a model to describe the knowledge needed to perform tasks on different systems. GOMS stands for Goals, Operators, Methods and Selections rules. A GOMS model will contain a description of the Methods to complete goals. Methods will be one or many Operators that user can perform. If many Methods can be used to reach a Goal, Selection rules determine the most appropriate one. The purpose of the model is to estimate the peak execution time and learning time for a task. NGOMSL (Natural GOMS Language) is a structured, program-like natural language to represent the Methods and Selection rules. (Kieras, 1996)

To estimate learning time (the time needed to learn how to operate the interface), is determined by the total length of the methods, calculated by the number of NGOMSL statements in the model of the interface. This is the amount of procedural knowledge needed to operate the system for *all of the possible tasks* under consideration. Estimating execution time of a task (the time to complete a task with no errors) is determined by the amount and content of NGOMSL statements that must be executed to complete that particular task. (Kieras, 1996). These include the primitive external times of the Operators. For example, the primitive external operator times are 0.1 s for a mouse button press or release (Card et al., 1983), 0.2 s for a mouse move (Gong & Elkerton, 1990), and 0.8 s for a mouse wheel scroll (Ramkumar et al., 2017).

A model for measuring learnability (Lee & Sah, 2020) was developed based on the NGOMSL and experimented in practice with a website usability test. The model consists of a mathematical formula for calculating the learning level after repetitive use of a

system. The learning level is a number between 1 and 0, 1 being “perfect learning” (user can perform at maximum efficiency) and 0 being “never learning” (user can never complete a task). The learning level is assessed for each task separately.

The formula for the learning level is derived partly from the difference of “expertise time” (estimated perfect time) and “actual time” (the real task time of test users), denoted as the “learning deficit”. The expertise time was estimated using the NGOMSL model. The relationship between learning deficit and learning level is showcased in figure 2.

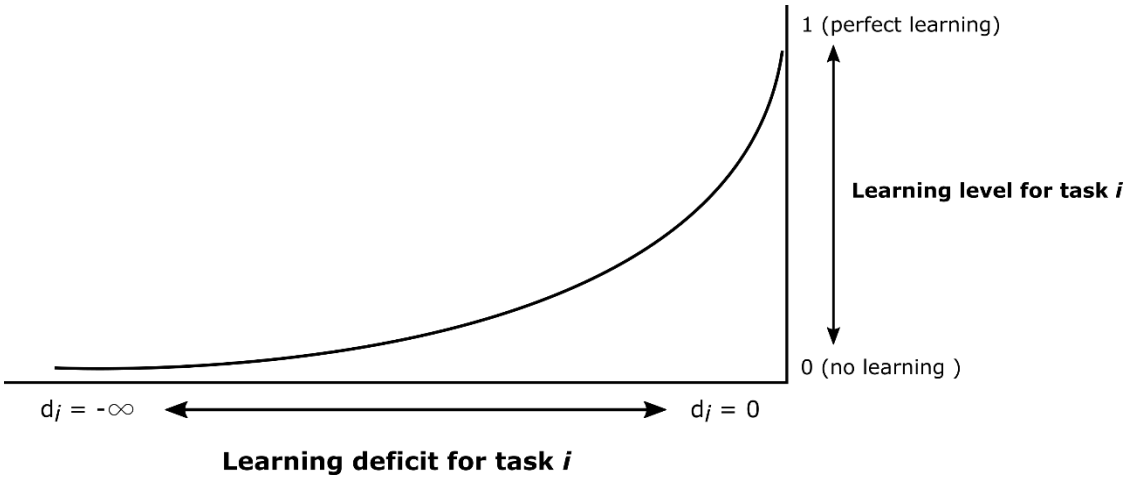

Figure 2. *Learning deficit curve (Redrawn based on Lee & Sah, 2020)*

The usability test was conducted on a website with four varying states of aesthetics and usability, with 64 users. The usability was tampered with intentionally, by making navigation inconsistent. Everyone completed a sequence of two tasks five times. On finishing the first and the last task, participants were asked to fill out a usability questionnaire, providing information of perceived usability and user satisfaction. In repetitive use during the sample users’ first use session, perceived usability and user satisfaction remained relatively stable, while the effects of usability level on learnability became limited and weak. (Lee & Sah 2020)

The weak effect of usability level on learnability might be explained by the way the usability level was changed in the experiment: inconsistent navigation (Figure 3). The tasks were not heavy on navigation, so it can be that memorizing the way to the location required by the tasks wouldn’t hinder performance on the subsequent trials.

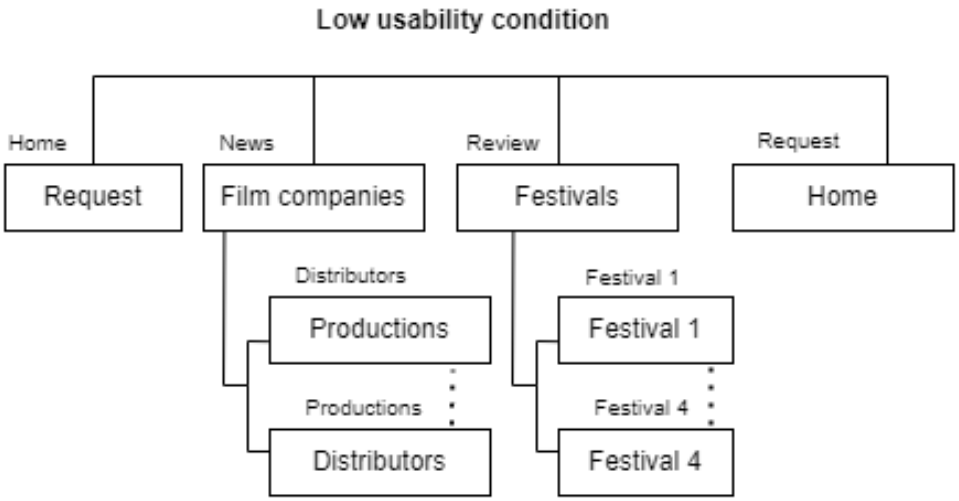

Figure 3. *Low usability sitemap (Redrawn based on Lee & Sah 2020)*

However, there is a problem for using the NGOMSL model for VR. A GOMS model can predict learning sensibly only if the lowest-level operators used in the model are ones that can be assumed the user already knows how to do, and for which stable time estimates are available (Kieras, 1996). When thinking about a new VR user, there's a chance neither of these are true. Furthermore, there's a need to determine the time estimates for many interactions that happen in the 3D virtual world, for example selecting menu items or manipulating objects, and this can be complicated. In fact, no VR studies utilizing NGOMSL model was found.

