Development of tool for remote asynchronous usability and user experience testing of desktop applications

Group 1082: Asger Møller Eriksen

10th semester Engineering Psychology

amer18@student.aau.dk

Aalborg Universitet

Institut for Elektroniske Systemer

AALBORG UNIVERSITET
STUDENTERRAPPORT

Engineering Psychology
Department of Electronic Systems
Fredrik Bajers Vej 7B
9220 Aalborg

Title:
Development of tool for remote asynchronous usability and user experience testing of desktop applications

ECTS:
30 ECTS

Semester:
10. semester

Theme:
Applied Engineering Psychology

Project period:
Spring 2023 4. semester -
Master in Engineering Psychology

Projectgroup:
1082

Group members:
Asger Møller Eriksen

Supervisor:
Rodrigo Ordoñez

Company supervisor:
Jeremie Gay

Page numbers:
77

Date of Completion:
June 2, 2023

Abstract:

As technology continues to evolve, the demand for intuitive and user-friendly software experiences is also on the rise. The aim of this project was to enhance the usability of desktop applications by exploring and developing methods for asynchronous remote user testing. To accomplish this, a case study of a desktop slicing software, REALvision Pro, was selected. Two studies were conducted: study 1 used asynchronous remote usability testing (ARUT), while study 2 employed traditional in-person testing as a benchmark. Study 2 yielded significantly richer data and identified more usability issues, including high severity ones, compared to study 1. The usability issues identified in study 2 were of higher quality due to the availability of different data sources. However, study 1 encountered limitations, such as the ineffective implementation of critical-incident reports and log data collection, resulting in no findings from these methods. The discussion highlighted methodological differences between the two studies, resulting in the identification of different types of usability issues. In conclusion, while the developed tool for asynchronous remote testing offers advantages such as resource reduction, scalability, and flexibility, in-person testing provides richer data and identifies more usability issues, which ultimately outweighs these benefits.

Preface

This report has been prepared by Asger Møller Eriksen, a 4th semester student in the Master's program of Psychology Engineering (Produkt- og Designpsykologi) at Aalborg University. The project was conducted in the period from February to June 2023.

The purpose of the project is to develop a method for evaluating the usability of desktop applications remotely and asynchronously.

I would like to express my gratitude to the users who participated in this research project. Their involvement was essential, as this project would not have been possible without their contributions. I would also like to extend my appreciation to my supervisor, Rodrigo Ordoñez, for providing assistance, insights, and feedback throughout the duration of this project. Additionally, I would like to thank Create It REAL for their collaboration, assistance and provision of necessary resources.

Guide to reading

The project consists of a main report and an appendix, which can be found after the reference list. Additionally, an annex is provided as a ZIP file. An overview of the contents of the attachment can be found in Appendix C.

For referencing sources in the report, the Harvard method is used, where the sources are referenced first by last name and then by year.

The reference list includes the author, title, edition, publication year, and publisher for each reference. Internet sources are also indicated with a link.

Furthermore, the report contains figures and tables, which are numbered according to the chapter they are located in and in sequential order.

Contents

1 Introduction

In today's digital era, software applications have become an integral part of our daily lives, catering to various needs and preferences. From productivity tools and communication platforms to creative software and hobbyist applications, software plays a significant role in enhancing our personal and professional endeavors. As technology continues to evolve, the demand for intuitive and user-friendly software experiences has become increasingly paramount.

Exceptional usability is essential for software applications to stand out in a highly competitive market. A well-designed and user-friendly application not only improves the overall user experience but also boosts productivity, reduces errors, and increases user satisfaction. On the other hand, products with poor usability lead to frustration, confusion, unintended incidents and wasted time for users [Hertzum (2020)].

Usability holds the key to the success of software applications, irrespective of their domain. The International Organization for Standardization (ISO) defines usability as the *"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"*. Effectiveness measures the *"accuracy and completeness with which users achieve specified goals"*, efficiency evaluates the *"resources used in relation to the results achieved"*, and satisfaction encompasses the *"extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations"* [ISO (2018)].

Recognizing the importance of usability in software applications, this project aims to explore ways of enhancing usability as efficiently as possible. While traditional in-person usability testing has been the standard approach for a long time, there is a growing interest in remote usability testing methods due to their numerous advantages. These advantages include, among others, a reduction in resources spent, a greater geographical reach, and the creation of a more authentic test environment [Dray and Siegel (2004), Andreasen et al. (2007), Brush et al. (2004)].

Two types of remote usability testing methods exists: synchronous and asynchronous testing. Synchronous remote usability testing (SRUT) involves real-time interactions between participants and researchers, often conducted through video conferencing tools. While SRUT allows for immediate feedback and dynamic collaboration, it comes with inherent limitations, such as scheduling constraints and the need for virtual presence during the testing sessions.

In contrast asynchronous remote usability testing (ARUT) allows participants to engage with the software application at their convenience, without the need for real-time coordination or scheduling conflicts. This flexibility not only accommodates participants from different time zones but also eliminates the need for researchers to be present during the testing sessions. By leveraging technology and automated data collection, valuable insights can be gathered from users worldwide while minimizing the resources and time required to conduct the usability tests.

Furthermore, the asynchronous nature of ARUT allows for continuous data collection over an extended period. Participants can engage with the software application over several days or weeks, providing feedback as they encounter various features and scenarios. This prolonged exposure to the software application offers a comprehensive understanding of usability issues that may arise in different usage contexts, providing valuable insights for iterative design improvements. By automating the data gathering process, ARUT presents a significant opportunity to enhance the efficiency and scalability of usability evaluations.

However, it is important to acknowledge that remote usability testing also has its limitations and potential disadvantages. Addressing these challenges and developing an effective tool to reliably detect usability issues in software applications is the primary goal of this project.

1.1 Research questions

This project aims to develop a tool for evaluating the usability of desktop applications that can be deployed remotely and asynchronously. Consequently, the following research question has been formulated:

What methods of asynchronous remote user testing can be employed to effectively detect potential usability issues in a desktop application?

1.2 Case

The research in this project will be based on a 3D printing desktop software called REALvision Pro (RvP). RvP is a subscription based computer-aided manufacturing (CAM) software used for preparing 3D models for 3D printers [REAL (n.d.)]. In the context of 3D printing, this type of software is commonly referred to as a 'slicer' because it *slices* the 3D model into thin layers that represent the layers used for printing the model. RvP is developed by Create It REAL (CiR). Since there are already established slicers in the market, such as Cura, PrusaSlicer, and Simplify3D, CiR aims to offer a slicer that surpasses them in various aspects. One of these areas of focus is to provide the most user-friendly slicing software. The development team at CiR is committed to continuously improving the usability of RvP, along with other factors that contribute to a positive user experience. Therefore, RvP aligns well with the objectives of this project.

As part of the official launch of RvP, a beta program is being conducted to thoroughly test the program for any critical bugs or issues. The primary objective is to invite a large number of users to try out RvP free of charge, in exchange for their feedback and usage data. These beta testers will serve as research participants for the first study of this project.

RvP is designed as a desktop version of a Software as a Service (SaaS) platform. This implies that the program is a subscription-based application (available on a yearly or monthly basis) that is downloaded and installed on a PC, as opposed to running on a website like many other SaaS products. SaaS is a model where a software vendor develops a web-based application and handles its hosting and operation for customers to access via the Internet. Instead of purchasing the software outright, customers pay based on their usage [Bidgoli (2010)].

Literature review 2

The literature review will contain an in-depth investigation of existing methods and techniques within the field of usability testing. This is done to gain insight into relevant methodologies in relation to the aim of the project.

2.1 Usability testing

Usability testing (UT), also known as usability evaluation, is a process that involves testing the usability of a product with a representative group of users. This typically involves collecting empirical data by observing users as they interact with the product and perform think-aloud (TA). Observers take notes on noteworthy events during the test. The TA protocol, also referred to as a verbal protocol, is a popular method for evaluating the usability of a system. It encourages participants to vocalize their thoughts, experiences, actions, and emotions while engaging with the interface [Alhadreti and Mayhew (2018)]. There are two primary variants of TA: concurrent and retrospective TA. In retrospective TA, users are asked to verbalize their thoughts after completing the tasks, while in the more common concurrent TA method, users express their thoughts while performing the tasks [McDonald et al. (2012)].

A post analysis is then performed by the evaluator in order to establish a list of potential usability issues on the basis of the observed users interactions and verbalization [Clemmensen et al. (2009)]. UT is typically run on a relatively small number of participants, since multiple studies have shown that as low as five users can detect most of the usability issues [Nielsen (2000), Virzi (1992)]. UT is typically conducted within a laboratory setting, where the observation is taking place [Hartson et al. (1996)]. This means that the participant and the observer is in the same spatial location.
UT can be used to test a specific hypothesis or be employed as an iterative cycle of tests used to expose usability issues. Both share the goal of improving the usability of a product. More specifically great usability entails that the product is useful, easy to learn and use, effective and efficient, as well as satisfying for the end user [Rubin and Chisnell (2008)]. The formal method, involving formulating and testing of a hypothesis, can be an inappropriate approach to testing usability as the purpose of UT is to derive informed decisions on improvements of the product. Therefore Rubin and Chisnell (2008) recommend the more informal approach which is an iterative process that is used to identify usability deficiencies, the cause of them and how to accommodate these issues.

UT can be further divided into four types: formative, summative, validation and comparison tests [Rubin and Chisnell (2008)].

- **Formative (exploratory)** studies are typically conducted in the beginning of the design process of a product. Nielsen (1993) provides the following description of the goal of a formative test: *"... to learn which detailed aspects of the interface are good and bad, and how the design can be improved."*. Formative tests are often performed several times as part of an iterative evaluation process. Elements studied during a formative study can include: usability issues, frustrating aspects of the product, errors or mistakes commonly made and improvements between product iterations [Albert and Tullis (2013)]. Products at this stage of the development are often prototypes or mockups. Formative UT can however still be used to test developed products after deployment [Hartson and Castillo (1998)]
- **Summative (assessment)** tests are performed early or halfway through a product development cycle. The objective is to assess the users performance of actual tasks as well as the overall quality (eg. level of usability) of the product. Summative tests are also used to evaluate how well a product meets its objectives. Emphasis is more on user behaviour rather than the thought process. Therefore the product has to be somewhat functional when performing a summative test [Rubin and Chisnell (2008)].
- **Validation (verification)** tests are conducted in the later stages of a product development cycle. The intention with this type of test is to measure whether the usability of a product reaches certain benchmarks. The test can also be used to verify that previously discovered usability issues has been resolved [Rubin and Chisnell (2008)].
- **Comparison** tests can be deployed at any point within the product development cycle in conjunction with any of the previous three methods. As the name suggests this type of study focuses on comparing different interfaces or components within a product. Comparison between similar types of products can be performed, in order to test which is preferred [Rubin and Chisnell (2008)].

2.1.1 Remote usability testing

Remote usability testing (RUT) distinguishes from conventional UT in terms of a spatial (and sometimes temporal) separation of the evaluator and the user, hence the addition of 'remote'. RUT is defined by Hartson et al. (1996) as:

> *"... usability evaluation wherein the evaluator, performing observation and analysis, is separated in space and/or time from the user."*

If the evaluator is separated only in space but not in time from the user, the test is synchronous. This means that the evaluator can observe and communicate with the user as well as receive data in real time. If the user and evaluator is separated both in space and time the method is asynchronous. Asynchronous remote usability testing (ARUT) is limited to data gathering methods like user logs, activity diaries, self-reports and automated clickstream collection [Dray and Siegel (2004)].
With the remote element of RUT compared to conventional UT a range of different advantages as well as disadvantages occur.

2.1.1.1 Advantages of remote usability testing

Compared to conventional UT, RUT can offer several advantages. First, it can significantly reduce the cost of the study by eliminating the need for transportation and physical facilities [Dray and Siegel (2004)]. This also makes it easier to access and recruit geographically dispersed users from around the world, resulting in a larger and more diverse participant pool [Andreasen et al. (2007)].

According to Brush et al. (2004), RUT allows participants to test the system in their usual environment, providing a more authentic evaluation of the interface. They also compared UT and SRUT methods and found that participants preferred the convenience of SRUT. Additionally, results from Sauer et al. (2019) suggest that the presence of an observer can negatively influence task difficulty. Therefore, it can be assumed that RUT, and even more so ARUT, which involves limited contact between observer and participant, may further reduce this influence.

2.1.1.2 Limitations of remote usability testing

One notable disadvantage of RUT is that it generally detects fewer usability issues [Bruun et al. (2009)]. However, SRUT has often proven to be just as effective as traditional in-person UT [Brush et al. (2004), Andreasen et al. (2007), Bruun et al. (2009)]. It is worth noting that results vary as some studies have shown that ARUT can outperform SRUT in terms of the number of usability issues identified [Alhadreti and Mayhew (2018)].

The results regarding the time spent on SRUT studies vary. Some studies show that the time spent is comparable to in-person UT [Andreasen et al. (2007)], while others claim that it increases the time, particularly in the data analysis phase [Olmsted and Gill (2005)]. This will naturally depend on the type and scale of the study and analysis conducted. On the other hand, ARUT aims to shift more of the effort from the evaluator to the user, potentially reducing the amount of time spent by evaluator. This will however typically increase the time spent by participants, as they have to fill out questionnaires, critical incident reports etc. after task completion [Andreasen et al. (2007), Bruun et al. (2009)].

A limitation of remote testing methods is the reduced ability to communicate with participants, which can lead to a decrease in the amount of qualitative information obtained. For instance, in a study conducted by Olmsted and Gill (2005), when participants dropped out of the test, the researchers were unable to gather detailed explanations regarding the reasons for the participants' decision to abandon the study. Furthermore, remote testing makes it difficult to interpret contextual information and nonverbal cues, resulting in a further reduction of qualitative data [Dray and Siegel (2004)].

Additionally, when conducting remote studies, the control over environmental distractions is significantly reduced, which can potentially affect participant performance [Sauer et al. (2019)].

2.1.1.3 Data gathering methods for remote usability testing

In this section some of the most prevalent methods for gathering data during RUT are described.

Observations

When performing RUT, the data collection methods of in-person UT can still be applied. For SRUT, where the participant and observer are physically separated, an observation requires the setup of a live screen sharing tool. Additionally, the TA protocol requires live audio communication to take place [Hertzum (2020)]. On the other hand, in ARUT, observation and TA are not transmitted in real-time. As the observer and participant are also temporally separated, it is not possible for the observer to prompt the participant for verbalization. However findings suggests that the use of such unmoderated TA sessions do not affect the results [Hertzum et al. (2015)].

Questionnaires

Questionnaires can either be administered after each task or when the user is done interacting with the product. The former method is used to compare the difficulty of each task while the latter can be used to create an overall measure of the perceived usability of a product. Both approaches are useful when attempting to measure the impact of new product iterations. The primary aim of questionnaires is to collect participants' preference data as well as the difficulty of performing tasks, which can clarify the strengths and weaknesses of the product. This can help identify which areas of the product that need improvement. One of the most prominent ways of constructing a questionnaire, is by using rating scales. Rating scales often measure parameters like the ease of use, usefulness and satisfaction of using a product [Albert and Tullis (2013)].

Open-ended questions, where respondents are allowed to add verbatim comments as a response, are a way of gathering more detailed information on how to improve a product. According to Albert and Tullis (2013) this is the best method for understanding possible usability issues when performing automated (ARUT) usability studies. Furthermore, they recommend applying this method in conjunction with rating scales, conditionally prompting participants to provide a verbatim comment only when they provide a low rating. However, a downside to this combination is that participants may adjust their ratings to avoid subsequent open-ended questions. [Albert and Tullis (2013)].

When administering questionnaires to multiple participants, it is important that each question is asked in exactly the same way [Rubin and Chisnell (2008)]. Furthermore, research has demonstrated that standardized usability questionnaires generally offer a more reliable assessment of usability, in contrast to custom-made questionnaires [Hornbaek (2006)]. Below are some of the most prominent standardized usability questionnaires:

- The **System Usability Scale (SUS)** is a widely used questionnaire designed to assess the usability of a system or product [Albert and Tullis (2013)]. It consists of a set of ten statements that participants rate on a 5-point Likert scale, reflecting their agreement or disagreement with each statement. The SUS provides a quick and reliable measure of perceived usability, capturing aspects such as ease of use, learnability, and user satisfaction [Brooke et al. (1996)]. Its simplicity and effectiveness make it a popular tool for evaluating and comparing the usability of different systems [Peres et al. (2013)].
- **Post-Study System Usability Questionnaire (PSSUQ)** is a questionnaire used to measure a user's perceived satisfaction with a computer system or application. The latest version consists of 16 statements that participants rate on a 7-point likert scale (plus a 'NA' response option). The items can be divided into three sub-scales; *System Usefulness* (6 items), *Information Quality* (6 items) and *Interface Quality* (3 items). The last item in the questionnaire concerns the overall satisfaction with the system in question. Each item go from 'Strongly Agree' (1) to 'Strongly Disagree' (7) with no labels in between [Sauro and Lewis (2012)].
- The **Software Usability Measurement Inventory (SUMI)** is a tool used to assess the usability of software applications. It contains 50 3-point Likert-scale questions that cover five dimensions; Efficiency, Affect, Helpfulness, Control, and Learnability. The SUMI helps identify areas for improvement. It is commonly used in research and industry to evaluate and enhance the overall user experience of software products [Sauro and Lewis (2012)].
- The **Questionnaire for User Interaction Satisfaction (QUIS)** is a questionnaire used to evaluate users' subjective satisfaction with specific aspects of the human–computer interface. It comes in different lengths, but the short version comprises 27 9-point bipolar Likert-scale questions that assess five aspects of the user's experience; Overall Reaction, Screen, Terminology/System Information, Learning and System Capabilities [Chin et al. (1988)]. With its comprehensive range of dimensions, the QUIS provides valuable insights into the user's perception of the system's interaction quality. The questionnaire is however not free to use [Sauro and Lewis (2012)].

Log data

Log data is an indirect feedback data gathering method that automatically collects detailed user activity logs of a system. This method can be used to provide insight into the usage pattern and behavior, as well as the overall performance of a system. Measurements from this method typically consists of individual keystrokes (eg., menu selection or navigation) with appurtenant timestamps and user ID [Millen (1999)]. This information can then be used to construct the individual user flows, how successful they were, how long they took in performing a task and where potential errors or mistakes appear. Furthermore time and performance measures can be used when comparing future designs [Scholtz (1999)] or certain benchmarks. Drop off rates are often recorded as part of the log data, which shows where in the process a user stops their interaction and quits the software/website Albert and Tullis (2013). This metric can reveal where in the chain of interactions an issue or deficiency is present.

User-reported critical incidents

User-reported critical incidents is a method that tries to capture usability issues, by enabling users to report the issues in a structured form based on several questions. According to Hartson and Castillo (1998) a critical incident in the context of formative usability testing is defined as: "an occurrence during user task performance that indicates something (positive or negative) about usability". Below is a list of things that can be included in a critical incident report [Castillo et al. (1998)]:

- URL (in the case of a webpage)
- The task being performed when incident occurred
- The users expectations of what was supposed to happen when incident occurred
- A detailed description of the incident and why the user thought it happened
- A description of how the user recovered from the incident, if it was possible
- A description of how to reproduce the incident, if possible
- A rating of the incidents severity
- Further comments, suggestions or possible solutions to the incident

Incidents are typically reported by pressing a button on an interface which opens a separate window where users can enter a report about the incident [Castillo et al. (1998)]. However Hartson and Castillo (1998) showed that the majority of participants preferred that the report tool was an integrated part of the application being evaluated.

Forum

Forum is a method where participants collaborate on reporting usability issues on a shared online platform. Bruun et al. (2009) tested this method where participants were instructed to first check if the experienced usability problem was already reported by another user. In case they could not find a similar issue, they were asked to create a description of the problem together with a rating of severity. Finally if a similar issue was already reported by another user, they were asked to either post an agreement or a disagreement (and reason) with the issue description.

Study 1: Asynchronous remote usability test 3

The following chapter contains a description of the first usability evaluation, as well as a presentation and discussion of the results.

3.1 Aim of study

As described in Chapter 1.1, the overarching aim of this project is to develop an asynchronous remote usability testing method for effectively evaluating the usability of desktop applications. The purpose of the tool is to assess the usability level of REALvision Pro (RvP), identify usability deficiencies, critical issues, and bugs, and utilize this information to enhance the software's usability. Therefore, the objective of this study is to identify and assess the severity of existing usability issues within RvP. The following research question has been formulated for this study:

Which usability issues exist in RvP and how severe are they?

3.2 Methods

This study aims to detect currently existing usability issues within RvP. Therefore the study will be taking a exploratory approach. Moreover this type of usability test is intended to be continuously repeated as part of an iterative development process with the purpose of eliminating as many usability deficiencies as possible.

Typically, in usability evaluation studies, participants are assigned specific tasks to complete on a system. This approach helps control variables and provides a basis for comparison between subjects or products. However, in this study, since there are no preconceived notions about where usability issues may arise, participants will be encouraged to freely explore and use the software as they wish. By allowing users to pursue their own tasks, it is expected that their willingness to participate in the study will increase. This could also result in more areas and functions of the software being explored and tested. Consequently, this approach may uncover usability issues that might not have been discovered if participants were limited to a set of predetermined tasks. Additionally, this format enables participants to extensively use RvP, both for personal and professional purposes, rather than relying on other slicers. The use of uncontrolled tasks is possible because RvP is fully developed and functional, allowing users to engage with the software in a natural and unrestricted manner.

Since the evaluation will be performed asynchronously, data will have to be collected automatically. Therefore, the following three methods of data gathering have been selected: Questionnaire, User-reported critical incidents, and Log data.

3.2.1 Questionnaire

Questionnaires (open-ended questions specifically) are an effective way of collecting detailed information about potential usability issues, when data is collected remotely and automatically [Albert and Tullis (2013)]. Therefore this method has been selected as the primary method for data collection. Version 3 of the Post-Study System Usability Questionnaire has been selected because of its low number of items, free availability, high reliability and relevance in context of RvP [Sauro and Lewis (2012)]. Contrary to the best alternative usability questionnaire, the System Usability Scale (SUS), PSSUQ also specifically examines the interface and information of a system. In order to collect the most relevant data, several changes have been made to the original PSSUQ.

The PSSUQ is exclusively a rating scale that results in a series of numerical scores. However, as the primary intention is to have participants provide detailed information about the causes of bad ratings (i.e., usability issues), open-ended questions have been added to each individual rating scale. Open-ended questions will only be triggered if a respondent provides a score of 3 or higher, indicating a level of dissatisfaction. This conditional approach is used to reduce the amount of time spent completing the questionnaire by the respondent. Since a score of 1-2 is assumed to indicate satisfaction, it is assumed that respondents will not have any critical issues to report. Furthermore, items 5, 6, 9, and 12 can lead to two different open-ended questions, depending on the given rating. For example, the open-ended question for item 5 will end with *"less comfortable to use"* when the item receives a rating of 3-4 (slightly agree - neither agree nor disagree). However, if the item is scored 5 or higher (slightly disagree - strongly disagree), the subsequent open-ended question will end with *"uncomfortable to use"*. The same difference in phrasing applies to items 6, 9, and 12 as well. Asking respondents to provide reasoning behind their ratings is similar to the approach originally described in the PSSUQ instructions by Lewis (1995), where respondents are encouraged to add elaborating comments to each answer.

The PSSUQ is already fairly long with 16 rating scales and with the addition of an open-ended question to each of those 16 scales the questionnaire would contain a total of 32 questions. In order to decrease the amount of time it takes to complete the questionnaire, in turn increasing the chance of full completions, all items (except one) from the *Information Quality* sub-scale were removed. These

items were removed based on the assumption that items from the other subscales were more relevant for the aim of the study. According to Sauro and Lewis (2012) the practice of editing the PSSUQ by adding and removing items, to a limited extent, is a viable option. The one item from the *Information Quality* subscale that was kept, relates to the users ability to recover from mistakes (Item 8 in the revised version of PSSUQ). This item is kept as it is very relevant to the usage of RvP.
One of the items from the second iteration of PSSUQ has been included as it can help uncover usability issues related to the effectiveness of RvP (Item 3 in the revised version of PSSUQ) [Lewis (1995)].
Finally, several terminological changes have been made. Firstly, 'the' (right before 'tasks') from item 3 and 4 has been removed since there are no predetermined tasks that users are asked to perform. Furthermore, 'and scenarios' has also been removed from item 3 and 4, as 'tasks' is the appropriate term. Lastly, 'the/this system' has been replaced with 'REALvision Pro', as this removes any uncertainty about which system is being referred to. The revised version of PSSUQ is presented below:

1. **Overall, I am satisfied with how easy it is to use REALvision Pro.**
   a) *Please describe which aspects of REALvision Pro were the least easy to use.*

2. **It was simple to use REALvision Pro.**
   a) *Please describe which aspects of REALvision Pro were the least simple to use*

3. **I was able to effectively complete tasks using REALvision Pro.**
   a) *Please describe which aspects of REALvision Pro that made completing tasks less effective*

4. **I was able to complete tasks quickly using REALvision Pro.**
   a) *Please describe which aspects of REALvision Pro that could have been quicker*

5. **I felt comfortable using REALvision Pro.**
   a) *Please describe which aspects of REALvision Pro were less comfortable to use*
   b) *Please describe which aspects of REALvision Pro were uncomfortable to use*

6. **It was easy to learn to use REALvision Pro.**
   a) *Please describe which aspects of REALvision Pro were less easy to learn*
   b) *Please describe which aspects of REALvision Pro were difficult to learn*

7. **I believe I could become productive quickly using REALvision Pro.**

   a) *Please describe which aspects of REALvision Pro that hinder you in becoming productive quickly*

8. **Whenever I made a mistake using REALvision Pro, I could recover easily and quickly.**
   a) *Please describe why you were unable to recover from mistakes easily and quickly*

9. **The interface of REALvision Pro was pleasant.**
   a) *Please describe which aspects of the interface were less pleasant*
   b) *Please describe which aspects of the interface were unpleasant*

10. **I liked using the interface of REALvision Pro.**
   a) *Please describe which aspects of the interface you didn't like*

11. **REALvision Pro has all the functions and capabilities I expect it to have.**
   a) *Please describe which functions and capabilities REALvision Pro lacked*

12. **Overall, I am satisfied with REALvision Pro.**
   a) *Please describe which aspects of REALvision Pro were less satisfying*
   b) *Please describe which aspects of REALvision Pro were unsatisfying*

It is required to provide a rating for each item in order to proceed with the questionnaire. However, respondents are able to skip the comment sections if they do not want to or have no information to provide.

3.2.2 User-reported critical incidents

The second method for gathering usability data is a form where users can report critical incidents that occur when using RvP. In this study, the form will be used to collect information about bugs and issues that users may experience. The form contains the following questions:

1. **What happened?**
2. **What did you expect to happen?**
3. **Describe how one can reproduce the bug/issue with a list of steps**
4. **Additional comments or suggestions?**
5. **Upload files (screenshots, workspace files, screen recordings, etc.**

It is required to answer the first question in order to submit a report. The full form can be seen in Appendix A.1. Users can submit as many reports as they wish. It was not possible to integrate the report form as a part of the software in this study; it was developed using Google Forms.

3.2.3 Log data

During the beta session, log data was collected using Mixpanel, an online service that enables companies to track user behavior and retention measures on SaaS or mobile platforms [Mixpanel (n.d.)]. The purpose of tracking log data is to analyze general behavior patterns, identify unreported bugs, and locate potential usability issues through user flows. One way to approach the latter could be to examine the steps prior to high drop-off rates. Additionally, log data will be used to support reported usability issues.

3.3 Participants

As mentioned in Chapter 1.2, the study will be performed on a group of beta testers. The use of beta users is a widely used method for involving users in the development process of a product [Millen (1999)]. Participants for UT has to be representative of the target audience [Rubin and Chisnell (2008)]. Therefore previous or current users of REALvision Online was contacted with a recruitment email in order to ensure that participants have an interest and likely experience with 3D printing and slicers, thus achieving the desired target group of 3D printing users.

A total of 32 users signed up for the beta program. However, only 10 unique users ended up using RvP at least once, making these 10 participants the only participating beta testers who provided data for this study. The demographic information collected from the 10 participants is presented below. The data was collected when signing up for the beta program (see demographic questionnaire in Appendix A.2).

User ID	Age (years)	3D printing experience	Frequency of slicer usage	Usage context (Personal or professional)
1	36-45	$>$ 2 years	Daily	Both
2	36-45	$>$ 2 years	Daily	-
3	26-35	$>$ 2 years	2-4 times a week	Personal
4	36-45	1-2 years	Daily	Personal
5	36-45	$>$ 2 years	2-4 times a week	Personal
6	56 or older	3-12 months	Daily	Personal
7	56 or older	3-12 months	2-4 times a week	Personal
8	56 or older	$>$ 2 years	2-4 times a week	Personal
9	46-55	0-3 months	Daily	Both
10	36-45	0-3 months	2-4 times a week	Personal

Table 3.1. The table contains demographic information about the beta testers. An error occurred for user 2 who was able to sign up without providing usage context information.

3.4 Equipment

The following equipment and material was distributed to the participants:

- REALvision Pro 4.1.0
- Online questionnaire (SurveyXact) described in chapter 3.2.1

3.5 Procedure

Between March 17th and March 21st, recruitment emails were sent to 5,932 people who had signed up for REALvision Online.

Upon entering the beta program landing page, users were prompted to complete a set of demographic questions (see the list in appendix A.2). Once users submitted their request to join the beta program, their requests would be evaluated and either accepted or rejected. Rejections occurred if users were under 18 years old or did not have access to a 3D printer. Participants who met the criteria for acceptance received an introductory email containing instructions for installing RvP, guidelines for reporting bugs and issues, and encouragement to provide feedback via the support email (see email in appendix A.3). Users who were invited to participate in the beta program were informed that, in exchange for access to RvP, they would be required to share their usage data and provide feedback on the product.

Between March 20th and March 28th a total of 10 unique users launched and used RvP. The questionnaire was distributed on March 28th, after two days of no activity from any of the beta testers.

3.6 Results

No critical incidents were reported.

Upon analyzing the log data it became clear that a lot of tracking points were missing in order to be able to analyze detailed user flows. Missing tracking points include object manipulation functions (movement, scaling, rotation, flipping and support generation), changes in printer and material settings and opening of user manual and help center. Furthermore more participants are needed in order to determine general behaviour patterns. Log data was however used to confirm that no crashes had occurred.

Four of the 10 beta testers responded to the usability survey. All four completed the survey. Results from the questionnaire consists of rating scale scores and verbatim comments related to each item. The identity of the respondents of the questionnaire are labelled 'R1', 'R2', 'R3' and 'R4'. Note that there is no connection between user IDs from table 3.1 and the ID assigned to respondents of the questionnaire in the following chapters.

3.6.1 Rating scale scores

Since only four participants have rated the items, any conclusions based on these scores should not be regarded as definitive

Due to the removal of five items from the *Information Quality* subscale, a score for this subscale cannot be calculated. Additionally, as item 3 is not part of the original PSSUQ, its score is not included in the calculation of the mean *System Usefulness* subscale score.

To interpret the scores, a set of reference data points is required. As described in Sauro and Lewis (2012), when interpreting rating scores, the best reference point is one's own data from previous tests. However, if no previous scores are available, the next best reference is the norms of PSSUQ responses collected from other studies. As there is no existing reference data for RvP, the optimal reference point is the data collected by Sauro and Lewis (2012) based on 21 studies and 210 participants. All means from this set of normative data fall below 4, which implies that 4 should not be considered an average score, but rather the threshold between a positive and negative attitude towards an item. Table 3.2 presents scores from the questionnaire along with means from the data collected by Sauro and Lewis (2012).

	Data from the present study		**Normative data [Sauro and Lewis (2012)]**	
Item	Mean (items)	Mean (subscales)	Mean (items)	Mean (subscales)
1	3.5	3.42	2.85	2.80
2	3.25		2.69	
3 (4)	3		3.16	
4 (5)	3.75		2.66	
5 (6)	3.75		2.27	
6 (7)	3.25		2.86	
7	-	-	3.7	3.02
8	2.25		3.21	
9	-		2.96	
10	-		3.09	
11	-		2.74	
12	-		2.66	
13 (9)	3.75	4.08	2.28	2.49
14 (10)	4.5		2.42	
15 (11)	4		2.79	
16 (12)	3.5	-	2.82	-

Table 3.2. The table presents means of rating scale scores from the questionnaire in the present study and data gathered by Sauro and Lewis (2012). Both individual item scores and subscale scores are presented. Green cells indicate the lowest (thus more positive) mean between the two sets of data. Note that item 3 from the questionnaire used in this study is excluded from the table as it is not a part of PSSUQ version 3. As the questionnaire used in this study was modified, the item numbers do not correspond with the item numbers of the original PSSUQ version 3. Therefore the corresponding item numbers from the questionnaire used in this study is presented in parenthesis next to the item numbers from the original PSSUQ version 3.

Item 3 *(I was able to effectively complete tasks using REALvision Pro)* from the questionnaire used in the present study received a mean rating of 3.75.

Notably only two items were rated better (lower) when comparing means with the reference data. Firstly item 8 (*Whenever I made a mistake using REALvision Pro, I could recover easily and quickly*) was rated lower (2.25) than the average score (3.21). Secondly item 3 (item 4 from the questionnaire used in the present study) (*I was able to complete tasks quickly using REALvision Pro*) was also rated lower (3.0) compared to the reference data (3.16), by a small margin. The rest of the items received a higher average score than the reference data, indicating less satisfaction with RvP than the average system.

The mean rating of the subscales are also rated lower compared to the reference data. The *Interface Quality* subscale has the lowest rating of 4.08, which is above the midpoint of 4.

In the line chart below, the responses for each respondent are presented.

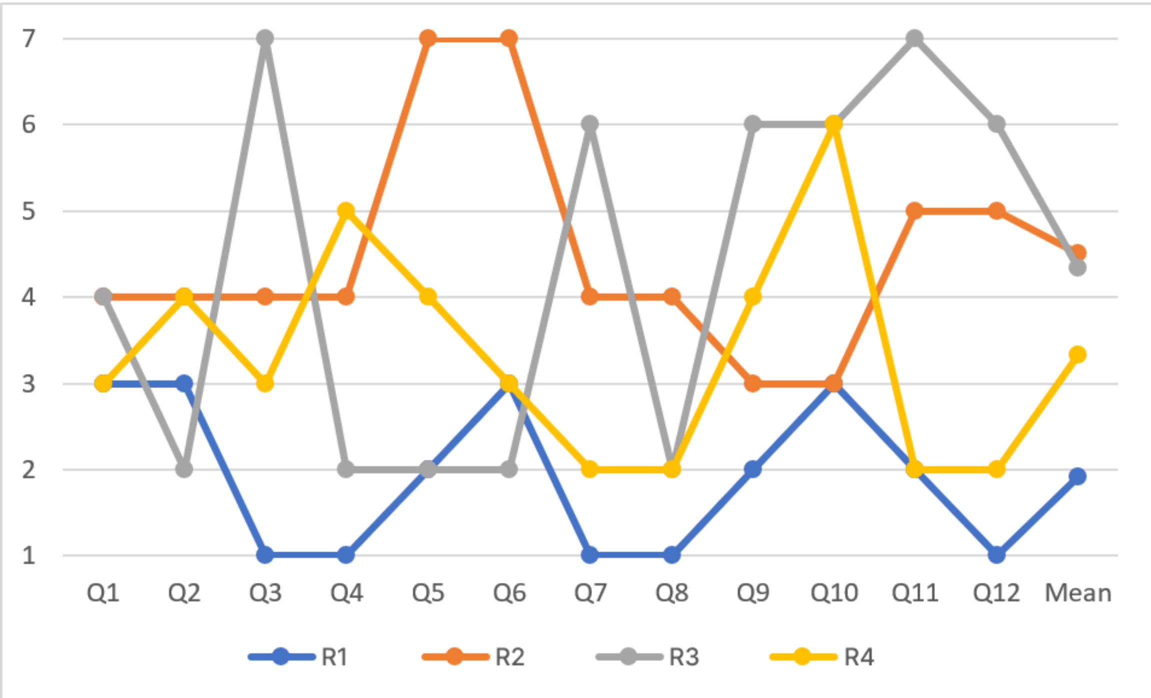


Figure 3.1. The figure presents a line chart of the item scores for the individual respondents. The respondents mean ratings (across all items) are the following: R1: 1.92, R2: 4.5, R3: 4.33, R4: 3.33.

Looking at Figure 3.1, it is evident that R1 has shown the most positive response towards the system, as none of the item ratings exceed 3. This respondent has the lowest mean rating of 1.92. In contrast, R2 rates only two items below the neutral midpoint of the scale, while scoring four items above 4, indicating a more negative attitude towards the system. Additionally, R2 has the highest mean rating of 4.5. R3's ratings exhibit more variation, with an average rating of 4.33, leaning slightly towards the negative side. R4 rated only two items above 4, while giving seven items ratings below 4. R4 had a positive mean rating of 3.33.

3.6.2 Usability issues

Based on the open-ended question responses a total of 10 usability issues were identified. Furthermore 8 feature requests and 7 comments regarding the interface were recognized. These are listed below followed by the respondent and question ID in parenthesis. All verbatim responses can be found in the annex within the 'Study1' folder.