Automation methods

Articles regarding the usability evaluation of virtual reality applications were searched from the IEEE Xplore and ACM Digital Library databases, using the search term:

("automating" OR "automation") AND ("UX" OR "Usability" OR "User experience" OR "learnability") AND ("testing" OR "evaluation") AND "virtual reality"

No studies claiming to have invented a way to automate UX or usability testing on virtual reality technology were found. However, two studies are presented next, containing approaches that could be utilized in the VR context, due to them being able to navigate and evaluate 3D environments.

A recent study by Fernandes et al. (2021) focused on finding a way to automate UX testing. It highlights the fact that it's common to run automated functionality tests at

each iteration of a system but that can't be done for UX. The study agrees testing with users can be highly effective in gathering information about a system's pain points but deems it very time-consuming.

An intelligent agent with an affective model was proposed as a solution to start automation of UX tests. Such an agent was used to evaluate the UX of a simple 3D maze game. The agent was coded to interact with the door-opening buttons in the game, choosing randomly to explore any open door, and continues exploring every known location until finding the goal.

The affective model used (based on Core Affect theory of emotions) assumes that emotion is derived from an initial affective state. An affective state consists of two dimensions, Valence (negative vs positive) and Arousal (calm vs exiting). While an emotional state cannot be defined with this information alone, it's a starting point of emotions.

In practice, both dimensions were assigned a value between –5 and 5 and in the beginning, they start from 0 and are modified according to the following rules: Whenever the agent accomplishes the main objective or sub-objective, Valence increases by 1. If the agent doesn't accomplish anything for 10 seconds, Valence decreases by 0.4. When the agent finds an interactable object, Arousal increases by 1. If it doesn't find any new interactable for 10 seconds, arousal decreases by 0.4.

Another notable study by Gordillo et al. (2021) researched automated playtesting as a part of game development process. It involved having curiosity driven agents go through complex 3D levels involving sequential jumps, elevators and climbable obstacles. The aim of this automated testing was gathering data to map unreachable areas in the levels, identify exploits and unintended mechanics and visualize ramifications of design choices.

Both of these studies propose an interesting way to automate testing. Though, in the context of an office productivity app like Glue, this method may not be able to provide suitable insight. Perhaps testing the 3D environment for flaws/glitches could be done using this approach. The agents would need to have a model to navigate and understand UI and other features as well.

Expert Evaluation

Usability and learnability of a system can be evaluated by experts. Cognitive walkthrough is a method in which reviewers from various backgrounds examine a pre-

defined task flow from the perspective of a novice user. While this method doesn't require recruitment of real users, it relies on the knowledge of the evaluators to make correct judgements. (Salazar, 2022)

When evaluating a task, it is walked through by stopping at each discrete step (any prompt or screen that comes up). What constitutes as a single step is roughly determined beforehand between the reviewers. For example, a login screen could be seen as 3 steps or 1 step. According to Salazar (2022), four questions will be discussed between reviewers at each step:

1. Will users try to achieve the right result?
2. Will users notice that the correct action is available?
3. Will users associate the correct action with the result they're trying to achieve?
4. After the action is performed, will users see that progress is made toward the goal?

Then, the group will determine whether a novice user could pass at that step and make reasonings why that is the case. If the answer to any of these questions is no, that step is marked as a failure. Once all steps and tasks of a flow have been examined, a summary of fail points is written (Salazar, 2022)

Cognitive walkthroughs are most effective for systems with complex, new or unfamiliar workflows and functionalities (Salazar, 2022). That makes a cognitive walkthrough a good alternative for evaluating the learnability of VR systems.

2.6 Usability and learnability testing guidelines

In this subchapter, literature related to best practices during the test sessions will be discussed. Due to the little amount of literature on specifically VR and learnability, some VR usability studies, and non-VR learnability studies will be considered as well.

A paper by Salvatore & Christina (2008) outlining 'Simple guidelines for testing VR applications' includes testing guidelines that are simple enough to be understood by "non-experts" yet based on scientific information. The guidelines, coined as a VR usability testing handbook, consists of 12 directives:

- Research Question (What areas to pinpoint in the study)
- Ethics (Potential hazards and informing the participants of them)

- Evaluation Method (Types of data to collect and methods for collection)
- Setup (Testing environment and technological setup)
- Participants (How many participants to recruit and target groups)
- Forms (Documents for participants to fill)
- Schedule (Estimating session durations)
- Test monitor and spectators (Roles of personnel in test session)
- Test plan (What to include in a test plan document)
- Pilot study (A mock-up test to reveal flaws in the test plan)
- Formal study (Contents of the study such as amounts of tasks)
- Results and presentation (Results analysis and conclusion)

A more recent work by Hertzum (2020) provides guidelines on usability testing, albeit not focusing on VR technology. This work and the article by Salvatore & Christina (2008) will be summarized and contrasted in this section, including findings from other relevant literature.

2.7.1 Designing and planning a study

Designing and planning includes outlining which product features to test, who the test users will be, how the test will be organized (task list, moderation), data collection methods (logging, questionnaires) and when and where the test will take place (Hertzum, 2020). These should be written out into a test plan that can be used to communicate the test to stakeholders. According to Salvatore & Christina, the test plan should include:

- Purpose
- Test objective
- Target participant profile
- Test design
- Task list
- Test environment/equipment
- Test moderation personnel
- Evaluation measures

Designing and planning the test should be led by a user experience expert, but members of other roles have much to give (Hertzum, 2020). As an example, designers and developers know what the product is supposed to do and what should be tested and help desk personnel know the issues that already commonly exist. The plan outlines the resources needed for the test and can be used to track the test progress (Salvatore & Christina, 2008).