Usability issues

1. The lack of choice in advanced settings, particularly on print speeds made completing tasks less effective (R4 (Q3.a))
2. Was not able to reduce the amount of spent filament enough (R2 (Q8.a))
3. Difficulty changing layer height (user did not know how to apply custom layer height) (R4 (Q3.a))
4. Did not understand the program well (R2 (Q10.a))
5. Spent a lot of time trying to figure out how to upload STL file (user was looking under 'File' in the top tab) (R2 (Q2.a))
6. Confused about link between contour and infill, but especially retraction (R4 (Q1.a))
7. Does not understand the value of REALview (R4 (Q2.a))
8. The arrangement of information made completing tasks less effective (R4 (Q3.a))
9. Graphical user interface was not comfortable to use (R4 (Q5.a))
10. Found some printer settings well hidden and in unexpected location (R1 (Q1.a))

Requested features

1. Need free YouTube tutorials covering basics (R2 (10 times))
2. User wants better visualization of supports and overhangs (R4 (Q4.a))
3. User wants "A LOT" more options for fine tuning slicing and printing (R3 (Q3.a))
4. User wants to be able to select and fine tune support structures (Tree or Organic) (R3 (Q11.a))
5. "Place an icon of the nozzle at the level of the printed line to make the simulation even more meaningful" (R4 (Q4.a))
6. User wants to be able to add material to an existing object (R4 (Q6.a))
7. User wants to be able to send start G-code commands tailored to the specific filament types (R3 (Q11.a))
8. Too few options to tailor the interface (R3 (Q9.b))

Interface comments

1. Too much of the same tone (R4 (Q9.a))
2. Too white (difficulty seeing) (R4 (Q9.a))
3. Lack of separation between functions (R4 (Q9.a))

4. Color scheme is a bit bland (R1 (Q10.a))
5. Opt for dark mode or light mode with standard windows colors (R1 (Q10.a))
6. Too limited (R3 (Q10.a))
7. Wants more advanced options at hand, "rather than having to dig deep into menus to find them" (R3 (Q1.a))

3.6.2.1 Severity rating

To determine the most important and urgent usability issues, a severity rating will be assigned to each identified issue based on its impact on the user experience. The chosen severity rating system is single-dimensional, consisting of three levels: low, medium, and high. Low severity issues may cause annoyance or frustration but do not lead to task failure. They have a small impact on efficiency and satisfaction. Medium severity issues lead to difficulty in completing a task but do not result in task failure either. They have a moderate impact on effectiveness and often efficiency and satisfaction as well. High severity issues, on the other hand, result in task failure and are therefore given the highest priority. They greatly affect effectiveness, efficiency, and satisfaction [Albert and Tullis (2013)]. When performing a usability evaluation study where participants are not given specific tasks to complete, like in the present study, it is difficult to determine whether a usability issue leads to task failure or not. Therefore, it is also challenging to determine if a usability issue is of high severity or not. The decision to use a single-dimensional severity rating system was primarily due to the low rate of respondents, as severity rating systems with multiple dimensions often include the frequency of reported usability issues. Assigning ratings based on frequency could potentially skew the severity of issues, especially when none of the usability issues are mentioned by more than one respondent.

High severity

It is uncertain whether usability issue 2 directly led to task failure as it is unclear whether the user was unable to complete the task due to the inability to reach the desired amount of filament spent or for other reasons. Additionally, the same respondent reported quitting the program before completing any tasks due to frustration with the lack of instructions (R2 response to Q3.a). As a result, it was determined that none of the usability issues were of high severity.

Medium severity

The following four usability issues were categorized as medium severity. Firstly, issue 1 as settings were missing in order for the user to accomplish the desired outcome. This had a negative impact on the effectiveness of the task.

Issue 2 is also assigned medium severity as the user was not able to reduce the amount of filament enough to achieve the desired outcome. This issue negatively affected the effectiveness and efficiency of the task.

Issue 3 made it difficult for the participant to specify the desired layer height. The participant was unaware that it was possible to apply a custom layer height, as the same respondent requested this option (R4 response to Q4.a). This issue might have hindered the user from achieving the desired layer height value, thus negatively impacting the effectiveness of the task.

Issue 4 is unspecific as it refers to the program as a whole. It is the same respondent that requested free YouTube tutorials in order to learn how to use the software (Feature request 1). The issue has had a negative effect on efficiency, satisfaction and likely also effectiveness. The issue also indicates a very low level of learnability for this specific user.

Low severity

The remaining usability issues were given a low severity rating due to their low impact on task completion difficulty. Most of these issues are related to confusing terminology or navigation

Issue 5 caused frustration for the respondent as they spent a lot of time locating the 'add model' button. None of the other respondents reported experiencing difficulty uploading a file. Additionally, this issue is not likely to recur as the respondent probably remembers how to upload a file in subsequent attempts. The issue reduced efficiency and satisfaction, but did not affect effectiveness since the end result was unaffected.

Issue 6 resulted in confusion regarding the meaning and link between several technical terms. The issue most likely had a negative effect only on efficiency.

Issue 7 relates to confusion about the value of a feature called 'REALview'. It is uncertain whether this issue has had a negative effect on effectiveness, efficiency, or satisfaction. However, it clearly indicates a lack of communication regarding the functionality of the feature.

Issue 8 points out that the arrangement of information negatively affects the effectiveness of the task. It is not clear whether the issue has actually had an effect on effectiveness or whether the respondent is referring to efficiency.

Unfortunately, Issue 9 is unspecific regarding which aspects and components of the interface the respondent was not comfortable using. The issue has likely had a negative effect on both efficiency and satisfaction.

Finally, Issue 10 is categorized as confusing navigation, as the respondent reports that settings are located in unexpected and hidden locations. This issue has had a negative effect on efficiency, as the respondent has spent time locating the settings.

3.7 Discussion

Multiple of the reported usability issues were general and did not include specific examples or suggestions (like issue 5: "Did not understand program well" and 8: "The arrangement of information made completing tasks less effective"). Therefore the study could have benefited from a debriefing interview with the participants, in order to clarify some of the reported comments. This would have been difficult to implement due to the study being asynchronous, but could have provided valuable information. This approach is also a part of the procedure in the original instructions on how to use PSSUQ (which was administered in person) [Lewis (1995)].

As no critical bugs or issues were reported, this could indicate either that no bugs or issues were experienced, the purpose of the report was unclear or that participants did not know how to report them or did not want to. When looking at the log data, no program crashes were registered, which is the most critical type of bug. This is however also the only type of bug that can be registered as a dedicated event. It is unknown if any other bugs have occurred during the beta test session.

The Bug/issue report could also be expanded to a more general critical incident report feature, which includes all types of incidents that the users might experience. By doing this, possible uncertainties related to the terms 'bug' and 'issue' might be eliminated.

Optimally the critical incident report should be integrated within the program, so users have the shortest possible path when reporting an issue. This has also shown to be preferred by participants [Hartson and Castillo (1998)]. Furthermore users were only made aware of the location of the report once (hyperlink to google forms), in the welcome email they received when registering for the beta program. This could result in participants forgetting about the existence of the report or how to access it.

If the critical incident report was integrated within RvP, an introduction window explaining when and how to report a critical incident could be added as well. This would likely increase the likeliness of users reporting incidents.

The questionnaire was also administered separately from RvP by providing the beta testers with a link to SurveyXact via email. The questionnaire was distributed two days after the last beta tester had been active. This delay between the distribution of the questionnaire and activity from the beta testers could be problematic, as respondents' experiences, emotions, and memory of encountered issues with RvP may no longer be fresh in their minds. Thus, the data might contain fewer details due to this delay. The solution to this issue would be to integrate and display the questionnaire within RvP when specific events occur. Events that would trigger the questionnaire could be finishing actions like exportation of G-code or exiting the program.

Just like the incident report, integrating the questionnaire into RvP would likely also facilitate a higher participation rate, as participants would be presented with the questionnaire instead of having to check their email and then having the option to decide whether or not to open the link to the questionnaire. Integrating a questionnaire or a critical incident report within a piece of software does however require a substantial amount of resources from programmers.

To gather more information about usability, one approach is to include two specific items in the questionnaire: "What is your preferred slicing software?" and "I would prefer to use this slicer instead of the one I am currently using." The first item should be presented at the beginning, while the latter should be placed at the end of the questionnaire. If respondents rate the second item with a score of 3 or higher (indicating disagreement), they will be asked to provide the reasons behind their response, similar to the other items. These additional items will help collect valuable insights into what features or aspects other slicers may be performing better. Furthermore, knowing respondents' preferred slicer can provide valuable insights into how reported issues are being addressed in their respective programs.

A major fault in the design of the study was the inability to connect the identity of the beta testers with the respondents of the questionnaire. This omission excluded the option of comparing the ratings and verbatim comments of the respondents with their behavior within RvP and demographic information. Such a comparison could have supported or expanded upon the results of the study.

Rating scale scores

As mentioned earlier, it is important to emphasize that the interpretation and conclusions drawn from the questionnaire ratings should not be considered definitive due to the low number of respondents. This variation is evident in figure 3.1, where responses significantly differ across several items. For example, in item 3, one respondent provided a score of 1, while another respondent gave a score of 7. Such variations underscore the necessity of a larger number of responses to establish more reliable mean ratings for RvP.

When looking at values in table 3.2, it becomes clear that the system generally receives a lower than average rating, with the exception of item 4 and 8.

The low rating of item 4 *(I was able to complete tasks quickly using REALvision Pro)*, which only slightly surpasses the mean from the reference data by 0.16, indicates a level of satisfaction among the respondents regarding the speed of task completion in RvP. However, upon examining the raw scores, it becomes evident that opinions are divided. Two respondents rated the item positively (ratings: 1

and 2), while the other two respondents rated it neutrally (4) and slightly disagreed (5), see figure 3.1. Therefore, it would be incorrect to conclude a consensus among the respondents regarding satisfaction with the speed at which they were able to complete tasks using RvP, despite the low mean rating of the item.

The low rating for item 8 *(Whenever I made a mistake using REALvision Pro, I could recover easily and quickly)* could indicate a sufficient level of usability when it comes to recovering from mistakes. However, the low rating of this item could also be a result of very few mistakes occurring in general during the beta session. This aspect, in itself, is positive, but the rating may not accurately reflect how easy it actually is to recover from mistakes when they do happen. Furthermore, when examining the log data, it becomes apparent that not a single participant utilized the undo function. This observation further supports the assumption that only a few mistakes occurred during the beta session. Therefore, the interpretation of this item could have benefited from observing how users interacted with the software, which would have provided insights into the frequency of mistakes and the ease with which they were fixed.

The items within the *Interface Quality* subscale all receives the highest ratings of all the items in the questionnaire. Initially this would indicate that the participants are generally dissatisfied with the interface of RvP, although by a small margin. However similarly to item 4 the opinions of the respondents are divided for all three items within this subscale, see figure 3.1. This means that it would be incorrect to conclude a consensus among the respondents regarding dissatisfaction with the quality of the interface.

Study 2: In-person usability test 4

The following chapter contains a description of the second usability evaluation, with a subsequent presentation and discussion of the results.

4.1 Aim of study

Based on the discussion points highlighted in Chapter 3.7, it has become evident that conducting more in-depth tests, which involve direct observation and live communication with the participants, would likely result in a greater richness and level of detail in the obtained information. This, in turn, could lead to improved proposals for addressing the reported usability issues. The traditional method of usability testing involves observing the user's interaction with a system, typically using a think-aloud protocol (TA). In addition to gathering more detailed data, this study aims to address the main objective of the project, which is to determine the effectiveness of ARUT methods in testing the usability of desktop applications (as described in Chapter 1.1). Hence, a traditional UT will be conducted in this study as a benchmark to compare findings and assess the effectiveness of the methods used in study 1.

Similarly to the first study, the aim of this study is also to identify and assess the severity of existing usability issues within REALvision Pro (RvP). To guide this study, the following research question have been formulated:

Which usability issues exist in RvP and how severe are they?

4.2 Methods

The study will employ three types of methods: in-person usability testing (UT), questionnaire, and a debriefing interview. UT will serve as the primary source of data, aiming to highlight usability deficiencies. The primary purpose of the questionnaire is to reveal additional usability issues. Lastly, a debriefing interview will be conducted to follow up on observations made during the UT in order to gain a deeper understanding of the participants' behavior.

4.2.1 Usability test

One of the primary components of a UT is having the user engage with the product by completing a set of predetermined tasks [Hertzum (2020)]. Consequently, a list of tasks will be created for the participants to carry out.

4.2.1.1 Tasks

Since the study is exploratory, there are no major preconceived issues regarding the program. Therefore, the tasks will primarily be based on the most frequently used functions [Rubin and Chisnell (2008)]. However, certain fundamental aspects of RvP are designed differently from popular slicers like Cura and PrusaSlicer. These areas could potentially pose difficulties for the participants, as they only have experience with other slicers. Therefore, it would be beneficial to test for potential usability deficiencies related to frequently used functions, as well as areas that differ in design from other slicers. The following list contains the 13 tasks selected for the UT:

1. Select Ultimaker 2+ printer model
2. Import the STL file
3. Turn the model correctly
4. Double the size of the object and place it in the bottom left corner
5. Change the infill pattern to 'Triangular'
6. Increase the shell thickness to 4 mm
7. Select 'Coarse' layer height
8. Make the infill density of support structures 25 percent
9. Apply support to all overhangs, no matter the overhang angle
10. Rename the object to 'Model'
11. Slice, and preview the sliced model and ensure that support are in the desired places (if not then redo)
12. Check if the infill pattern is correct
13. Export and save the file on the desktop

The Danish translation of the tasks can be seen in appendix B.2.1.

Tasks 1, 2, and 13 are all necessary actions required to correctly transform an STL file into printable G-Code. As a result, these steps naturally need to have an excellent level of usability.

The same applies to the first part of task 11, which involves slicing the model. However, the second part of task 11, along with task 12, is included to assess whether participants can verify that their changes have been successfully applied to the sliced version of the model by utilizing the preview feature.

Tasks 3 and 4 are included because the method of specifying these settings differs from other slicers. Specifically, they lack an interactive tool that appears on the model when attempting to rotate, move, or scale it. Therefore, it is important to determine whether participants can easily perform these actions in RvP.

Tasks 5, 6, 7, and 9 are among the most commonly used functions in slicers because all of the respective settings are always visible in basic/simple mode[1] (such as in RvP, Cura, and PrusaSlicer). Additionally, these settings are typically covered in online beginner guides for slicers. These assumptions are based on several YouTube guides, including ones for Cura, PrusaSlicer, and general slicing tutorials [General (2020*a*), General (2020*b*), Revolution (2022), Tech (2021), Sanladerer (2020)].

Task 8 aims to test whether participants are able to locate settings that are only available when switching settings modes.

The goal of testing task 10 is to assess whether the current UI is adequate for participants to understand how to change the name of the model.

After completing the last task, participants will be asked the following question: *Is there something that you usually do, that was not included in the task list?*. This question is included to ensure that the most frequently used features are being tested.

Pilot study

A pilot study of the tasks was carried out to evaluate the clarity of the task formulations and identify possible issues. Two interns from Create It REAL were recruited for the pilot study. Their experience with slicing and RvP is limited, as they have only used it a couple of times. The pilot study resulted in several minor grammatical changes. Additionally, it became clear that task 4, which contains multiple subtasks, should be split into separate tasks. The revised list of tasks can be found in the appendix B.2.2. It is important to note that tasks going forward will be referenced in relation to the revised task list.

Task success

Since the study is formative in nature, the main focus will be on identifying usability issues, with less emphasis placed on user performance. Performance measurement is primarily employed in summative tests that aim to assess or compare the level of usability of completed products. However, the success rate of task completion will be collected to provide an overview of the program's most challenging aspects to use.

Task success is the most commonly used performance metric in evaluations. The success of a task can

[1]The setting modes in slicers determine the level of complexity of the settings displayed. Typically, slicers include basic/simple, advanced, and expert modes.

be measured in various ways. The simplest form is binary success, where a task is either completed successfully or not. However, binary success may lack detail when tasks can be partially completed, and this aspect becomes important to consider. Therefore, measuring task success in levels is also possible, allowing for different levels of success between complete success and complete failure. For this type of measurement, the different levels of success have to be clearly defined [Albert and Tullis (2013)]. Albert and Tullis (2013) describes two ways of defining the level of success: a method based on assistance needed and one based on the degree of problems experienced during task completion. The former breaks the level of success into six levels: complete success, with or without assistance; partial success, with or without assistance; and failure where the user either thought it was complete or gave up. The latter method consists of four levels categorized whether the user experiences: no problems, minor problems, major problems and failure/giving up.

Measuring task success solely in a binary manner would overlook situations that result in partial task completion. Additionally, since the study will already address the issues participants encounter, categorizing task success based on the severity of problems would be somewhat redundant. Therefore, the success of tasks will be measured using levels of success based on the assistance required, as it is important to determine which tasks users needed assistance with in order to complete them partially or successfully.

To establish the criteria for identifying when a user is receiving assistance, it is necessary to define the different types of assistance. Albert and Tullis (2013) identifies four types of assistance: (1) the moderator resets the system to the initial pre-task state; (2) the participant receives probing questions or the moderator restates the task; (3) the moderator provides helpful information or answers participant questions; (4) participants turn to an external source for assistance. These variations will be utilized during the evaluation to determine if participants have received assistance or not.

Assisting the participant

When evaluating the usability of products, it is crucial to provide assistance to participants in completing their tasks only as a final option, as this can significantly impact the test results. Therefore, assistance should be offered only when absolutely necessary. Some tasks, like importing an STL file, act as prerequisites for other tasks. In such cases, if a participant is unable to complete the task, assistance will be provided to ensure successful completion, in order to avoid hindering subsequent tasks. Assistance may also be provided if a bug/crash occurs or if the participant is about to give up on a task [Rubin and Chisnell (2008)].

The UT will consist of the following ways of gathering data; Observation, note-taking and TA.

4.2.1.2 Observation and note-taking

The user will be observed while executing the given tasks. During observation, interesting events will be noted. Additionally, notes can be based on the user's verbalization from the TA protocol. The screen and audio will also be recorded, allowing the interaction to be analyzed after the UT and reducing the need to note everything. Reviewing the recordings enables the detection of unnoticed issues or events of interest [Hertzum (2020)]. Moreover recordings will be used to evaluate task success.

4.2.1.3 Think-aloud protocol

When selecting the type of TA protocol to use, it is important to look at the difference in benefits and withdrawals of the methods. Concurrent TA has shown to negatively impact task performance (like accuracy and response time), as the user spend considerable cognitive resources on performing the TA while executing tasks. This is not the case with retrospective TA as the user is able to focus solely on completing the task(s) first and the TA after. Retrospective TA has however been been shown to not deliver as truthful a reflection, due to the user forgetting or fabricating information about their experience [Russo et al. (1989)]. Another significant drawback of retrospective TA is the increased time consumption compared to concurrent TA [Van den Haak et al. (2004)].

The concurrent TA protocol has been selected for the present study due to its lower time consumption, considering that the study also involves a questionnaire and interview, which collectively make the study quite time-consuming. Additionally, the potential drawback of worse task performance is not considered significant, as the study is formative and does not focus on the affected performance metrics.

4.2.2 Questionnaire

The questionnaire from the previous study will be reused to measure the participants' satisfaction with RvP and understand the reasons behind potentially low levels of satisfaction. The questionnaire will be administered to participants after they have completed the UT. However, Item 11, *'REALvision Pro has all the functions and capabilities I expect it to have'*, has been removed from the questionnaire. This decision is based on the fact that participants are required to complete predetermined tasks that are possible to accomplish, and thus the functions and capabilities needed to complete the tasks are already present in the current version of RvP. A description of the questionnaire and the list of items can be found in chapter 3.2.1.

4.2.3 Debriefing interview

After the UT is completed and the participants have filled out the questionnaire, a short debriefing interview will be conducted. This interview serves to address any unanswered questions, encourage participants to expand on their thought processes and statements, and provide an opportunity for them to explain the reasons behind observations noted by the moderator.

The first set of questions will be general, focusing on the participants' thoughts regarding the overall design of the program and how it compares to their preferred slicing software.

Subsequent questions will be based on observations and statements noted by the moderator, as well as responses provided in the questionnaire, requiring further explanation by the participant [Rubin and Chisnell (2008)].

Finally participants were asked if they had any final remarks, before concluding the study.