2.7.2 Ethics and test environment

Upon ethics, the possibility exists that VR technology triggers medical disorders such as PTSD, or epilepsy and for this reason, a disclaimer about the issue should be shown to participants (Pavuna, 2019). She states minimizing the risk of physical injuries is also a key factor in VR studies, adding that cleaning up the lab space of clutter and positioning the play space away from walls and sharp corners is vital. Even swivel chairs can be considered dangerous due to a possible disorientation (Pavuna, 2019). Users must be able to stop the test at any time for any reason (Salvatore & Christina, 2008).

Evaluators are reminded to acknowledge participants with eyeglasses. Some VR headsets do not go comfortably with glasses and for this reason participants could be suggested to wear contact lenses for the duration of the study. (Pavuna, 2019)

Obtaining informed consent for recording audio and video, and to participate in the study, is a standard practice in any usability study.

2.7.3 Participants

Pavuna (2019) writes that when testing a VR app, one should consider recruiting both current VR users and users that haven’t adopted the technology yet. This is because the current userbase are likely to be much more advanced and those users may encounter minor issues with more ease, leading evaluators to overlook those issues. This is endorsed by Joyce (2019) regarding non-VR learnability studies, stating it is vital to recruit participants with little to no experience using the system they’ll be testing. As a general rule, Salvatore & Christina (2008) advise to test user groups of varying backgrounds like age, gender. On the number of participants for VR usability studies, a number of approximately 23 participants is recommended (Salvatore & Christina, 2008).

2.7.4 Designing tasks

Test tasks serve the following purposes: First, prescribe what the users should try to achieve with the product, second, focus the study on certain features of the product and finally, provide a goal to assess the users' actions to. There are two types of tasks: open-ended and specific or close-ended tasks (Hertzum, 2020). Open-ended tasks are more flexible and encourage a wide range of possible responses. They invite participants to answer "with sentences, lists, and stories, giving deeper and new insights" (Farrell 2016, as cited by Dube, 2019). Open-ended tasks are more relevant when user needs of the product are not yet clearly defined. (Hertzum, 2020)

Specific or close-ended tasks focus on the exact features to be researched (Dube, 2019). Specific tasks are outlined with concrete detail to enable the user to solve the task within the context of a real-life situation. The task may be accompanied by a scenario that sets the scene for the task and makes it easier for the user to imagine the situation. For example: "A couple needs a truck that is suitable for all the furniture and belongings in their three-room apartment. Please find the total price the couple will have to pay for the truck." No matter the task type, test tasks should be short, and written in every-day language, rather than using the exact words found in the user interface. This allows the test to reveal whether the users can associate the user interface labels and buttons with the on-going task. The wording of the tasks shouldn't guide the users toward the needed actions. Making tasks available to the participants in writing is advised. (Hertzum, 2020)

Lindgaard & Chattratichart (2007) found that the number of test tasks correlate with an increased amount of previously unidentified usability problems. Though, this must be balanced with the lengthiness of the usability test. Roughly one hour is a typical length to avoid bias due to user fatigue (Hertzum, 2020). It can be rewarding to give a few users a different set of tasks than many users the same set of tasks, in other words widening task coverage (Lindgaard & Chattratichart, 2007).

2.7.5 Data collection

As earlier noted, Kaminska et al. (2022) argue the most reliable objective metrics for VR usability studies to be the user's biomedical signals, such as EEG signals and heart-rate tracking. For subjective data collection, questionnaires and interviews and think-aloud method can be used (Salvatore & Christina, 2008). A study by Othman et al. (2022) evaluated the usability of a virtual museum tour application for both VR and

non-VR versions of the application. The study mainly bases its learnability and usability conclusions on the participant's SUS (System Usability Scale) scale answers. They compare and contrast the answers based on gender, nationality, and previous usage experience of virtual tour applications. There seems to be a lack of observational findings in this study. In another study by Fussel et al. (2019) the usability of a Virtual Reality Tutorial was tested. They included a metric for learnability, which was simply users' ratings on task-specific difficulty.

Recording the screen and audio of the participants is recommended, because it reduces the need of thorough notetaking, although good notes are valuable in any case. (Hertzum, 2020). In VR usability studies, Pavuna (2019) also recommends recording the first-person view of VR headsets to be closely analysed later. Using a logging application in which notes can be inserted and buttons clicked to indicate a usability problem along with a timestamp is also advised (Hertzum, 2020).

The CUE (Components of User Experience) model separates UX to instrumental (pragmatic) and non-instrumental (hedonic) components. The meCUE 2.0 questionnaire was developed based on that model and it contains 34 Likert scale items. This questionnaire is flexible in the way that it has modules targeting the various sub-components, such as emotion, usability and usefulness and as so can be modified to the needs of different research (Minge & Thüring, 2018). In a learnability study, the usability and usefulness modules including items like "The product is easy to use", "It is quickly apparent how to use the product", and "The operating procedures of the product are simple to understand" are especially central.

2.7.6 During the test

A few evaluation protocols to use by the evaluator, such as the think-aloud method, are explained in the chapter 2.5, Methodologies.

Users should be made to feel at ease, as they might be anxious due to many reasons. Welcoming and thanking the user warmly should be the first action. Then, describing the test's objective and introducing the equipment and prototype will be done. The users should be reminded that it is the product, not them, that are being tested. They also only have to think about how they personally experience the product and shouldn't try to predict their colleagues' opinions. (Hertzum, 2020)

A well-prepared evaluator's script will help streamline the session. It includes welcoming words, explanations, tasks, and technical steps for both the evaluator and the participant.

2.7.7 Data analysis

Quantitative analysis can be done on the numerical findings extracted from the study. Statistical significance can be calculated for quantitative data, for example to validate a change in design. This can be done with methods such as finding the p-value or using the Chi-squared test. Though, quantitative analysis can produce misleading results, and qualitative data from observation is considered more reliable in usability studies (Nielsen, 2004). Quantitative data also requires a larger sample size and stricter study conditions to be more effective. (Budiu, 2017)

Qualitative analysis involves going through recorded data such as notes and recordings taken during the test and words said, actions taken, and emotions evoked by participants. It can be accompanied by thematical analysis, which is grouping occurred events in codes and themes.

A severity rating scale by Nielsen (1994) can be used to rate usability problems. The scale goes as follows:

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

The severity of a problem can be estimated with a combination of three factors:

- The frequency: How often does it occur?
- The impact: What are the ramifications of the problem when it occurs?
- The persistence: Is it learnt to deal with or avoid the first time it is encountered, or will it repeatedly disrupt the user experience?

In addition to rating the severity of the problems, they can be categorized. Goodman et al. (2009) categorized software learnability problems in four dimensions: Understanding task flow, awareness of functionality, locating functionality, and understanding functionality. A problem of understanding task flow is a situation where a user is aware that a set of actions must be taken to achieve a goal but does not know where to start.

Awareness of functionality is when the user doesn’t simply know a tool or feature exists. A problem with locating functionality is that the user is aware of a feature but cannot find access to it. Lastly, understanding functionality is when the user finds the correct feature, but is unable to operate with it.

3 VIRTUAL REALITY LEARNABILITY EVALUATION CASE STUDY

New users are having apparent difficulties in understanding how to operate the Glue platform, specifically in VR mode. This learnability study is directed to find problems arising on the first moments of usage. In this chapter, the process of the case study, performed with the previously discussed guidelines in mind, is explained.

3.1 Study design and plan

This study's main goal is to evaluate the initial learnability of the Glue VR mode. Potential learnability problems are noted and for the most severe problems, a solution will be suggested. To establish a reference point, the test will be completed by participants of varying VR experience. Due to the participants being geologically far away, the evaluation was decided to be performed remotely. While the remote factor complicates things, it was the most cost- and time-effective solution.

The participants are instructed to use their own VR headsets for the study, and to find a suitable space for the duration of the test either in their home or office. The evaluation sessions will be 1-on-1 sessions with the participant and evaluator. A video connection on a computer/phone will be set for the duration of the session. While Glue would have its own voice chat feature, not all tasks can be performed so that the Glue voice is active between the parties (for example, tasks in the private 'home' environment of the user). Using video and voice connections on non-VR devices also restricts the use of the VR headsets to just interacting with the Glue software.

The number of participants we decided to recruit was 10. Definitely a lower sample size than recommended, but with this number, the amount of work stayed within realistic limits. Two organizations provided participants for the evaluation, five participants joining from each. The other half had close to no experience of VR headsets, and the other group had at least basic experience, including trying out similar collaboration applications as Glue.

The recruitment begun by contacting the team leaders of the two organisations, explaining the scope of the study and kindly asking for five volunteers to participate in the study. Once they agreed, a briefing session with each organisation was held. After that, suitable times for the test sessions, were mapped out and agreed upon. The session

length was deemed to be maximum one hour, based on the recommendation by Hertzum (2020). The tests were conducted in a two weeks' timeframe.

Due to the remote factor, there was careful consideration on balancing the ease of setup on the participant’s end and the quality of recorded data and observation. Due to the recommended. A video connection will be used to observe the test and the first-person view from the participant’s headset is recorded for later analysis.

Having the study goal in mind, the most relevant features and interactions for users were mapped out in co-ordination with the Glue Collaboration staff, in more detail the representatives of the roles senior XR developer, head of product and chief experience officer.

- Moving around
- Rotating camera view
- Grabbing/moving objects
- Opening tablet menu
- Selecting UI items
- Entering a meeting space
- Muting and unmuting microphone
- Modifying avatar
- Sitting down on a chair
- Grabbing items
- Using/drawing on whiteboard
- Presenting files on large screen

To evaluate the learnability of these features, the following task list, presented in table 3, was created.

Table 3. Task list

Task #1	Put on your VR headset and log into your Glue account.
Task #2	Turn off your microphone.
Task #3	Change your team to “Petrus Test”.
Task #4	Go to space called “Scrum place”.
Task #5	Open your tablet menu.
Task #6	Grab the menu and move it around.
Task #7	Move to and sit down on the couch in the area. After, stand up again.
Task #8	Without physically turning your body or head, turn around 180 degrees.
Task #9	Create a whiteboard.
Task #10	Move closer to the whiteboard.
Task #11	Draw a circle and a square and triangle on it.
Task #12	After, erase only the circle.
Task #13	Clear the whole whiteboard.
Task #14	Create a sticky note and place it on the whiteboard.
Task #15	Locate the big screen in the space. Show a file in it.
Task #16	Give some time to modify avatar to their liking.

The think-aloud method will be used in this evaluation. Due to the participant’s VR view being hidden from the observer, the think-aloud method will allow the instructor to roughly determine which phase the participants are currently at.

Before the sessions, a background information form and a consent form were sent to the participants. The session begins by connecting on the video meeting and greeting the participants and thanking them for joining the study. It was explained that the study is part of the evaluator’s thesis, and the goal of the study was revealed. The participants were instructed to think aloud as much as possible during the evaluation. Before starting the test and putting on the VR headset, a quick guide on how to start recording the VR view will be shown to the participants, including guided images.

The tasks are given to the participants one-by-one. The evaluator will let the participant try out the controls and UI actions for a while, say, 2-5 minutes before assisting them, depending on the task and situation. If the evaluator assists the participant on a task, that task is denoted a failure for that participant.

After the tasks are done, or the time is up (5 minutes before the end of the session), the participant is instructed to stop the VR recording and take of their headset, after which a short interview will be given to the participant. The following questions were asked:

- How did doing basic tasks in Glue VR feel?
- Was anything difficult or inconvenient? (A situation or problem that occurred can be cited here)
- If any, which features felt well-made or convenient?
- Do you have any other feedback about Glue VR or the session overall, including my (the moderator's) performance.

After this, the participant was reminded that just two things need to be done after the session is over – sharing the screen recording of the VR view and filling the post-test questionnaire. Detailed instructions on how to accomplish these are to be emailed to the participants right after the session. At the end of the session, goodbyes will be said, and gratitude of their volunteering will be once again expressed.