4.3 Participants

There are two criteria for participating in the test. The first criterion is having no prior experience with RvP. The second criterion is a minimum of 3 months of experience with slicers. The latter requirement ensures that the participants are representative users, such as 3D printing hobbyists or professionals, who are likely to be familiar with the terminology used and functions tested in the UT. Initially, it was desired to have a total of five participants, as research has shown that this number is sufficient to uncover most of the existing usability issues [Nielsen (2000), Virzi (1992)]. However, only three subjects could be recruited for the study. All three participants are Danish males. Demographic information about the participants is presented below in table 4.1.

ID	Age (years)	3D printing experience	Preferred slicer	Usage context (Personal or professional)
P1	18-25	1 year	Cura	Personal
P2	56 or older	9 - 12 months	Cura	Personal
P3	46-55	3 - 6 months	Creality (Cura)	Personal

Table 4.1. The table contains demographic information about the participants.

4.4 Equipment

The following equipment and material was used during the test:

- ASUS VIVOBOOK S15 S510UN-BQ146T 15.6" LAPTOP
- Huawei P20 (Debriefing audio recording)
- OBS Studio 29.1.1 (Screen and audio recording during UT)
- REALvision Pro 4.1.5[2]
- Wireless computer mouse
- Online questionnaire (SurveyXact) described in section 3.2.1

[2] There are no changes to the interface in this version compared to the older version used in study 1.

- Consent declaration (including demographic questions)

4.5 Procedure

The study took place in the homes of the participants, in order to reduce the logistical effort required to participate [Rubin and Chisnell (2008)]. Furthermore, this allowed participants to remain in their usual work environment, as all three participants were 3D printing hobbyists. All equipment used in the study was portable and could fit in a single backpack.

Since all tests were conducted on the same laptop, it was necessary to manually reset all the settings to default before each test, as RvP does not reset the settings automatically upon exiting. Therefore, the following actions were taken after each test:

- Reset printing settings
- Select basic printing settings
- Select Creality printer model
- Apply all material types in preview
- Delete saved G-code file

Figure 4.1 shows the four sections of the study.

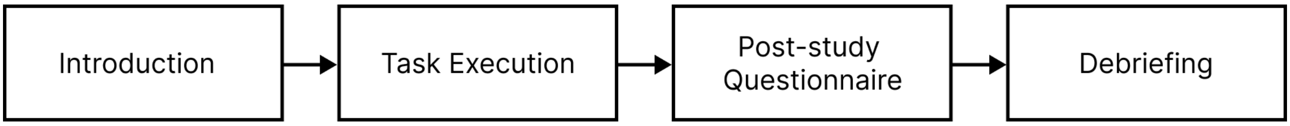


Figure 4.1. The figure presents the procedure of the usability test.

Introduction

Participants were initially handed the consent declaration, which also contained a short description of the test, along with demographic questions. Afterward, they received a more detailed explanation of the purpose and procedure of the test, which can be seen in appendix B.1. They were then shown an example of how to perform TA, where the moderator demonstrated a task involving changing the background on the desktop. Finally, participants were asked if they had any questions before booting RvP.

Task execution

When participants were ready to begin the UT, OBS Studio recording was started, and tasks were verbally given one at a time. Reading the tasks aloud instead of handing them on a piece of paper ensured that the participants understood the objectives and allowed the moderator to control the pace of the test [Rubin and Chisnell (2008)]. During the UT, notations of observations and statements of interest were made. After completing all the tasks, participants were asked if any of their frequently used functions were not included in the tasks. The UT was then concluded.

Post-study questionnaire

Following the UT, participants were presented with the online questionnaire. While filling out the questionnaire, they were able to ask clarifying questions about the items.

Debriefing

When participants had completed the questionnaire, an audio recording was begun and participants were asked a set of debriefing questions, before the test was concluded.

4.6 Results

During the first test RvP crashed during the execution of task 12. Here the participant was asked to preview the sliced model (showing all the individual layers). When the function was clicked the program crashed. The program was restarted and the button was clicked again, resulting in the program crashing again. Therefore it was decided to stop the task section of the test.

The crash is due to the laptops low performance, as the feature works on computers with higher performance. It was not possible to reschedule the tests or acquire a portable computer with a higher performance before the next tests, so it was decided to change task 12 and 13. Task 12 was reduced so it only contained 'Slice the model' and not the 'preview and check that the support is correct' part. Task 13 was changed to 'Check if the infill pattern is Triangular, without clicking 'Show preview''. This final version of the task list can be seen in appendix B.2.3.

The crash was not discovered before the tests began, because the procedure and pilot studies was being performed on a different computer (than the test laptop), with better performance.

The identity of the participants is labelled 'P1', 'P2' and 'P3'.

A debriefing session was not conducted with P2 as the UT dragged on for longer than expected, and the participant had time constraints that prevented us from proceeding with the session.

4.6.1 Rating scale scores

Since only three participants have rated the items, any conclusions based on these scores should not be regarded as definitive

The same procedure from chapter 3.6.1 is used in the analysis of rating scale scores for this study. This includes comparing the results to a set of reference data points collected by Sauro and Lewis (2012) and presenting the individual responses. The only difference from the previous study is the exclusion of item 11 from the questionnaire. Therefore likewise the subscale *Information Quality* a value for the subscale *Interface Quality* will not be calculated.

Table 4.2 presents scores from the questionnaire along with means from the data collected by Sauro and Lewis (2012).

	Data from the present study		**Normative data [Sauro and Lewis (2012)]**	
Item	Mean (items)	Mean (subscales)	Mean (items)	Mean (subscales)
1	4.33	4.39	2.85	2.80
2	4.67		2.69	
3 (4)	4.67		3.16	
4 (5)	4.67		2.66	
5 (6)	3.67		2.27	
6 (7)	4.33		2.86	
7	-	-	3.7	3.02
8	4[1]		3.21	
9	-		2.96	
10	-		3.09	
11	-		2.74	
12	-		2.66	
13 (9)	3	-	2.28	2.49
14 (10)	3.33		2.42	
15 (11)	-		2.79	
16 (12)	4.33	-	2.82	-

Table 4.2. The table presents means of rating scale scores from the questionnaire in the present study and data gathered by Sauro and Lewis (2012). Both individual item scores and subscale scores are presented. Green cells indicate the lowest (thus more positive) mean between the two sets of data. Note that item 3 from the questionnaire used in this study is excluded from the table as it is not a part of PSSUQ version 3. As the questionnaire used in this study was modified, the item numbers do not correspond with the item numbers of the original PSSUQ version 3. Therefore the corresponding item numbers from the questionnaire used in this study is presented in parenthesis next to the item numbers from the original PSSUQ version 3.
[1]P3 responded 'N/A' to item 8, and the score is therefore only based on the response from P1 an P2.

Item 3 *(I was able to effectively complete tasks using REALvision Pro)* from the questionnaire used in the present study received a mean rating of 4.

Upon reviewing the mean scores from table 4.2, it is evident that none of the items from the questionnaire in the present study received ratings lower (positive) than the means obtained from the reference data set. In the line chart below, the responses for each respondent are presented.

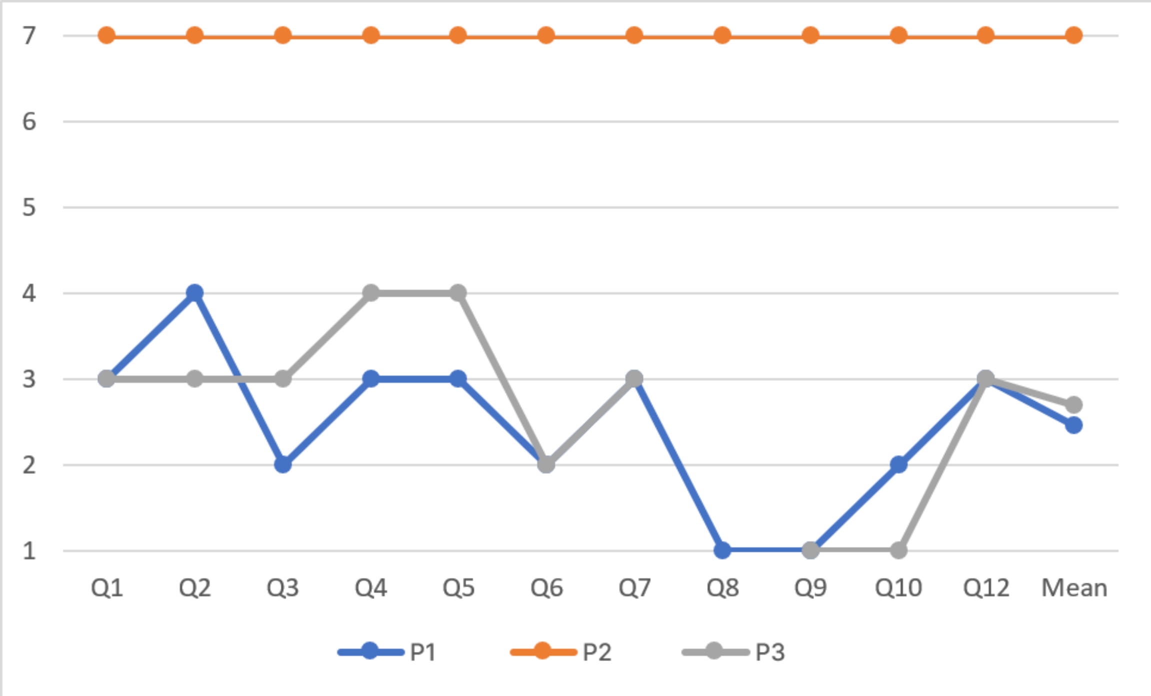


Figure 4.2. The figure presents a line chart of the item scores for the individual respondents. The respondents mean ratings (across all items) are the following: P1: 2.45, P2: 7, P3: 2.7. Note that item 8 only contains two scores as P3 selected the 'N/A' option.

Figure 4.2 shows that the respondents have different opinions about the RvP. Both P1 and P3 provided positive responses to most items, with some neutral ratings and no negative scores. In contrast, P2 responded with the highest score for every item.

Table 4.3 presents the level of task success each participant achieved during the UT.

	P1		P2		P3	
Task	Success	Ass. (Type)	Success	Ass. (Type)	Success	Ass. (Type)
1	Complete	No	Complete	No[2]	Complete	No[2]
2	Complete	Yes (3)	Complete[1]	Yes (3)	Complete	Yes (3)
3	Complete	No	Complete	No	Complete	No
4	Complete	No	Complete	No	Complete	No
5	Failure	User thought task was completed	Complete	Yes (3)	Failure	User thought task was completed
6	Complete[1]	Yes (3)	Complete	Yes (3)	Complete	No[2]
7	Complete	No[2]	Complete	No	Complete	No
8	Complete	No	Complete	Yes (3)	Complete	No
9	Complete	No[2]	Complete[1]	Yes (3)	Complete	No[2]
10	Failure	Gave up	Complete[3]	Yes (3)	Complete	Yes (2)[4]
11	Complete	No	Complete	Yes (3)	Failure	User thought task was completed
12	Complete	No	Complete	No	Complete	No
13	-	-	Complete	No	Failure	User thought task was completed
14	-	-	Complete	No	Complete	No

Table 4.3. The table provides an overview of the participants' level of success in executing each task, categorized based on the assistance they required. The first column beneath each participant indicates the type of success achieved. The second column specifies whether the participant received assistance (and which type of assistance), gave up, or mistakenly believed the task was completed.
[1]The participant was not able to successfully complete the task without assistance.
[2]The participant asked to be reminded of task information. However, as the requested information does not help completing the given task, it is not considered as receiving assistance.
[3]P2 only partially completed the task due to the preview function causing the program to crash. However he correctly explains how he would have completed the task if he had been able to view the preview. Therefore his task success is determined to be complete.
[4]Restating of task in order to remind participant of the second part of the task.

4.6.2 Usability issues

Usability issues will be extracted by listening to the participants verbalization during the UT in conjunction with reviewing the recorded footage of their interactions and notations made during the observation. Furthermore the questionnaire responses and the interviews will be reviewed as well.

To identify usability problems from the data recommendations from Hertzum (2020) will be followed. This book presents a comprehensive list of situations, which typically arise from usability issues, that should be analyzed within the test data. These situations include various aspects, such as users expressing negative emotions or surprise towards the system, requiring assistance to accomplish tasks, performing incorrect actions to complete a task, overlooking unnoticed information, or experiencing mental overload. Furthermore, usability-related situations may involve the product's non-compliance with standards or conventions, as well as a lack of trustworthiness and credibility. During concurrent TA a disconnect between what the user thinks and what happens in the interface is a clear usability issue as well.

Another consideration when analyzing the data is whether a difficulty experienced by a participant is severe enough to be considered a usability issue. This depends on how central the feature, at which the issue occurred, is to the product. "*For the central product features, smaller difficulties qualify as usability problems*" [Hertzum (2020)].

Considering the frequency of an issue as a criterion for qualifying it as a usability issue is advised primarily for studies involving a larger number of users. Tests conducted with a smaller number of users generally do not determine usability issues based on frequency [Hertzum (2020)].

Each usability issue will be labeled with a number (followed by the task in which the usability issue occurred), which will be used to reference the specific issue. All the identified issues can be seen in appendix B.3. Transcriptions of the UT and the debriefing interviews, as well as responses from the questionnaire can be found in the annex within the 'Study2' folder.

A total of 61 usability issues were identified, out of which 44 are unique. Among the participants, P1 encountered 14 usability issues, P2 experienced 28, and P3 encountered 19.

Usability issue 56 (*P3 expects a single click on the object (in the object list) to activate renaming*) will not be considered, as both P1 and P2 were using double-click. Therefore, it is presumed that the general expectation among users does not involve activating the renaming function with a single click.

4.6.2.1 Severity rating

Each of the identified usability issues will be assigned a severity rating to help prioritize their importance and urgency. The severity rating is based on the impact that each issue has had on the user experience. For rating the usability issues, the severity rating system from the ARUT, described in chapter 3.6.2.1, will be utilized.

Unlike the ARUT, participants in this study are performing defined tasks, which makes it easier to determine if an issue is of high severity. Although the participants are performing the exact same tasks this time, the frequency of an issue will not directly influence its severity rating due to the limited number of participants. However, the frequency will be taken into consideration when assessing the overall impact of the issues.

Due to the high number of usability issues, the rationale for each assigned rating will not be presented, unlike in the results from study 1. Table 4.4 presents the severity and frequency of the identified usability issues.

Severity (Total)	Experienced by one participant	Experienced by two participants	Experienced by three participants
High (8)	12, 36, 57, 60	3, 14	1, 7
Medium (10)	8, 11, 16, 26, 27, 33, 41, 46	34, 35	
Low (25)	4, 5, 9, 10, 13, 22, 23, 24, 25, 28, 37, 39, 40, 42, 43, 45, 48, 52	18, 20, 29, 30, 32	2, 6

Table 4.4. The table presents the number of usability issues categorized within the three levels of severity, as well as the frequency of each usability issue.

4.7 Discussion

Not being able to test the 'Show preview' feature, due to it crashing RvP has had a negative effect on the authenticity of the tests, as all three participants attempted to use the preview function. This was reported by all participants to be an essential part of their usual procedure when using slicers. It is therefore unfortunate that the usability of such a frequently used feature is not being tested, considering it is an action they always perform. By using the intended laptop in the pilot test, the crash could have been discovered and a laptop with better performance could have been acquired in time for the actual UT.

During the UT, participants occasionally received assistance when they should not have. This occurred when participants asked questions and the moderator inadvertently failed to reject them, or when the moderator assumed that participants had given up before they formally expressed it, resulting in premature hints being provided. The option of receiving assistance should have been more strictly controlled in order to reduce the amount of noise within the data.

Another aspect that should have been more strictly controlled was the requirement for participants to clearly express when they believed that they had completed a task. This expectation was quickly forgotten by the participants, despite being informed about it during the test introduction. However, the moderator inadvertently failed to remind them to do so, which may have led the participants to believe it was unnecessary. This lack of confirmation from the participants could result in the moderator prematurely moving on to the next task, potentially overlooking unidentified usability issues.

Comparing the volume of qualitative data (comments) derived from the questionnaire between the two studies in this project, it is evident that this study resulted in considerably less data. Besides the first study having one more respondent, this could be due to several of the following factors:

- During the UT in study 2, the participants provided a large volume of qualitative information. This may have led them to believe that they had already provided all the information they could, resulting in a perception of repeating themselves when providing comments in the questionnaire.

In contrast, the questionnaire was the sole method (the only one used anyway) for participants in study 1 to provide qualitative feedback.

- Participants in study 1 may not have been aware that they were able to skip the comment sections of the questionnaire, leading them to believe that they were obligated to provide a response. Similarly, P1 and P3 from study 2 may not have been aware of this option. However, P2 was informed that he only had to respond to the questions he found relevant. As a result, he provided a total of 4 responses out of the 11 comment sections.
- P3 provided the same response for 6 items, significantly reducing the variability in his data.

The qualitative responses from the questionnaire did not contribute to the identification of usability issues that were not discovered during the UT. This suggests that the comment sections in the questionnaire are redundant when conducting a UT that includes TA.

It became evident during the UT that fulfilling the roles of both moderator and observer was challenging at times. Having multiple researchers dedicated to each role would have yielded more valuable insights, as it was difficult to collect an adequate amount of notes while simultaneously moderating the test. Consequently, there were fewer notes available for the final analysis, ultimately impacting the quality of the results.

It has also been reported that involving multiple evaluators in the data analysis process leads to a higher number of identified usability issues, which is referred to as the *evaluator effect* [Hertzum and Jacobsen (1999), Jacobsen et al. (1998)]. Therefore, if more than one evaluator had analyzed the data, it would likely have increased the quantity of identified usability issues.

Initially, task 5 was included primarily to assess participants' ability to use the 'Move' function. However, it became apparent that the real challenge of the task was determining the front side of the print bed. While all participants were able to move the object without difficulty, both P1 and P3 incorrectly identified the right side of the print bed as the front. This identification issue exposed a usability issue (6, 19, 49) and also uncovered a bug in the system.

During the UT with P2, the printing settings were inadvertently not reset, leading to the shell thickness value already being set to 6mm when task 7 (*Set the shell thickness to 6 mm*) was reached. P2 was then asked to change the value to 5mm instead, uncovering usability issue 23. This issue would not have been discovered if the settings had been properly reset.

Task 10 (*Apply support to all possible overhangs, no matter the overhang angle, except for the eyes*) posed significant challenges during the usability test. P1 gave up on the task, and P2 was unable to complete it without using the unavailable preview. P2 also commented during the test at 00:22:01, saying, *"Now we're definitely getting into something advanced."* This task involved combining two tasks into one, requiring the moderator to remind P3 about the second part. Dividing the task into two separate tasks could be beneficial, particularly isolating the task of removing support from the model's eyes.

Rating scale scores

The means from table 4.2 indicate a generally lower level of satisfaction with RvP compared to the average system. However, it is important to examine the individual scores provided before making any definitive assumptions about the general attitude towards RvP. Upon observing figure 4.2, it becomes evident that the overall attitude towards RvP is not negative, as only P2 has given a mean score higher than 4. Additionally, the scores provided by P2 demonstrate the highest level of dissatisfaction towards RvP, which significantly increases the mean rating for each item. It should be noted that P2 also encountered the highest number of usability issues, which likely contributed to the extreme ratings given. On the other hand, both P1 and P3 reported a positive mean rating, indicating that the majority of participants had a positive attitude towards RvP. Interestingly, their ratings show some correlation, with differences never exceeding 1 scale point.