3.2 Data collection and analysis

The user will be instructed to record their screen inside their VR headsets. The audio dialogue between the evaluator and the participant will also be recorded. The evaluator will make notes during the evaluation as well. Both quantitative and qualitative metrics are recorded, but a greater focus will be on the qualitative data, since the sample size is small. This means that the reliability of the quantitative data is weaker.

The quantitative metrics to be recorded are time on task, task completion rate and errors made. A custom questionnaire, consisting of learnability-oriented questions combining items from the SUS-scale and the MeCue 2.0 questionnaire, will be used to evaluate usability and learnability subjectively. It uses a 7-point Likert scale, ranging from 1 (Completely disagree. The options are presented in table 7. For qualitative data, observational notes and video/screen recordings will be used. The recordings will be examined by the evaluator after the sessions. The approximate time to complete tasks

and whether help was needed are noted. Additional task-related comments from participants and problems observed with the aid of the first-person VR screen recording were also written down.

Objective metrics such as percentage of task completion and mean times of relevant groups will be calculated. A list of learnability/usability problems will be created, and each problem will be assigned a severity value. The answers to the post-test questionnaire are analysed with respect to VR experience level.

4 EVALUATION RESULTS

In this section, the results from the evaluation sessions are revealed. Participants came from differing VR-backgrounds. One had advanced VR experience (ID 5), one joined with intermediate experience (ID 7), four had basic experience (IDs 4, 6, 8, 9) and four had none (IDs 1, 2, 3, 10). The participants self-rated their VR experience on a scale of none – basics – intermediate – advanced. In table 4, task fail percentages and VR usage experiences of participants are laid out.

Table 4. Task fail statistics and VR experience

ID	1	2	3	4	5	6	7	8	9	10
Task fail %	21	42	14	21	0	23	8	8	27	100
Moderate fails	2	3	1	1	0	3	0	0	3	0
Severe fails	1	2	1	2	0	0	1	1	0	1
VR experience	no	no	no	no	advanced	basics	intermediate	basics	basics	no
Experience with another social VR app	no	no	no	no	yes	yes	yes	yes	yes	no
Glue VR experience	no	no	no	no	yes	no	yes	no	no	no

In table 5, the times in seconds to complete a task and whether moderate help (light tint) or extensive help (dark tint) was required. NA denotes that the task was not completed for some reason, for example the user clearly finding that functionality already, or software error. In table 6, the mean times to complete the tasks without any assistance are shown. The times are from two groups of participants, one group with no VR experience and the other with at least some experience. Answers to the post-test questionnaire are portrayed in table 7.

Task #1, logging in, was removed from the analysis, as recording had to be stopped due to cybersecurity reasons during this task.

Table 5. Task-specific times per participant

ID	1	2	3	4	5	6	7	8	9
TASK #2	5	5	5	40	5	80	5	5	25
TASK #3	15	15	60	25	20	25	15	10	15
TASK #4	30	15	15	22	5	25	15	10	35
TASK #5	20	15	20	5	2	20	120	20	10
TASK #6	80	480	240	180	4	15	10	35	10
TASK #7	150	120	45	42	15	45	25	120	120
TASK #8	5	40	15	40	5	NA	5	NA	NA
TASK #9	10	15	30	30	3	15	15	15	5
TASK #10	210	90	5	5	3	NA	10	10	NA
TASK #11	110	70	60	30	10	35	30	140	45
TASK #12	10	15	5	5	5	15	10	8	10
TASK #13	5	85	5	130	10	45	15	8	20
TASK #14	180	NA	120	128	15	25	40	30	45
TASK #15	10	NA	50	10	NA	NA	NA	NA	NA

Table 6. Mean task completion times

Task	Mean task times, no VR experience (IDs 1-4)	Mean task times, basic to advanced VR Experience (IDS 5-9)
2	12,0	5,0
3	27,0	15,0
4	17,4	18,0
5	12,4	13,0
6	Failed by all	14,8
7	43,5	20,0
8	10,0	5,0
9	21,3	10,6
10	5,0	7,7
11	67,5	52,0
12	8,75	9,6
13	5,0	19,6
14	154,0	31,0

Table 7. Post-test questionnaire results

Question	Mean	SD
The product is easy to use.	5,1	1,0
It is quickly apparent how to use the product.	5,5	0,9
I consider the product extremely useful.	4,6	1,7
The operating procedures of the product are simple to understand.	5,3	0,7
I think that I would need the support of a technical person to be able to use this system.	3,3	1,8
I needed to learn a lot of things before I could get going with this system.	2,3	1,0
I found the system very cumbersome to use.	3,3	1,6
I would imagine that most people would learn to use this system very quickly.	5,1	1,1

From the overall mean scores, we seem to have positive feedback in terms of learnability. There were 8 respondents to this questionnaire, it was decided not to be sent to the participant who wasn't able to complete any tasks and one participant did not answer the questionnaire.

'The product is easy to use' item was answered 5 or more by all but one participant. An interesting fact is that the one advanced VR user (ID 5) answered 5 on this whereas three of the users with basic experience answered 6. This user could be weighing his predictions of other user's experience in this answer, or perhaps comparing to other similar VR products. None answered below 5 for the items 'It is quickly apparent how to use the product' and 'The operating procedures of the products are simple to understand'.

One user (ID 2) rated a 6 on expecting to need the support of a technical person, and a 3 on the 'The product was easy to use' item. Despite this, they rated a 6 on whether they imagine other people would learn to use the system quickly. This user ranked the lowest of all, 2, on considering the product extremely useful. ID 8 found the system the most cumbersome to use, rating it a 6. This was clear from the observational findings, seeing the user bend to awkward physical positions when using the product.

4.1 Occurred learnability problems

The usability problems closely related to learnability are listed here. They are categorized into four categories: Understanding task flow, awareness of functionality, locating functionality and understanding functionality.

Problem 1: Learning how to move seems to be difficult (*Locating functionality*)

Description: Many participants had a hard time realizing how movement happens. They mess around with controls, but the thumbstick press seems to be well hidden. Even some users with other social VR experience didn't get it.

"Doesn't work like in Altspace VR (A similar VR collaboration application)" (ID 6)

Task #: 7 | Severity: 3

Proposed solution: Change movement so that it happens by tilting the thumbstick – this way it corresponds to the movement in most games and other VR apps.

Problem 2: Users confused about the UI navigation method (*Understanding functionality*)

Description: The method to 'physically' push the UI buttons seems to confuse users. I think there are two reasons as to why: First, the underlying navigation method in both headsets works by pointing at things with a laser and pressing trigger. Second, the feedback is not clear. When moving a finger slowly towards a button the in-app menu, a 'hover' highlight is enabled. Many users seem to think that this means the button has been activated. Once they realize nothing is happening, they try to fiddle with the triggers in order to press the button, sometimes even switching hands to try to press the button. I believe this method also is more straining for users.

"Press next to continue, how do I do it?" (ID 4)

"I need getting used to not having to press trigger when pressing keys" (ID 4)

"My physical table gets in the way of my keyboard" (ID 1)

"The poking system has to be removed" (ID 5)

"Sometimes it's unclear whether my press is registered" (ID 9)

Task#: General | Severity: 3

Proposed solution: Think about adding ‘laser pointer’ controls as an option. (The ability to control UI menu and other objects from afar by pointing towards them and pressing the trigger or grab buttons)

Problem 3: The voice search microphone icon is misleading (*Locating functionality*)

Description: After joining any meeting Space, the menu stays in the Spaces-tab. When asked to turn off their microphone, the users are confused, as the option is only available in the Apps tab. They instead click on the microphone icon present in the spaces-tab, which triggers a voice-search function. Upon pressing this icon, the color change to orange causes users to mistakenly believe their mic is muted.

Task #: 2 | Severity: 2

Problem 4: Users unable to easily understand how to grab and move the UI (*Understanding functionality*)

Description: The menu needs to be grabbed from a specific angle. Some users had the right idea and pressing the correct button to grab the menu, but they failed to do it. Since this is the only way to move the menu, coupled with problem 5 (The UI objects spawn too far), users may become stuck.

Task #: General | Severity: 2

Problem 5: Users had trouble scrolling the avatar view (*Understanding functionality*)

Description: A user (ID 4) tried to scroll in the avatar view using the scroll bar, but didn’t realize they should drag from the window itself.

Task #: 16 | Severity: 2

Problem 6: Troubles going a step back in the UI menu *(Locating functionality*)

Description: A user (ID 2) got stuck for a while when they navigated to the Locations tab as they were unable to return to the main view. The apps icon (9 squares) was hard to find.

Task #: 4 | Severity: 1

Problem 7: ‘Locations’ confused with spaces (*Locating functionality*)

Description: Some users mistakenly navigated to the Locations menu, when asked to move to a meeting room. Perhaps the ‘maps’ icon is misleading.

Task #: 4 | Severity: 1

Problem 8: Advanced color selection options cause confusion (*Understanding functionality*)

Description: Some users found the color selection to be too much to take in.

“I don’t know what these HSV values are” (ID 4)

Task #: 16 | Severity: 1

Problem 9: The ‘clear whiteboard’ -icon’s purpose is misunderstood (*Understanding functionality*)

Description: The icon to clear the whole whiteboard is deemed to look like a folder, or a ‘new file’ icon, causing confusion.

“At first I thought it’s related to folders” (ID 4)

Task #: 13 | Severity: 1

Problem 10: Some individuals don’t realize the whiteboard can be grabbed (*Awareness of functionality*)

Description: The affordability of being able to move the whiteboard goes unnoticed at times, and users need to take time to adjust their position using the teleportation. To be fair, the task description did not instruct participants to do so.

Task #: 9 | Severity: 2

Problem 11: The laser-selection feature confuses users (*Undestanding task flow)*

Description: While trying to draw (by pointing towards the whiteboard from a distance), a user mistakes the laser (which appears as the grip button is held) as a way to draw on it, so the user keeps pressing the trigger and spawning a separate ‘manipulate object’-menu. The user doesn’t seem to acknowledge the menu and keeps trying. One time the menu is clipping through the whiteboard, so that only part of it is visible. This resulted into the user accidentally deleting the object via the same menu.

Task #: 11 | Severity: 1

4.2 General usability problems

Problem 12: UI objects (UI menu, Whiteboard, keyboard) spawn too far

Description: In some instances, the UI menu or keyboard is too far from the user to reach it. This caused users to sometimes have to go outside of their VR playing space in order to interact with the menu. The same happened with the whiteboard. When it is created, the whiteboard spawns far from the user. Users had to physically stretch in an uncomfortable position in order to interact with some whiteboard elements. In another case, the keyboard was partly hidden by the rails in the login space.

Task #: 1, 11, 14 | Severity: 4

Proposed solution: Calculate spawn distance to a certain length from the VR Headset, so that it can surely be reached without leaning or crouching.

Problem 13: Participants keep dropping the grabbable whiteboard tools (pen and eraser)

Description: It takes many tries for most participants to understand one has to hold the grab button in order to hold the pen.

Task #: 11 | Severity: 2

Problem 14: Accidentally triggering the clapping emote

Description: Many users, even the same user in different tasks across the session, accidentally starts clapping.

“And again I’m clapping” (ID 9)

Task #: General | Severity: 2

Problem 15: The user’s avatar is not apparent before going to the profile

Description: Many users were surprised that they had an avatar in the first place. A user who had previously created an avatar in Glue had forgotten how their avatar looks.

“I didn’t even remember my avatar looked like this” (ID 6)

Task #: 16 | Severity: 1

Problem 16: A sticky note added to the space doesn’t clearly stand out from the UI menu

Description: One user (ID 6) had to take a moment to realize that a sticky note was created after he pressed the button. The note blends into the underlying UI menu so that the newly created 3D note doesn’t clearly stand out.

Task #: 14 | Severity: 1

Problem 17

Description: Next-button too small in login screen.

“Hard to hit the next-button” (ID 5)

Task #: 1 | Severity: 3

Problem 18: While trying to grab the UI menu, people accidentally press buttons on the UI menu.