When comparing the number of usability issues encountered with the age of the participants, there appears to be a correlation. The youngest participant, P1, encountered the lowest number of usability issues (14), while the oldest participant, P2, experienced the highest number of 28. However, it is important to note that this observation is based on a small sample size of only three participants, so it may be a random occurrence. Nonetheless, it is still worth noting. Another possible explanation for the low number of usability issues encountered by P1 is that he has the most experience with 3D printing, although the difference in experience among P1 and P2 is relatively small.

Several points of interest were noted for P2 during the UT:

- P2 expressed a need for a step-by-step guide or wizard that would assist in navigating through the essential steps of model preparation. He emphasized the importance of knowing where and when to apply support to the model.
- Another observation during P2's UT was that he consistently hovered over icons and buttons, waiting for the pop-up information boxes to appear. This allowed him to read the functionality before clicking and seeing the outcome.
- When needing to return to the main interface, P2 stated at 00:22:01: *"Now the question is whether one dares to close the window. I probably can't break anything here, I'm counting on the fact that I can't."* This statement reflects the cautious approach and fear of unintentionally causing any issues.

These observations indicate that P2 is a cautious user who takes great care not to make mistakes or disrupt the system. His actions are deliberate, and he approaches tasks with hesitancy, seeking certainty about the outcomes of his actions. While this may not explain the extremely negative ratings provided, it does suggest that P2 requires an exceptionally good user experience in order to feel comfortable with a system.

The amount of variety within the scores presented in figure 4.2 showcase the need for more respondents in order to reliably draw definitive conclusions about the general level of satisfaction of RvP.

General discussion 5

In this chapter a comparison of the usability testing methods used in study 1 and 2 will be discussed.

Purely looking at the results, it is evident that study 2, which used a traditional form of in-person usability testing, contributed significantly more rich data compared to study 1, despite involving one less participant. Study 2 yielded a total of 44 unique usability issues, while only 10 were derived from study 1. These findings align with the results of similar comparative studies comparing traditional in-person and asynchronous remote usability testing [Andreasen et al. (2007), Bruun et al. (2009)]. However, it should be noted that the primary methods for data collection in these studies differ from the one used in study 1. Throughout ARUT studies, the prevalent method for users to report encountered usability issues is through critical-incident reports. Although study 1 also included this method, it was implemented ineffectively, resulting in no reports being made. Plausible reasons for this are discussed in greater detail in Chapter 3.7.

One of the advantages of ARUT is the ability to scale the sample size without significantly increasing resource requirements, thus allowing for a larger amount of data to be gathered. However, due to the low number of respondents (4) in study 1, this advantage was not utilized. Therefore, it remains uncertain whether the number of usability issues identified in study 1 could have approached that of study 2 if a larger sample size had been available.

Another factor that could have influenced the lower number of reported usability issues is the duration of RvP usage. If it turns out that participants from study 1 spent less time using RvP, this could result in encountering fewer usability issues. An indication of this can be seen in the response of respondent 2 from study 1, who stated in the questionnaire that they quit the app in frustration after struggling to import an STL file (Q3.a, Q7.a, and Q8.a). These comments suggest that the respondent was stuck for a significant amount of time, exploring only a limited portion of the program. However, it was not possible to determine the exact amount of time participants in study 1 spent using the program due to a flaw in the study's design, which prevented linking the time spent in RvP with the questionnaire responses.

When comparing the methodologies of the usability tests used in study 1 and study 2, a significant difference lies in the more controlled setting of study 2, where participants were required to complete a predetermined set of tasks. This approach allows for more comparable results and ensures that the usability of various aspects within RvP is tested. However, this approach may also overlook issues that arise when attempting tasks not included in the predetermined set.

When examining the unique usability issues discovered in the two studies, it becomes apparent how challenging it is to compare the findings. Most usability issues identified in study 1 are general and do not specify particular interactions but rather encompass broad aspects, such as usability issue 9: *'Graphical user interface was not comfortable to use'* or 10: *'Found some printer settings well hidden and in unexpected location'.* In contrast, usability issues identified in study 2 typically focus on specific interactions with particular elements of the UI. The latter type of definitions provide more detailed information, making it easier to translate those issues into concrete design proposals. In fact, it is nearly impossible to determine how to address usability issues like number 8 (*The arrangement of information made completing tasks less effective*) or 9 (*'Graphical user interface was not comfortable to use'*) from study 1. However, it is worth noting that usability issues 1, 2, 6, and 7 from study 1 were not discovered during study 2, as they pertain to aspects of RvP that were not explored in study 2 due to the selected set of tasks. This aligns with the aforementioned concerns regarding the moderator's influence on which areas of the program are tested for usability.

When comparing the findings, only one usability issue can be confidently said to be repeated among the results from the two studies. This is usability issue 5 (*Spent a lot of time trying to figure out how to upload STL file (user was looking under 'File' in the top tab)*) from study 1 and usability issues 1, 15 and 44 (*Difficulty figuring out how to import an STL file.*) from study 2. Due to the vague definition of usability issues 4, 8, 9, and 10 from study 1, it is unclear if they are related to any specific usability issues from study 2. For instance, usability issue 27 (*Unable to locate 'support infill density' which is hidden inside 'Advanced' mode*) from study 2 could potentially be the same usability issue encountered when a user from study 1 reported that *'The arrangement of information made completing tasks less effective*' (usability issue 10). While this scenario is unlikely, it highlights the importance of providing detailed descriptions of usability issues, as there is no way of knowing which information is being refereed to in usability issue 10 from study 1.

Usability issues 3, 5, and 6 from study 1 are specific issues defined similarly to those discovered in study 2. The user who reported usability issue 5 also provided an explanation of their expectations, which is valuable information when attempting to address the issue. However, determining the underlying reasons behind issues 3 and 6 is challenging without any video or audio material available. Therefore, proposed solutions to these issues would have to rely on assumptions about what went wrong.

Now that the findings of the two studies have been compared, it is important to consider the difference in required resources associated with them. As mentioned before, studies have consistently shown that the in-person usability evaluation method outperforms the asynchronous remote method in terms of results [Andreasen et al. (2007), Bruun et al. (2009)]. However, the former method requires significantly more resources from the researchers' perspective, including the presence of a moderator and the increased amount of data that needs to be analyzed by an evaluator. The balance between resources and results must always be considered in relation to the subject and goal of the research.

However, when comparing the current iteration of the developed usability testing tool from study 1 with the methods utilized in study 2, it becomes evident that the drawbacks of the additional resources spent in study 2 have been overshadowed by the sheer quality and quantity of the findings.

Conclusion 6

The research question posed in this project was: *"What methods of asynchronous remote user testing can be employed to effectively detect potential usability issues in a desktop application?"* The aim was to explore ways of enhancing the usability of software applications by developing a tool for evaluating the usability of desktop applications that can be deployed remotely and asynchronously, with the purpose of increasing efficiency and reducing the resources required for conducting usability evaluations.

During the project, two studies were conducted: study 1, which employed asynchronous remote usability testing (ARUT), and study 2, which used traditional in-person usability testing. Study 2 served as a benchmark to measure the quality and quantity of data gathered during study 1. Both studies aimed to identify the usability issues within REALvision Pro and assess their severity. The findings from study 1 resulted in a total of 10 usability issues, none of which were of high severity, while study 2 provided a total of 44 usability issues, including 8 that were categorized as high severity. These results indicated that study 2 yielded significantly richer data and identified more usability issues compared to study 1. This finding is consistent with previous comparative studies that have also highlighted the advantages of in-person testing in terms of data quantity.

Furthermore, the usability issues identified in study 2 were of higher quality compared to those reported in study 1. This was attributed to the availability of different data sources, such as think-aloud, video material, and debriefing, which provided more specific definitions of the issues and often included the underlying reasons behind them. This facilitates a more accurate and efficient process when proposing improved design changes.

However, this project was not without limitations. In study 1, the implementation of the critical-incident report method proved ineffective, resulting in zero responses. Additionally, the analysis of log data yielded no findings, primarily due to a lack of tracked functions and the small sample size. Furthermore, the small sample size and the unknown duration of participants' usage may have limited the number of reported usability issues in study 1. Consequently, the questionnaire remained the only source of data during study 1, which ultimately proved to be less effective than initially assumed.

The discussion also highlighted the differences in methodologies between the two studies, with study 2 employing a more controlled setting and predetermined tasks, while study 1 allowed participants to engage with the application freely. This difference in approach resulted in the identification of different types of usability issues.

In conclusion, while the advantages of asynchronous remote usability testing, such as scalability and flexibility, are recognized, the findings suggest that traditional in-person testing still provides richer data and identifies a greater number of usability issues. However, it is important to consider the specific objectives of the usability testing and the available resources when choosing between remote and in-person methods. Within the context of reviewing the usability of REALvision Pro, the quality and quantity of findings from the in-person usability test outweighed the increase in resources required compared to the asynchronous remote method used in study 1.

Future research on the method developed in study 1 should focus on improving the implementation of critical-incident reports as the primary data gathering method to address the challenges of participants providing vague responses. Additionally, conducting testing on a larger sample size and addressing the limitations of participation duration would provide further insights into the effectiveness of the tool.

Overall, this project contributes to the understanding of usability testing methods and provides insights into the challenges and advantages of conducting them remote and asynchronous. The developed tool for evaluating the usability of desktop applications remotely and asynchronously has significant room for improvement to enhance its capability of gathering data of high quality and quantity.

Bibliography

Albert, W. and Tullis, T. (2013), *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*, Interactive Technologies, 2nd edn, Elsevier Science, San Francisco.

Alhadreti, O. and Mayhew, P. (2018), Rethinking thinking aloud: A comparison of three think-aloud protocols, *in* 'Conference on Human Factors in Computing Systems - Proceedings', Vol. 2018- of *CHI '18*, ACM, pp. 1–12.

Andreasen, M., Nielsen, H., Schrøder, S. and Stage, J. (2007), What happened to remote usability testing?: an empirical study of three methods, *in* 'Conference on Human Factors in Computing Systems - Proceedings', CHI '07, ACM, pp. 1405–1414.

Bidgoli, H. (2010), *The handbook of technology management, supply chain management, marketing and advertising, and global management*, Vol. 2, John Wiley & Sons.

Brooke, J. et al. (1996), 'Sus-a quick and dirty usability scale', *Usability evaluation in industry* 189(194), 4–7.

Brush, A., Ames, M. and Davis, J. (2004), A comparison of synchronous remote and local usability studies for an expert interface, *in* 'CHI '04 Extended Abstracts on human factors in computing systems', CHI EA '04, ACM, pp. 1179–1182.

Bruun, A., Gull, P., Hofmeister, L. and Stage, J. (2009), Let your users do the testing: a comparison of three remote asynchronous usability testing methods, *in* 'Conference on Human Factors in Computing Systems - Proceedings', CHI '09, ACM, pp. 1619–1628.

Castillo, J., Hartson, H. and Hix, D. (1998), Remote usability evaluation: can users report their own critical incidents?, *in* 'CHI 98 Conference Summary on human factors in computing systems', CHI '98, ACM, pp. 253–254.

Chin, J., Diehl, V. and Norman, K. (1988), Development of an instrument measuring user satisfaction of the human-computer interface, *in* 'Conference on Human Factors in Computing Systems - Proceedings', Vol. 130202 of *CHI '88*, ACM, pp. 213–218.

Clemmensen, T., Hertzum, M., Hornbæk, K., Shi, Q. and Yammiyavar, P. (2009), 'Cultural cognition in usability evaluation', *Interacting with computers* 21(3), 212–220.

Dray, S. and Siegel, D. (2004), 'Remote possibilities?: international usability testing at a distance', *Interactions (New York, N.Y.)* 11(2), 10–17.

General, T. D. P. (2020*a*), 'Cura slicer settings part 1/2 - 3d printing 104'.
URL: *https://www.youtube.com/watch?v=6Lk4RyD0VbU*

General, T. D. P. (2020*b*), 'Cura slicer settings part 2/2 - 3d printing 104'.
URL: *https://www.youtube.com/watch?v=GNoKh3ylS48t=0s*

Hartson, H., Castillo, J., Kelso, J. and Neale, W. (1996), 'Remote evaluation: the network as an extension of the usability laboratory', pp. 228–235.

Hartson, H. R. and Castillo, J. C. (1998), Remote evaluation for post-deployment usability improvement, *in* 'Proceedings of the working conference on Advanced visual interfaces', pp. 22–29.

Hertzum, M. (2020), *Usability testing : a practitioner's guide to evaluating the user experience*, Synthesis lectures on human-centered informatics, 45, Morgan Claypool, San Rafael, California.

Hertzum, M., Borlund, P. and Kristoffersen, K. B. (2015), 'What do thinking-aloud participants say? a comparison of moderated and unmoderated usability sessions', *International journal of human-computer interaction* 31(9), 557–570.

Hertzum, M. and Jacobsen, N. E. (1999), The evaluator effect during first-time use of the cognitive walkthrough technique., *in* 'HCI (1)', Citeseer, pp. 1063–1067.

Hornbaek, K. (2006), 'Current practice in measuring usability: Challenges to usability studies and research', *International journal of human-computer studies* 64(2), 79–102.

ISO (2018), 'ISO 9241-11:2018(en) - Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts'.
URL: *https://www.iso.org/obp/ui/iso:std:iso:9241:-11:ed-2:v1:en*

Jacobsen, N. E., Hertzum, M. and John, B. E. (1998), Evaluator effect in usability studies: Problem detection and severity judgments, *in* 'Proceedings of the Human Factors and Ergonomics Society', Vol. 2, Sage Publications Ltd, Santa Monica, pp. 1336–1340.

Lewis, J. R. (1995), 'Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use', *International journal of human-computer interaction* 7(1), 57–78.

McDonald, S., Edwards, H. M. and Zhao, T. (2012), 'Exploring think-alouds in usability testing: An international survey', *IEEE transactions on professional communication* 55(1), 2–19.

Millen, D. R. (1999), 'Remote usability evaluation: User participation in the design of a web-based email service', *SIGGROUP Bull.* 20(1), 40–45.
URL: *https://doi.org/10.1145/327556.327616*

Mixpanel (n.d.), 'Mixpanel analysis'.
URL: *https://mixpanel.com/analysis*

Nielsen, J. (1993), Chapter 9 - international user interfaces, *in* 'Usability Engineering', pp. 237–254.

Nielsen, J. (2000), 'Why you only need to test with 5 users'.
URL: *https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/*

Olmsted, E. and Gill, M. (2005), In-person usability study compared with self-administered web (remote—different time/place) study: Does mode of study produce similar results?, *in* 'Proceedings of UPA'.

Peres, S. C., Pham, T. and Phillips, R. (2013), 'Validation of the system usability scale (sus): Sus in the wild', *Proceedings of the Human Factors and Ergonomics Society* 57(1), 192–196.

REAL, C. I. (n.d.), 'Realvision pro'.
URL: *https://realvision.pro/*

Revolution, D. (2022), 'Prusaslicer beginner tutorial: Learn the basics'.
URL: *https://www.youtube.com/watch?v=*$_kIqMPNQNSw$

Rubin, J. and Chisnell, D. (2008), *Handbook of usability testing: how to plan, design, and conduct effective tests*, 2nd edn, Wiley, Indianapolis, Ind.

Russo, J., Johnson, E. J. and Stephens, D. L. (1989), 'The validity of verbal protocols', *Memory cognition* 17(6), 759–769.

Sanladerer, T. (2020), '3d printing basics: The slicer! (ep6)'.
URL: *https://www.youtube.com/watch?v=btLCKun-xFQ*

Sauer, J., Sonderegger, A., Heyden, K., Biller, J., Klotz, J. and Uebelbacher, A. (2019), 'Extra-laboratorial usability tests: An empirical comparison of remote and classical field testing with lab testing', *Applied ergonomics* 74, 85–96.

Sauro, J. and Lewis, J. R. (2012), *Quantifying the User Experience*, Morgan Kaufmann.

Scholtz, J. (1999), A case study: developing a remote, rapid, and automated usability testing methodology for on-line books, *in* 'Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers', Vol. Track2, IEEE, pp. 11 pp.–.

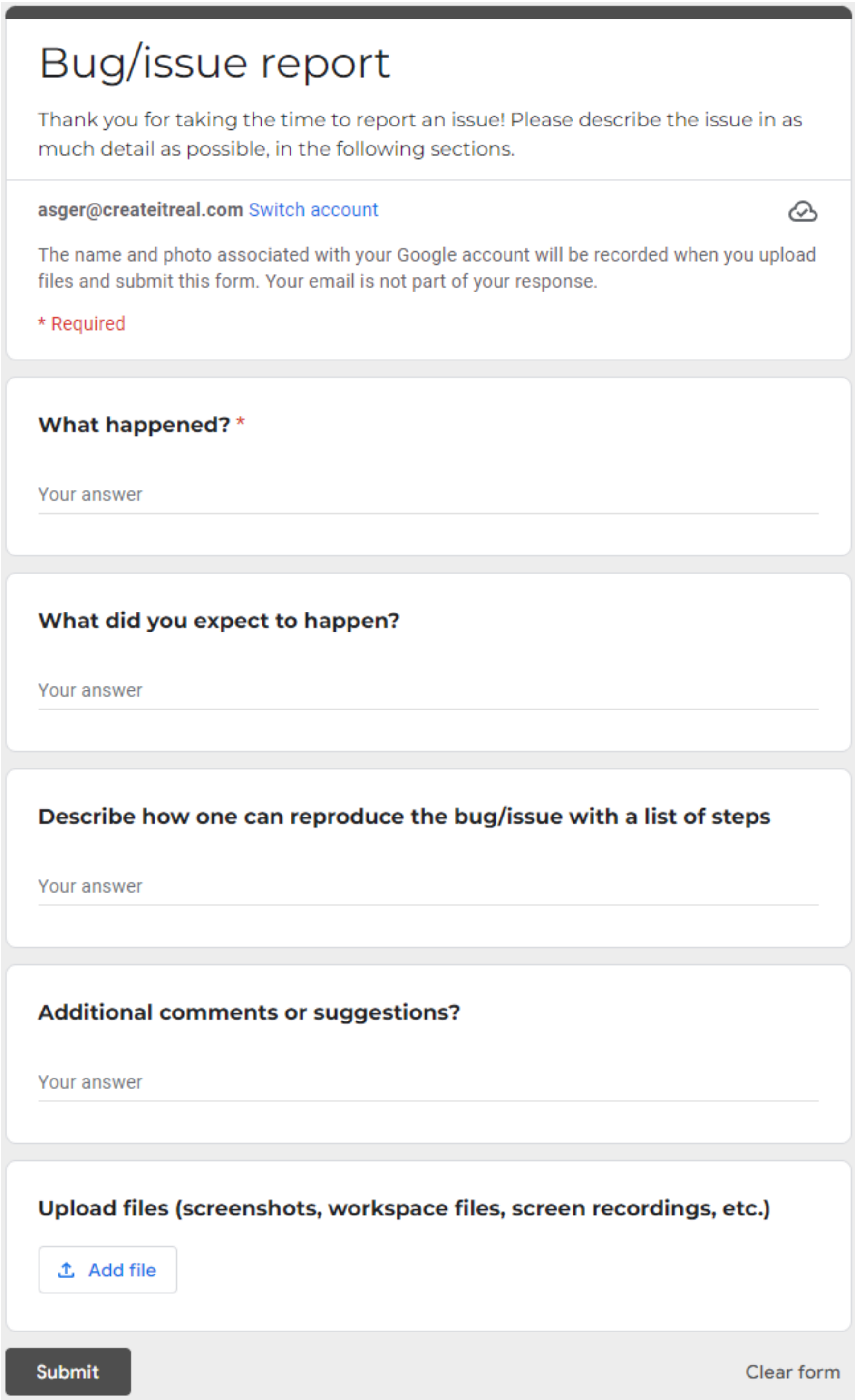
Tech, T. (2021), 'Beginners guide to 3d printer slicers'.
URL: *https://www.youtube.com/watch?v=yXNZ*$_2 - rHNE$

Van den Haak, M., de Jong, M. and Schellens, P. (2004), 'Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison', *Interacting with computers* 16(6), 1153–1170.