Description: A few users teleported to another space while trying to reposition the menu.

Task #: General | Severity: 1

Problem 19: The menu opens and closes in quick succession accidentally

Description: When opened from the wrist button, the press is registered too easily and users accidentally press the button two times, resulting in the menu quickly flashing.

Task #: General | Severity: 1

4.3 Differences between novices and more experienced users

As presented earlier, Participants with IDs 1-4 had no VR experience before, whereas IDs 5-9 had some VR experience, including the use of similar social applications such as AltSpace VR.

Looking at table 5, we can see that the novices had the most problems with task #6, grabbing the UI menu. Moving (task #7) caused problems for both groups. The participants did not think to press on the thumbstick to move. As ID 6 commented that ‘It doesn’t work like in Altspace”, the general movement system seems to be by tilting the joystick forward. It doesn’t seem promising if even experienced users need hints for movement – so the suggestion is to use a similar control method as most other VR games and apps.

From table 6, we can see that the mean times for tasks completed without assistance vary slightly in most tasks, excluding some greater differences like in task #14, where the mean for novices was 154 seconds vs the mean for the experienced group. The experienced group was faster in most tasks on average, but some tasks were quite even. This signals there is at least some efficiency to be gained by having more robust VR skills.

4.4 Evaluation Shortcomings

There were a few challenges that impacted the evaluation. Firstly, some participants were already logged into Glue when the test began, resulting in the first task being skipped. The participants should have been instructed to log out and log in again. The users who had to log themselves in faced difficulties in the login process, for example due to the ‘Next’ -button being hard to hit. This important issue could have been missed if all the users would have been already logged in when the test started. Additionally, one task (#15, presenting a file on the big screen) had to be forfeited for half of the users due to insufficient permissions, and there was no time to fix it between test sessions. The pilot test did not reveal this due to the fact that it was made on the evaluator’s account.

The evaluator's lack of experience with a specific VR headset caused problems for some participants, who then had to figure out solutions to headset-specific problems themselves. This added to their evaluation time and mental effort, skewing the results. Additionally, screen recording was lost from two of the evaluations, which could have provided useful insights into the participants' behavior and interactions. One of them was due to the screen recording automatically stopping when the headset is removed – the participant had to do something in their home and took off the headset for a moment, and the evaluator didn’t realize to remind them to restart the screen recording.

During the evaluation, the evaluator could have encouraged participants to talk aloud more. There were moments of silence where participants could have provided meaningful comments. In some tasks, the evaluator may have offered helped too early. Furthermore, the task order evolved during the evaluation: The camera rotating task was given before the moving around task, since commonly the users seemed to figure out camera rotation while finding out how to move. This switch may have impacted the accuracy of the results.

Nielsen (1992) notes that severity ratings from a single evaluator are too inaccurate to be trusted, indicating the importance of multiple evaluators and the need for inter-rater reliability. In addition, the sample size of 10 people was quite small, so the results must be considered carefully.

5 DISCUSSION AND CONCLUSION

In this chapter, the results of the learnability evaluation, usefulness of the collected learnability guidelines and the state of automation of learnability evaluation are discussed.

To address RQ1, asking how the learnability of a VR application can be measured, this research presents guidelines collected from related work for conducting a learnability study for a VR application. Due to a lack of research on specifically VR learnability evaluation, the guidelines were combined from learnability literature of other technologies (like flat-screen software) and general usability testing guidelines. Having the guidelines in mind, a learnability evaluation for the Glue VR platform was performed as a case study. To assess RQ2, inquiring the possibility of automating learnability evaluation without involving end-users, a literature review was made.

These guidelines helped streamline the designing and planning of the study. However, one aspect I'd like more information on is the instructions of qualitative analysis – what to look out for in the participant's behaviour in the screen and video recordings and how to connect these events to learnability. Now, I had to rely on my personal VR expertise to determine whether which events count as problematic. Also, VR evaluations conducted in a similar remote fashion as this study, where the participant has their own VR headset in their own home, is unlikely to have been studied in previous academic literature. The VR usage may cause physical damage to surroundings or people in the participants home, so perhaps a waiver for this issue should be signed by the participants, so the evaluator would not be liable for damages. Luckily, nothing like this happened in this study. It can also be argued that such a remote setup brings bias to the results. Each of the participants have a different environment, as opposed to a physical lab setup where the similarity of the environment is ensured. On the other hand, this setup is closer to the 'natural' environment of the users and helped reveal problems related to the smaller space available in home offices, such as problem 2 in chapter 4.1, where a user's physical table got in the way of the virtual keyboard.

The evaluation yielded eleven potential learnability problems and seven potential problems in other areas of usability. However, only a few severe learnability problems were found – these were related to movement controls in the virtual space and the use of the UI navigation modality. The learnability of the system was subjectively rated by the participants possessing varying VR expertise to be on the positive side.

Regarding RQ2 and the automation of learnability testing: No articles seem to exist, that claim to be able to fully automate virtual reality testing without user input. However, two articles were found that approached automating UX and usability testing from somewhat similar angles. The object of evaluation in both articles was a 3D game. The study by Fernandes et al. (2021) was based on an affective model that emulates the users' emotions, and it attempts to predict the arousal of players as they traverse through the game. The other study by Gordillo et al. (2021) involved 'crawling' through the levels with AI actors, with the intention to find exploits and unintended mechanics. This approach seems like it would work best in evaluating level design – but perhaps the affective model in the study by Fernandes et al. could be integrated to work with UI as well. Even the study by Kaminska et al. (2022) was not able to determine with suitable accuracy the usability problems using automated analysis with objective data gathered from actual users. This means that the automatic usability and learnability evaluation of any software, not to mention VR software, without any end-users, seems unfeasible at the time of writing this research.

Considering future research needs, to the best of my knowledge there is no academic literature that would provide guidelines on the procedural steps of VR learnability evaluation. This test was 1-on-1, which means a key usage situation of Glue, group meetings, was not evaluated. A testing session with a small group of people simultaneously using the platform could be valuable, possibly revealing more learnability problems. The collection of simple guidelines regarding VR usability studies by Salvatore & Christina (2008) is useful to both experts and novices of usability but is quite outdated – an updated list of guidelines with the latest VR technology would be in place. With regard to automating testing without end-users, I believe an way must be found to automate the analysis of objective usage data accurately, before moving on to predicting usability and learnability problems without any end-user input.

REFERENCES

Budiu, R. (2017). Quantitative vs. Qualitative Usability Testing. https://www.nngroup.com/articles/quant-vs-qual/

Corrêa Souza, A. C., Nunes, F. L. S., & Delamaro, M. E. (2018). An automated functional testing approach for virtual reality applications. Software Testing, Verification & Reliability, 28(8), e1690-n/a. https://doi.org/10.1002/stvr.1690

Dube, S. (2020). A Guide To Usability Testing Procedure And Creating Tasks For Successful Usability Testing. https://www.invespcro.com/blog/usability-testing-procedure-tasks/

Fernandes, P. M., Lopes, M., & Prada, R. (2021). Agents for Automated User Experience Testing. ICSTW, 247–253. https://doi.org/10.1109/ICSTW52544.2021.00049

Fussell, S. G., Derby, J. L., Smith, J. K., Shelstad, W. J., Benedict, J. D., Chaparro, B. S., Thomas, R., & Dattel, A. R. (2019). Usability Testing of a Virtual Reality Tutorial. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63(1), 2303–2307. https://doi.org/10.1177/1071181319631494

Grossman, T., Fitzmaurice, G., & Attar, R. (2009). A survey of software learnability: Metrics, methodologies and guidelines. 649–658. https://doi.org/10.1145/1518701.1518803

Hertzum, M. (2020). Usability Testing: A Practitioner's Guide to Evaluating the User Experience. Springer International Publishing.

Hillmann, C. (2021). UX for XR: User Experience Design and Strategies for Immersive Technologies. Apress L. P.

International Organization for Standardization. (2019). Ergonomics of human-system interaction. Part 210: Human-centred design for interactive systems (ISO Standard No. 9241-210:2019). https://www.iso.org/standard/77520.html

John, B., & Kieras, D. (1996). The GOMS family of user interface analysis techniques: Comparison and contrast. ACM Transactions on Computer-Human Interaction, 3(4), 320–351. https://doi.org/10.1145/235833.236054

Joyce, A. (2019). How to Measure Learnability of a User Interface. https://www.nngroup.com/articles/measure-learnability/

Kaminska, D., Zwolinski, G., & Laska-Lesniewicz, A. (2022). Usability Testing of Virtual Reality Applications—The Pilot Study. Sensors (Basel, Switzerland), 22(4), 1342. https://doi.org/10.3390/s22041342

Kao, D., Magana, A. J., & Mousas, C. (2021). Evaluating Tutorial-Based Instructions for Controllers in Virtual Reality Games. Proc. ACM Hum.-Comput. Interact., 5(CHI PLAY). https://doi.org/10.1145/3474661

Lee, S., & Sah, Y. J. (2020). Development of an Approach to Measuring Learnability Based on NGOMSL from Perspectives of Extended Learnability. International Journal of Human-Computer Interaction, 36(2), 199–209. https://doi.org/10.1080/10447318.2019.1625569

Lindgaard, G., & Chattratichart, J. (2007). Usability testing: What have we overlooked? 1415–1424. https://doi.org/10.1145/1240624.1240839

Marcus, A., & Wang, W. (2018). The MeCUE Questionnaire (2.0): Meeting Five Basic Requirements for Lean and Standardized UX Assessment (Vol. 10918, pp. 451–469). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-91797-9_33

Miller, G. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information.

Nielsen, J. (1994a). 10 Usability Heuristics for User Interface Design. https://www.nngroup.com/articles/ten-usability-heuristics/

Nielsen, J. (1994b). Severity Ratings for Usability Problems. https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/

Nielsen, J. (2004). Risks of Quantitative Studies. https://www.nngroup.com/articles/risks-of-quantitative-studies/

Ntoa, S., Margetis, G., Antona, M., & Stephanidis, C. (2021). User Experience Evaluation in Intelligent Environments: A Comprehensive Framework. Technologies (Basel), 9(2), 41. https://doi.org/10.3390/technologies9020041

Othman, M. K., Nogoibaeva, A., Leong, L. S., & Barawi, M. H. (2022). Usability evaluation of a virtual reality smartphone app for a living museum. Universal Access in the Information Society, 21(4), 995–1012. https://doi.org/10.1007/s10209-021-00820-4

Pavuna, A. (2019). A fundamental guide to user testing in Virtual Reality. https://medium.com/ostmodern/a-fundamental-guide-to-user-testing-in-virtual-reality-a71c85c764c0

Rafique, I., Jingnong Weng, Yunhong Wang, Abbasi, M. Q., Lew, P., & Xinran Wang. (2012). Evaluating software learnability: A learnability attributes model. ICSAI, 2443–2447. https://doi.org/10.1109/ICSAI.2012.6223548

Ramkumar, A., Stappers, P. J., Niessen, W. J., Adebahr, S., Schimek-Jasch, T., Nestle, U., & Song, Y. (2017). Using GOMS and NASA-TLX to Evaluate Human-Computer Interaction Process in Interactive Segmentation. International Journal of Human-Computer Interaction, 33(2), 123–134. https://doi.org/10.1080/10447318.2016.1220729

Salazar. (2022). Evaluate Interface Learnability with Cognitive Walkthroughs. https://www.nngroup.com/articles/cognitive-walkthroughs/

Salvatore, L. Christina, K. (2008). Simple Guidelines for Testing VR Applications. IntechOpen. https://doi.org/10.5772/5925

Tollmar, K., Gordillo Chaves, C. A., Gisslén, L. M., & Bergdahl, J. (2022). PLAYTESTING COVERAGE WITH CURIOSITY DRIVEN REINFORCEMENT LEARNING AGENTS.

Tullis, T. (Thomas), & Albert, B. (William). (2008). Measuring the user experience: Collecting, analyzing, and presenting usability metrics (1st edition). Elsevier/Morgan Kaufmann.