Virzi, R. A. (1992), 'Refining the test phase of usability evaluation: How many subjects is enough?', *Human factors* 34(4), 457–468.

Asynchronous remote usability test A

A.1 Critical incident report



Bug/issue report

Thank you for taking the time to report an issue! Please describe the issue in as much detail as possible, in the following sections.

asger@createitreal.com Switch account

The name and photo associated with your Google account will be recorded when you upload files and submit this form. Your email is not part of your response.

* Required

What happened? *

Your answer

What did you expect to happen?

Your answer

Describe how one can reproduce the bug/issue with a list of steps

Your answer

Additional comments or suggestions?

Your answer

Upload files (screenshots, workspace files, screen recordings, etc.)

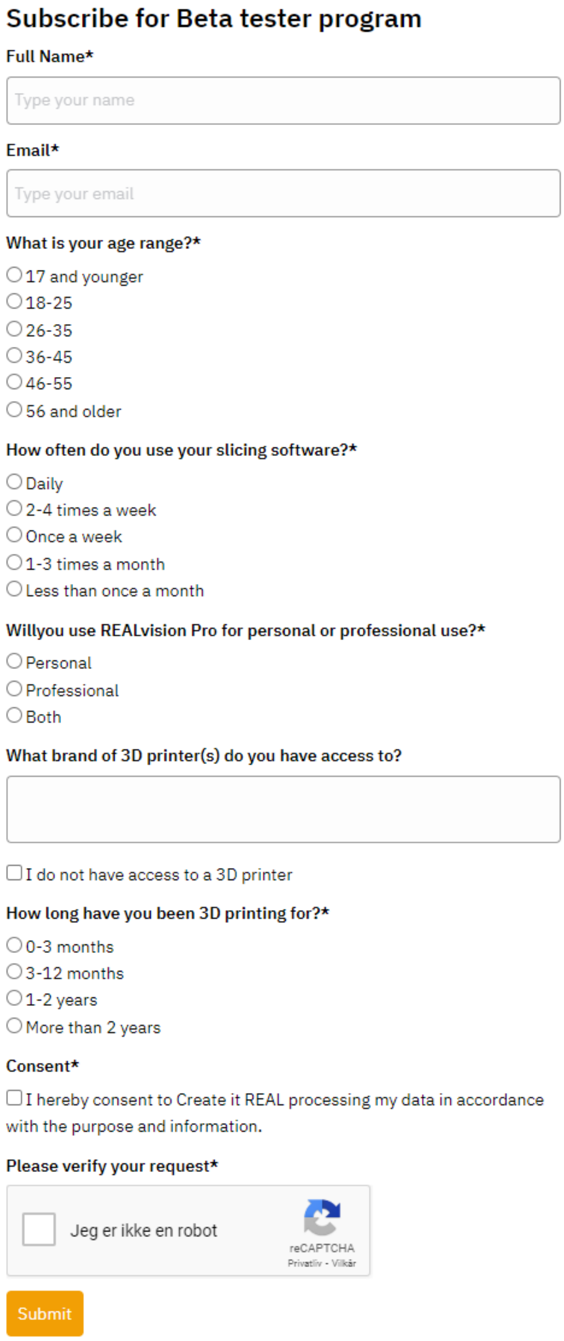
Add file

Submit

Clear form

Figure A.1. Critical incident report.

A.2 Demographic questionnaire



Subscribe for Beta tester program

Full Name*

Type your name

Email*

Type your email

What is your age range?*

- 17 and younger
- 18-25
- 26-35
- 36-45
- 46-55
- 56 and older

How often do you use your slicing software?*

- Daily
- 2-4 times a week
- Once a week
- 1-3 times a month
- Less than once a month

Willyou use REALvision Pro for personal or professional use?*

- Personal
- Professional
- Both

What brand of 3D printer(s) do you have access to?

- I do not have access to a 3D printer

How long have you been 3D printing for?*

- 0-3 months
- 3-12 months
- 1-2 years
- More than 2 years

Consent*

- I hereby consent to Create it REAL processing my data in accordance with the purpose and information.

Please verify your request*

Jeg er ikke en robot

reCAPTCHA

Privatliv - Vilkår

Submit

Figure A.2. Demographic questionnaire required to be completed in order to sign up for the beta program.

***Consent form**

This is a request for your consent to process your data. The processing aims to understand and develop tools that can track and measure the impact of changes to our slicing software REALvision Pro and identify areas within the software that need improvement. Data will also be anonymized and used in a Master thesis revolving around the subject above. You consent to the processing of the following data about you:

- REALvision Pro usage metrics and statistics
- Email
- Name
- Age
- Country
- 3D printing experience
- Context and frequency of 3D printer usage

Your data will be stored securely, and we will solely use the data for the above purpose. You always have the right to change your consent. If you wish to change your consent later on, you can revoke your consent by contacting us at the following email: beta@realvision.pro. The General Data Protection Regulation entitles you to obtain the information you find below.

The purpose of processing your data

The purpose of processing your data is to identify areas of improvement within our slicing software REALvision pro. This entails specific and general user experiences or functionalities that can be improved upon. The data will also be used to understand and develop tools for measuring remote asynchronous usability and user experience testing of computer platforms. Data will also be anonymized and used in a Master thesis revolving around the subject above.

We process the following personal data:

General personal data

- REALvision Pro usage metrics and statistics
- Email
- Name
- Age
- Country
- 3D printing experience
- Context and frequency of 3D printer usage

How we store your data

We will store your data as long as necessary for the data processing purpose for which we obtain your consent and follow the applicable legislation. We will then erase your data.

Your rights

You have several rights under the General Data Protection Regulation when we process your data. For example, you have a right to erasure and a right to data portability.

In certain cases, you have a right to access, a right to rectification, a right to restriction of processing, and a right to object to our processing of the personal data in question. Be aware that you cannot withdraw your consent with retroactive effect.

Do you want to complain?

If you believe that we do not meet our responsibility or do not process your data according to the rules, you may complain to the Danish Data Protection Agency at dt@datatilsynet.dk.

However, we encourage you also to contact us at beta@realvision.pro, as we want to do our utmost to accommodate your complaint.

Disclosure to and from third parties

Your data (or parts of your data) may be transferred to Mixpanel and Aalborg university.

Figure A.3. Consent form for signing up the beta program.

A.3 Introductory email

REALvision Pro

Dear Asger Møller Eriksen,

Congratulations! You have been selected to participate in the REALvision Pro beta tester program. We appreciate your interest in our product and are excited to have you on board.

To get started, please use the following link to download the installation file for the beta program: https://shop.realvision.pro/7wo9EkUN5n. Please note that this link is unique to you and should not be shared with anyone else.

Once you have downloaded the installation file, please follow the instructions to install REALvision Pro on your computer. You will then be able to access all the features and plugins.

As a beta tester, your feedback is incredibly valuable to us. If you encounter any bugs or issues while using the program, please use the following link to fill out our bug report form: https://forms.gle/uNfemMYcT62VbkoRA. This will help us identify and fix any problems as quickly as possible.

We encourage you to try out all the features and provide us with any feedback or suggestions that you may have. You can also send your feedback directly to us at beta@realvision.pro.

Thank you again for your participation in the REALvision Pro beta tester program. We look forward to working with you and making our product the best it can be.

Best regards,

Asger Moeller Eriksen

Sent to: asger@createitreal.com

Unsubscribe

Create It Real, Hjulmagervej 28, 9000 Aalborg, Denmark

Figure A.4. Introductory email that participants receive upon being accepted for joining the beta program.

A.4 Survey

A.4.1 Questions

Questions from the survey used in asynchronous remote usability evaluation. The questions are presented in the same order as in the survey.

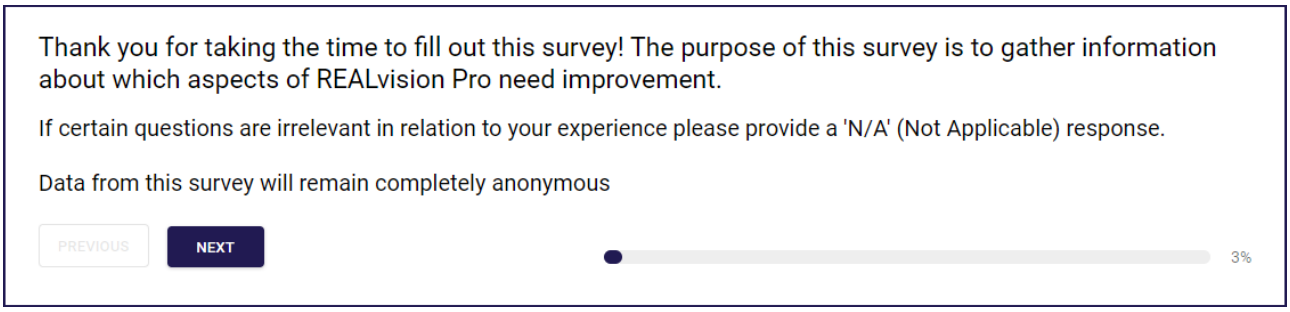


Figure A.5. Introduction to the survey.

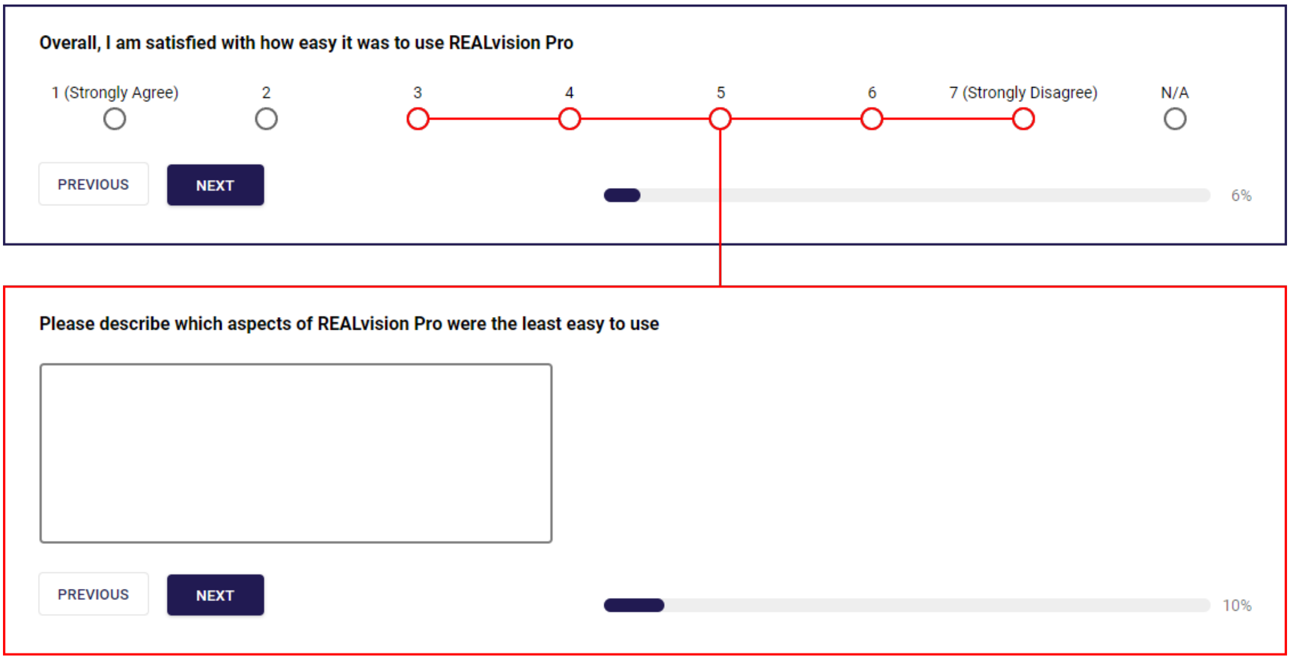


Figure A.6. Question 1. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

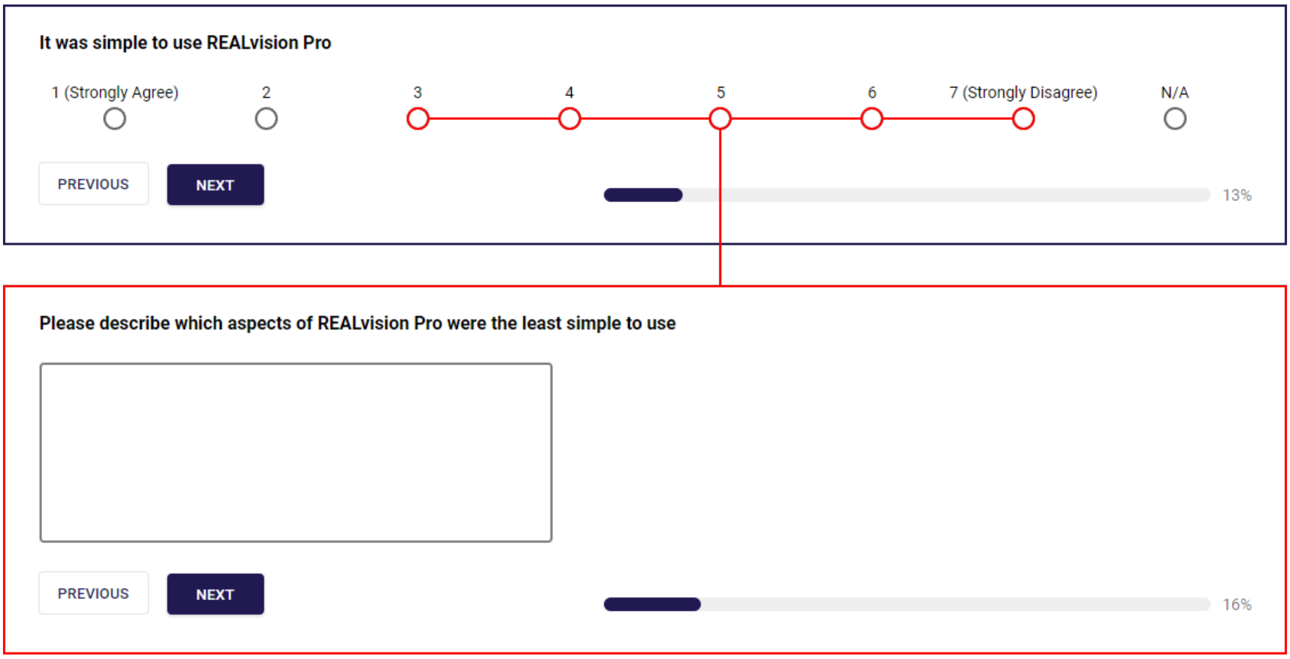


Figure A.7. Question 2. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

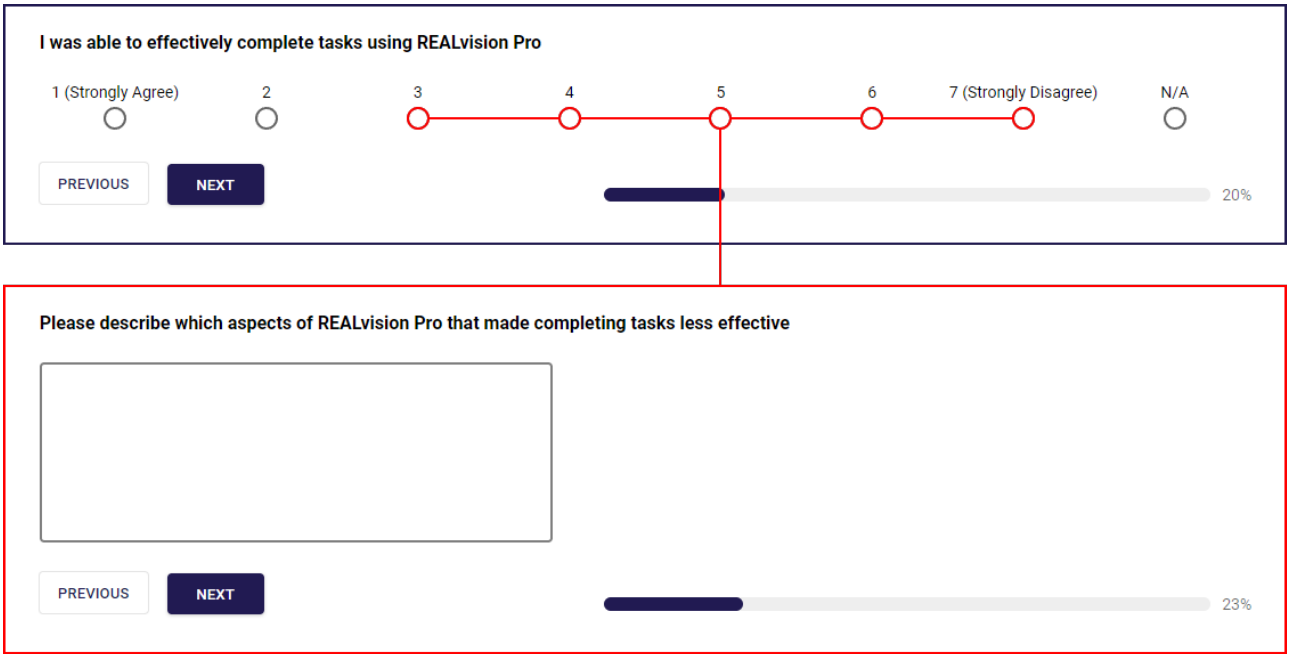


Figure A.8. Question 3. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

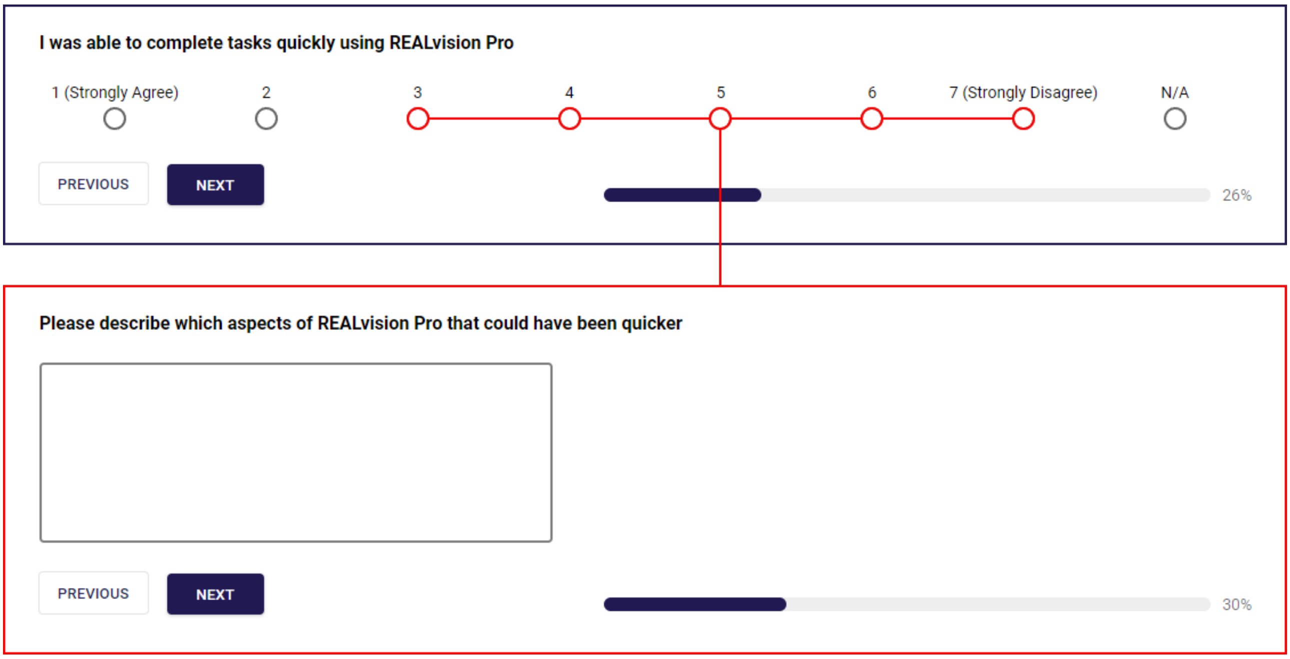


Figure A.9. Question 4. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

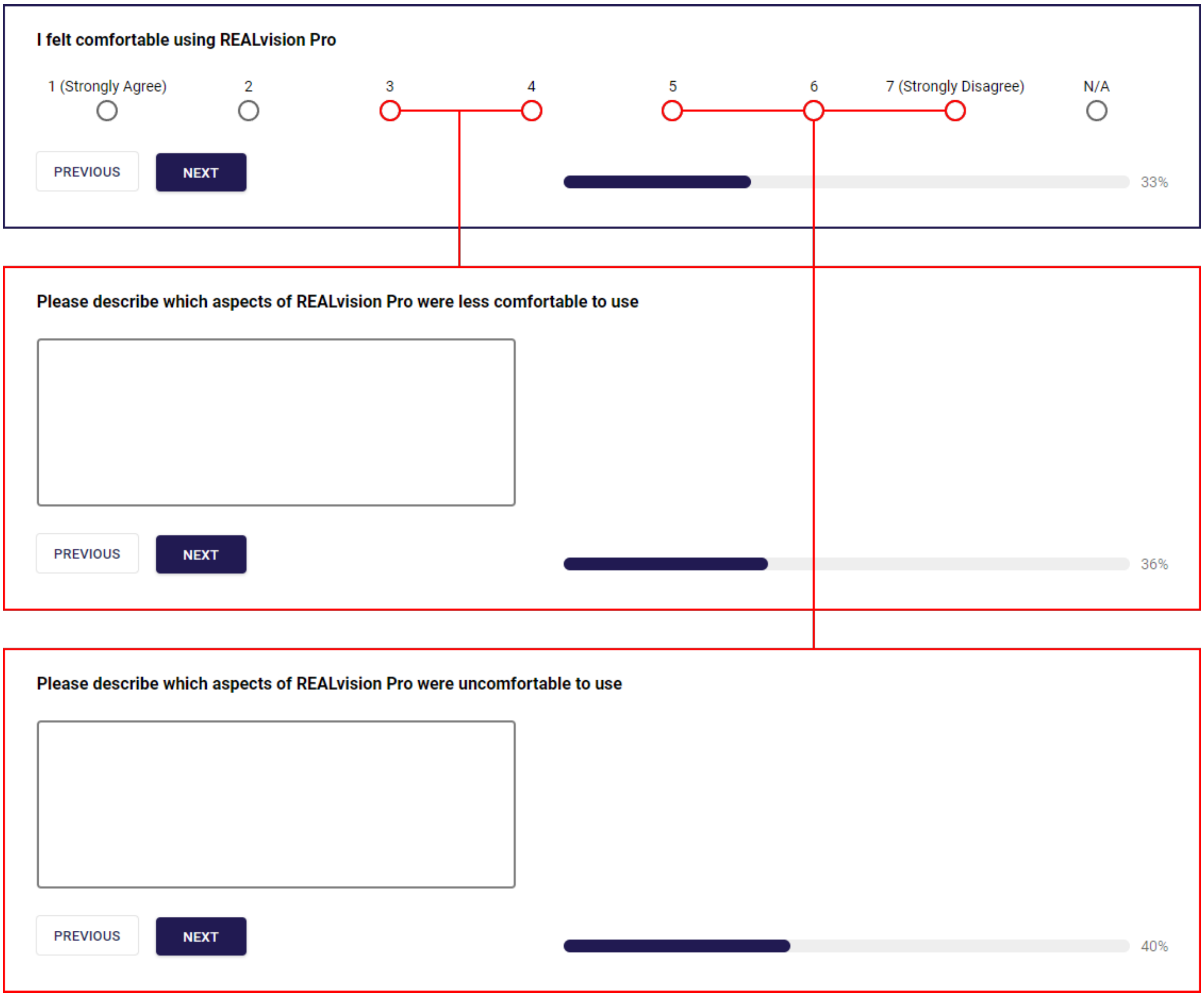


Figure A.10. Question 5. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

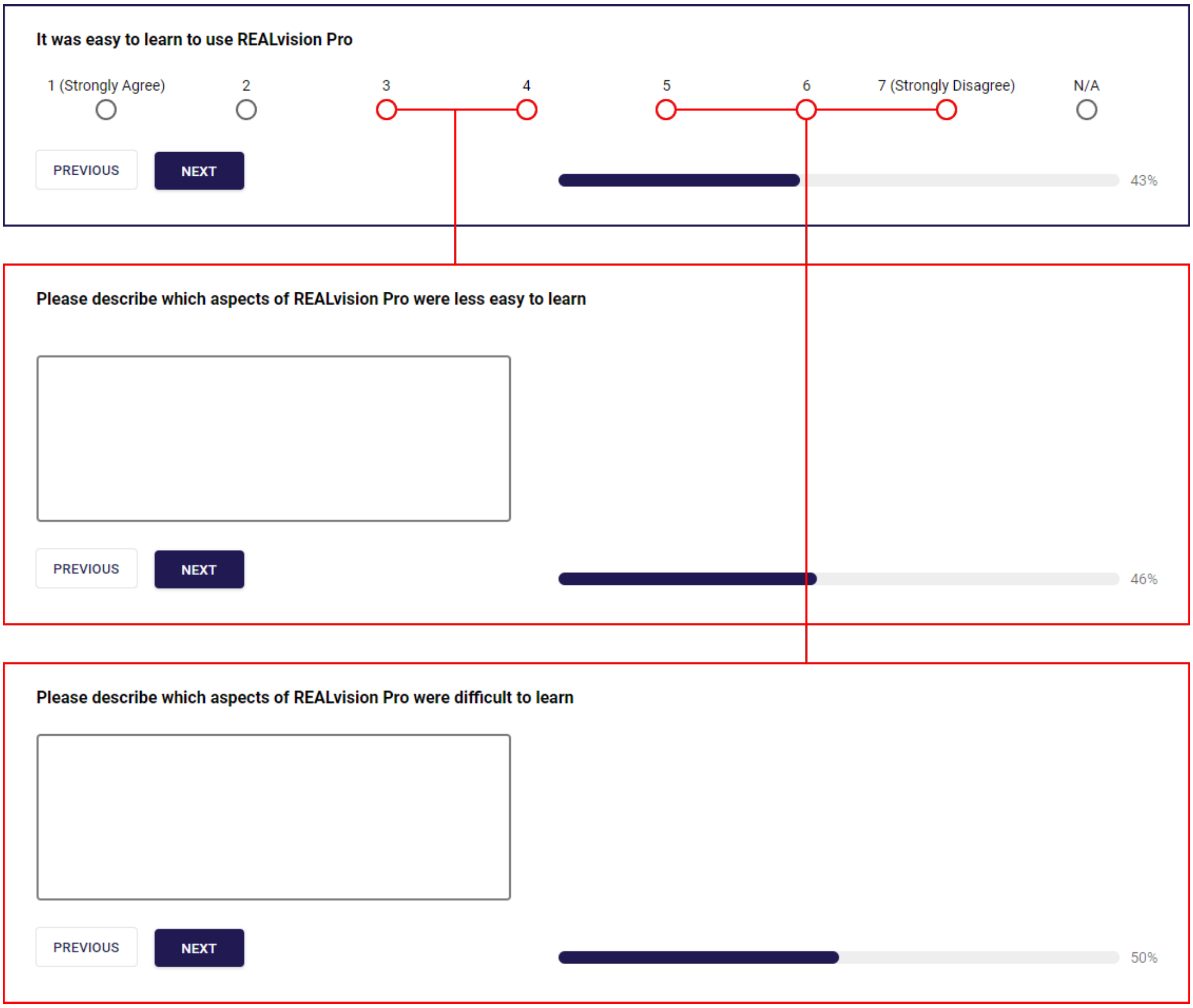


Figure A.11. Question 6. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

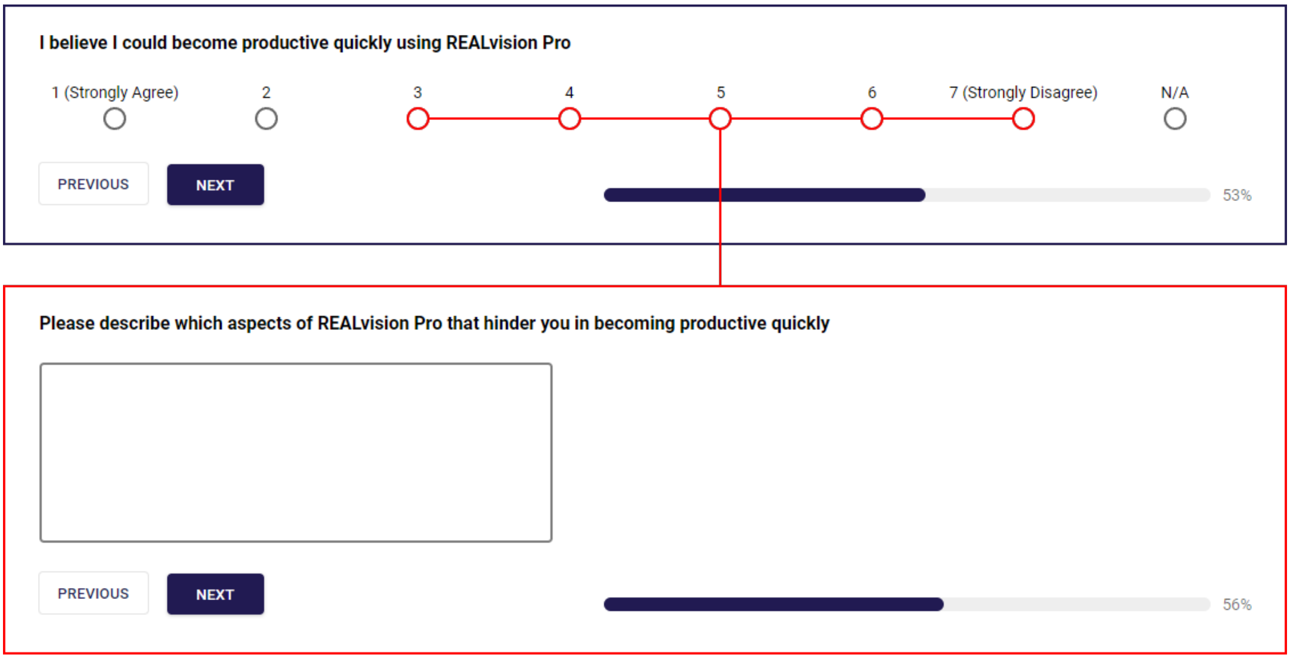


Figure A.12. Question 7. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

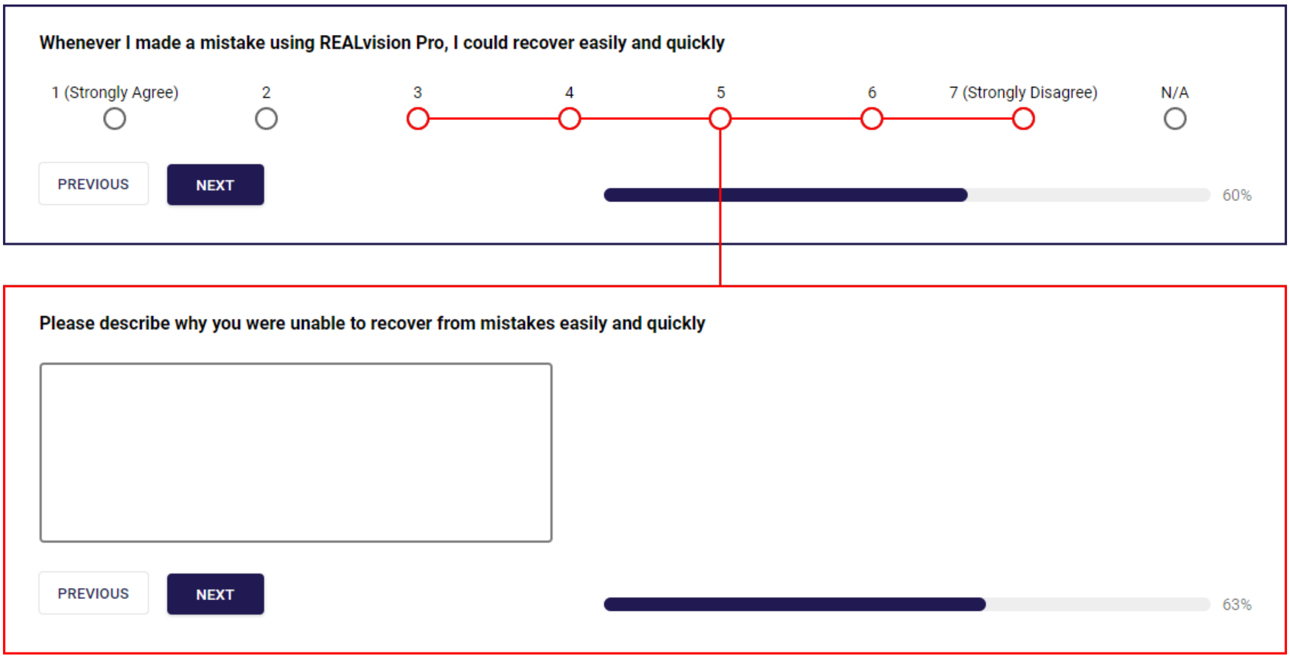


Figure A.13. Question 8. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

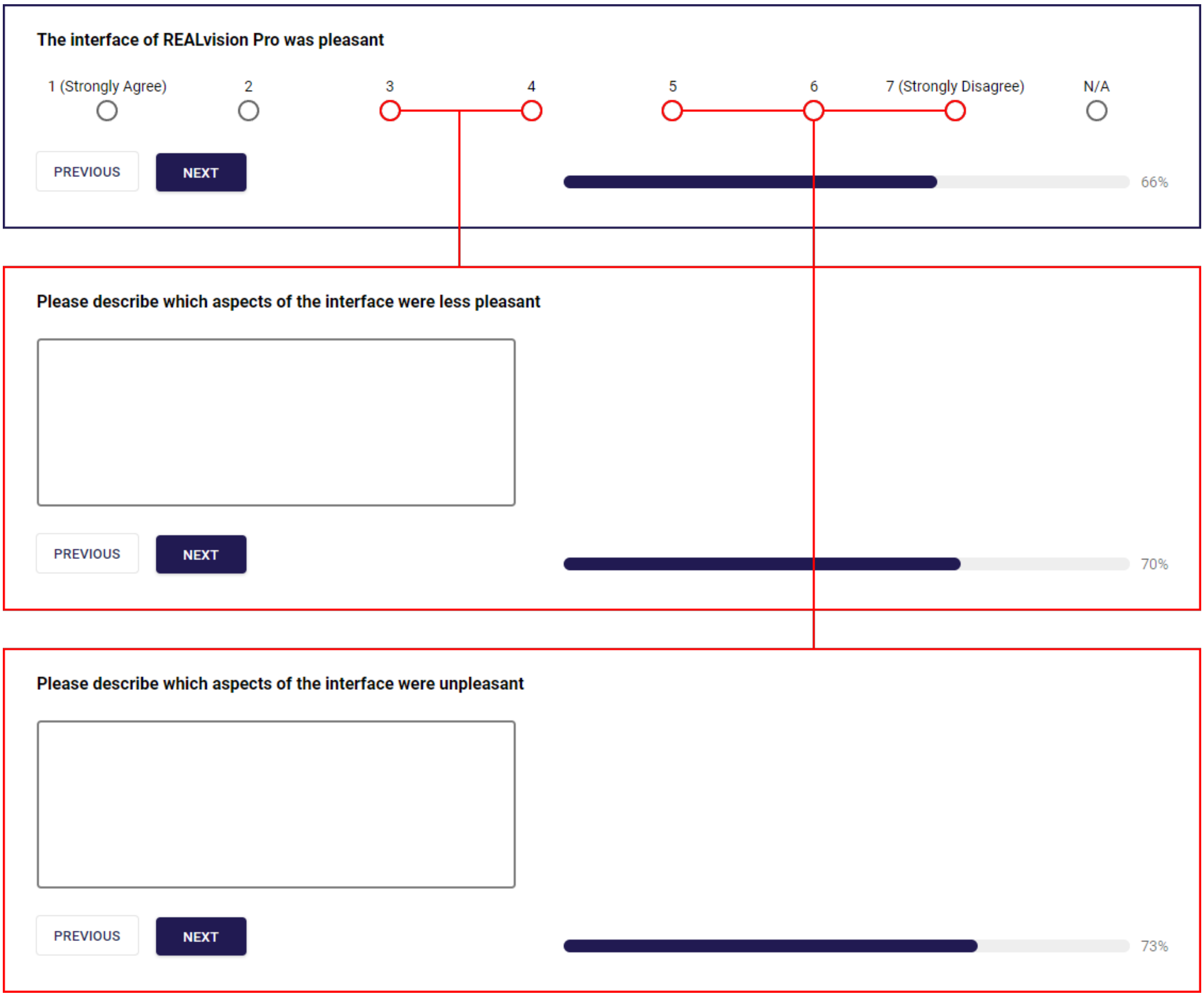


Figure A.14. Question 9. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

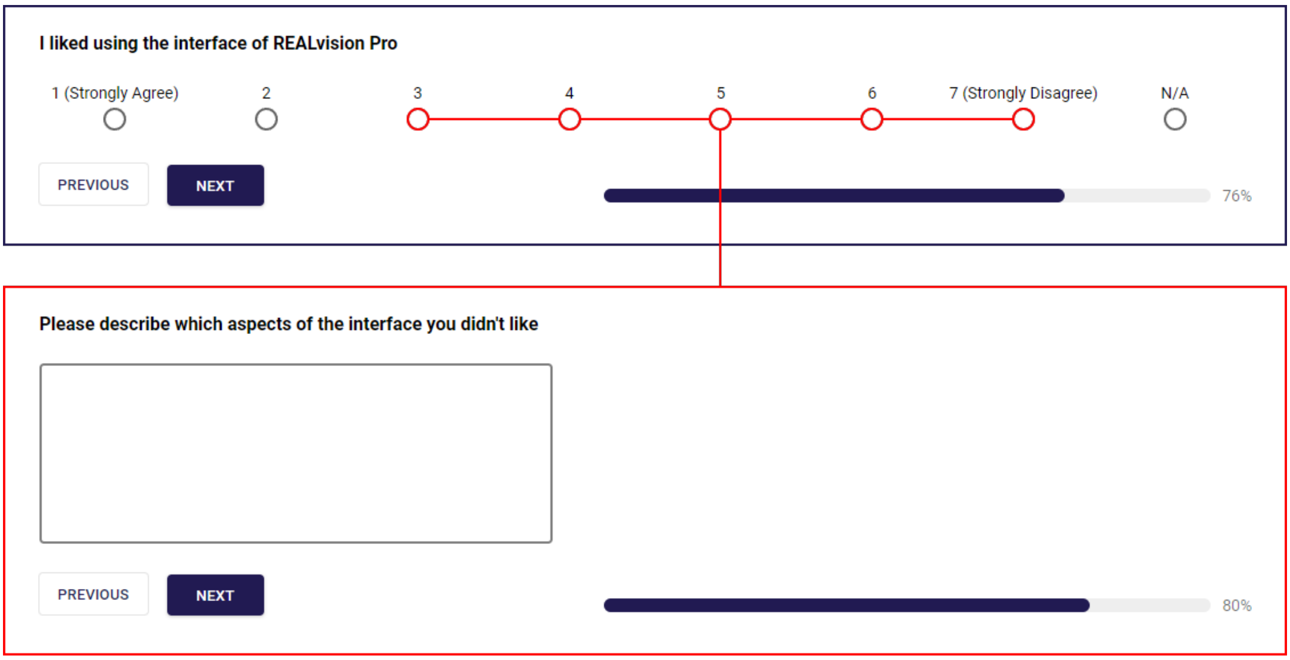


Figure A.15. Question 10. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

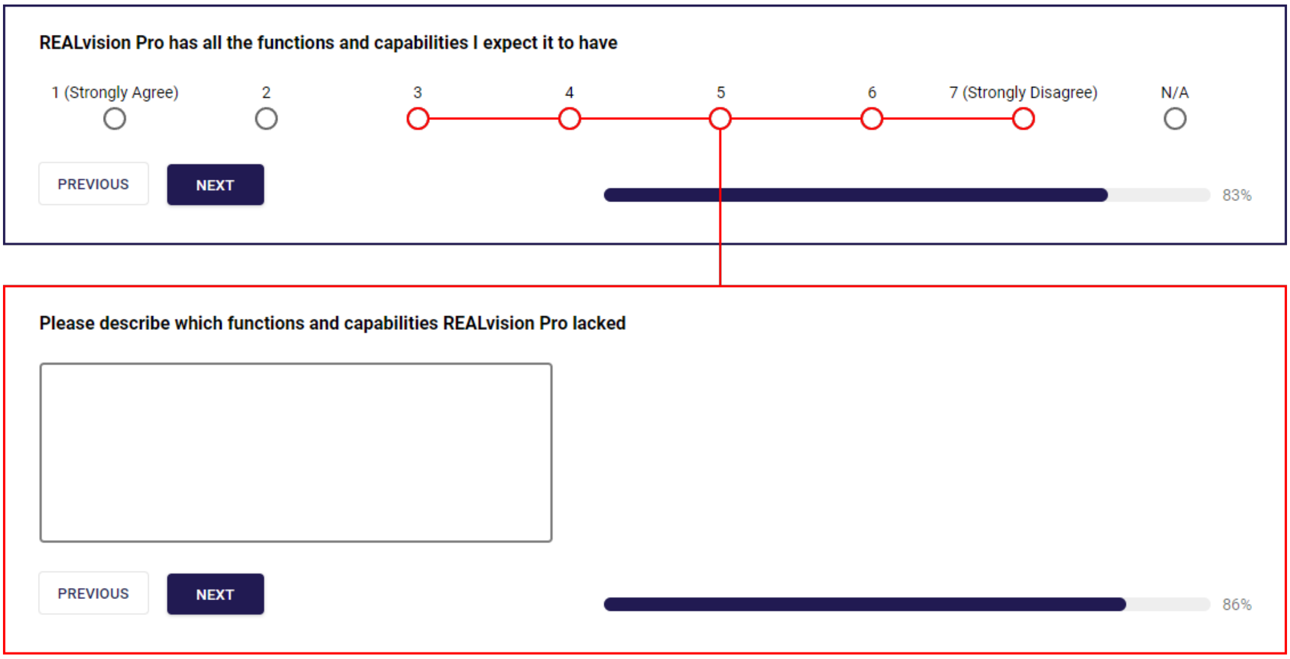


Figure A.16. Question 11. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

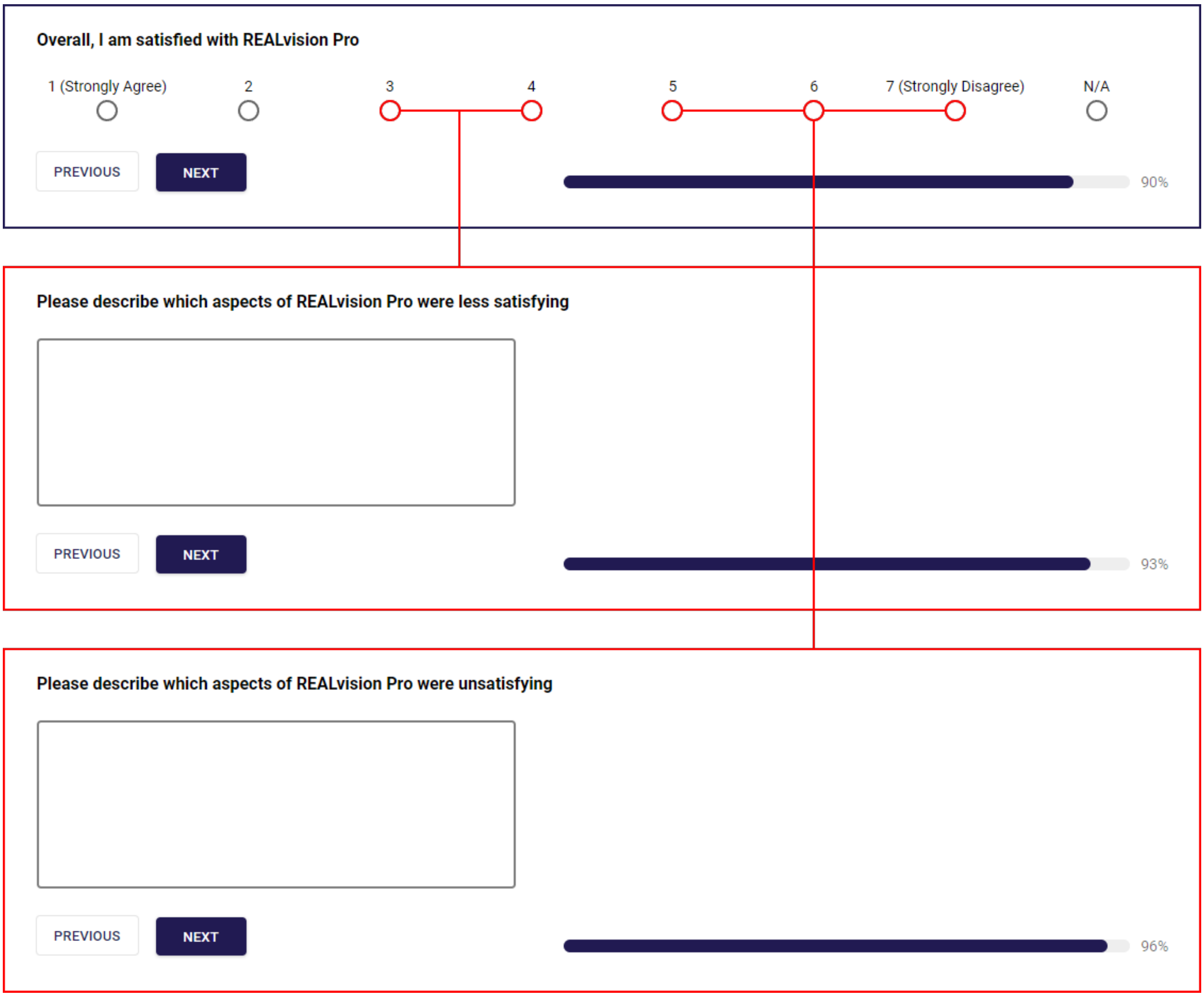


Figure A.17. Question 12. Selecting any of the response options marked with red will prompt a subsequent follow-up question (outlined with red).

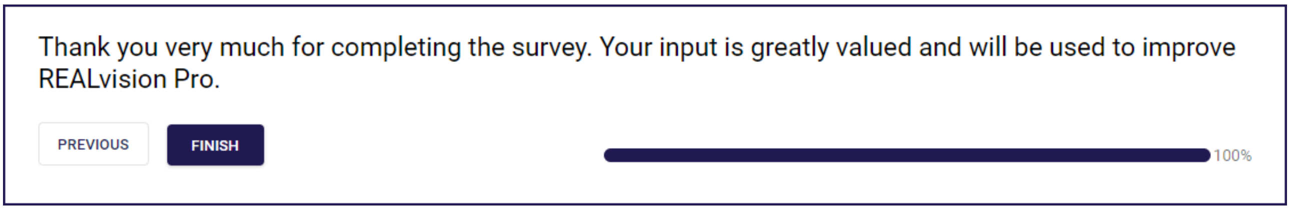


Figure A.18. Outro.

Traditional usability test B

B.1 Introduction

English version

You are going to perform a series of tasks in a slicer called REALvision Pro. I will present you with one task at a time. While you complete the tasks, you should engage in think-aloud, which means you should talk about what you are trying to do and what you are thinking and looking at. The more you speak, the better. This is to understand the reasons behind your actions.

I will provide you with an example of performing a task with think-aloud in a moment.

You should attempt to solve each task until you either solve it or reach the point where you would normally give up and seek help, such as online or from a colleague. I cannot assist you in performing the tasks, but please feel free to ask if the task is unclear. Let me know when you complete a task, using phrases like 'done' or 'finished.'

The purpose of the experiment is to evaluate the program, not you.

After the tasks, you will be given a questionnaire aimed at collecting your opinions about the program and the reasons behind them.

Finally, I will ask you some general questions and some related to your task performance or questionnaire responses.

Perform think-aloud example - Does it make sense?

Any questions?

Danish version

Du skal gennemfører en række opgaver i en slicer som hedder REALvision Pro. Jeg kommer til at stille dig én opgave ad gangen. Imens du udfører opgaverne skal du udføre think-aloud hvilket indebærer at du taler om hvad du forsøger at gøre samt hvad du tænker og kigger på. Jo mere du snakker, jo bedre. Det er for at forstå grunden bag dine handlinger.

Jeg kommer til at give dig et eksempel på en udførsel af en opgave med think-aloud om lidt.

Du skal forsøge at løse hver opgave indtil du løser den eller du når det punkt hvor du normalt ville give op og søge om hjælp, eksempelvis online eller fra en kollega. Jeg kan ikke hjælpe dig i udførelsen af opgaverne, men spørg endelig hvis opgaven er uklar. Udmeld når du løser en opgave, med eksempelvis 'færdig' eller 'done'.

Formålet med forsøget er at vurdere programmet og ikke dig.

Efter opgaverne vil du blive tildelt et spørgeskema, som har til formål at indsamle din holdning til programmet og hvorfor.

Til sidst vil jeg stille dig nogle generelle spørgsmål og nogle, som henvender sig til din udførsel af opgaverne eller din besvarelse af spørgeskemaet.

Udfør think-aloud eksempel - Giver det mening?

Nogle spørgsmål?

B.2 Task lists

B.2.1 Initial task list

1. Vælg Ultimaker 2+ printer modellen
2. Importer STL filen
3. Vend modellen korrekt
4. Gør modellen dobbelt så stor og placer den i det nederste venstre hjørne
5. Ændre infill mønstret til 'Triangular'
6. Indstil tykkelsen af skallen til 4 mm
7. Vælg lag tykkelsen der hedder 'Coarse'
8. Indstil infill densiteten for support strukturen til 25 procent
9. Put support på alle mulige overhangs, uanset vinklen på overhanget
10. Omdøb 3D modellen til 'Model'
11. Slice, og preview den slicede version af modellen og sikre dig at der er support på alle overhangs.
12. Tjek at infill mønstret korrekt
13. Eksporter filen og gem den på skrivebordet

B.2.2 Revised task list

English version

1. Select Ultimaker 2+ printer model
2. Import the model named 'Test model' to the workspace
3. The model is turned on the head, turn it correctly

4. Make the object twice as big
5. Move the model to the front of the printbed
6. Change the infill pattern to 'Triangular'
7. Set the shell thickness to 6 mm
8. Select 'Coarse' layer thickness
9. Set the infill density of support structures 25 percent
10. Apply support to all possible overhangs, no matter the overhang angle, except for the eyes
11. Rename the 3D model to 'My model'
12. Slice, and preview the sliced model and ensure that support is on all overhangs, except for the eyes
13. Check if the infill pattern is Triangular.
14. Export the file and save it

Danish version

1. Vælg Ultimaker 2+ printer modellen
2. Importer modellen, som hedder 'Test model' til programmet
3. Modellen vender på hovedet, vend den korrekt
4. Gør modellen dobbelt så stor
5. Flyt modellen hen til fronten af printpladen
6. Ændre infill mønstret til 'Triangular'
7. Indstil tykkelsen af skallen til 6 mm
8. Vælg lag tykkelsen der hedder 'Coarse'
9. Indstil infill densiteten for support strukturen til 25 procent
10. Put support på alle mulige overhangs, uanset vinklen på overhanget, på nær ved øjnene
11. Omdøb 3D modellen til 'Min model'
12. Slice, og preview den slicede version af modellen og sikre dig at der er support på alle overhangs pånær øjnene
13. Tjek at infill mønstret er Triangulært
14. Eksporter filen og gem den

B.2.3 Task list given to P2 and P3

English version

1. Select Ultimaker 2+ printer model
2. Import the model named 'Test model' to the workspace
3. The model is turned on the head, turn it correctly
4. Make the object twice as big

5. Move the model to the front of the printbed
6. Change the infill pattern to 'Triangular'
7. Set the shell thickness to 6 mm
8. Select 'Coarse' layer thickness
9. Set the infill density of support structures 25 percent
10. Apply support to all possible overhangs, no matter the overhang angle, except for the eyes
11. Rename the 3D model to 'My model'
12. Slice the model
13. Check if the infill pattern is Triangular, without clicking 'Show preview'.
14. Export the file and save it

Danish version

1. Vælg Ultimaker 2+ printer modellen
2. Importer modellen, som hedder 'Test model' til programmet
3. Modellen vender på hovedet, vend den korrekt
4. Gør modellen dobbelt så stor
5. Flyt modellen hen til fronten af printpladen
6. Ændre infill mønstret til 'Triangular'
7. Indstil tykkelsen af skallen til 6 mm
8. Vælg lag tykkelsen der hedder 'Coarse'
9. Indstil infill densiteten for support strukturen til 25 procent
10. Put support på alle mulige overhangs, uanset vinklen på overhanget, på nær ved øjnene
11. Omdøb 3D modellen til 'Min model'
12. Slice modellen
13. Tjek at infill mønstret er Triangulært, uden at trykke 'Show preview'
14. Eksporter filen og gem den

B.3 Usability issues

In this section all the identified usability issues are presented. Each usability issue is labeled with a number. Each issue is consists of a concise definition of the usability issue, followed by a description of the origin of the issue, included the task in which the issue occurred. Note that all statements from the participants have been translated from Danish to English. The original Danish statements can be seen in the annex within the folder 'Study2'

Usability issues identified for P1

1. Difficulty figuring out how to import an STL file.
   - (Task 2) In an attempt to import the STL file, P1 initial thought was to look under the 'File' drop-down menu in the menu bar. Here P1 clicked 'Open workspace file' without being able to locate the 'test model' file in the directory, as only workspace files (.rvw) were available. He then looks under 'Edit' afterwards. Then he looks in the left menu where the 'Add model' button is located, but fails to notice it. Finally P1 attempts to drag the file from the desktop into the program which works.

2. Initially unable to recognize that the model is selected.
   - (Task 3) P1 is not aware that the model is automatically selected from the beginning as P1 states at 00:03:04: *"I'm trying to select the model. I cannot do that, so now i am trying to go down here. Here i can see that it is under 'Objects'"*. He ended up continuing with the task at hand, probably without realizing that the object was selected all along.

3. Unable to correctly use the rotation menu.
   - (Task 3) P1 was not aware of how to use the + and - button under the rotate menu, as he simply wrote 180 degrees and pressed enter. He also preferred that click/draggable axis lines appeared on the object (mentioned in the debriefing as well).

4. Difficulty understanding how to rotate the model with the hover function.
   - (Task 3) P1 then discovered that an icon appeared when hovering the object, initially thinking that he could then drag the object around, but finally realizing that it was only by clicking on the object that it rotated. Also stating at 00:03:04: *"It was about trying things out"*. He later states during debriefing that he did not really understand the clicking function but managed to make it work anyways. He also mentions in the questionnaire (Q2.a) that it was cumbersome to rotate the model.

5. Unable to correctly resize the model due to bug.
   - (Task 4) P1 attempted to resize the model by clicking on the model and holding down the mouse in order to change the size by dragging. However a bug occurred which made the object as small as possible as soon as he clicked on it. (It appears that this bug can only occur the first time you attempt to scale the object by clicking on it, in a program instance.)

6. Identifies the wrong side as the front of the printbed.
   - (Task 5) P1 mistakenly believed that the front of the print bed was the direction in which the model was facing. The model is not oriented with the face towards the front of the printbed due to two things: (1) a bug specifically related to the Ultimaker 2+ printbed occurred, in which a model imported onto the Ultimaker 2+ printbed, is orientation 90 degrees clockwise, when looking from the top and down, in relation to the front of the printbed; and (2) when the model is rotated correctly, it is simultaneously facing the opposite way than when it was imported. These two factors means that the model is facing towards the right side of the printbed, thus causing confusion.

7. Unable to locate 'Printing settings' menu.
   - (Task 6) When trying to find infill settings, the first place P1 looked was in the printing profiles drop-down menu. Then he looked under 'File', then under the different manipulation functions (Flip and Support), then under 'Edit', then 'View', 'Plugins', 'Help', then Camera reset and REALview. Thereafter P1 gave up stating: 00:05:12: *"I don't think I can figure that out. No."* He was then told to try and press the gear icon besides the material profile selection, and figured out the rest himself.

8. Does not know how to see if support is enabled.
   - (Task 10) As the default value for the overhang angle is higher than any of the overhangs on the test model, no support is initially applied to the model even when customizable support is enabled. This resulted in P1 stating the following: 00:09:12: *"I'm looking for a place where I can turn it* ***(support)*** *on because I can't really see it on the model".* Support was in fact turned on, however nothing was applied due to the high overhang angle threshold. P1 is then looking inside printing settings menu for a way to enable support, initially without luck.

9. Expects preview to be shown immediately after model is sliced.
   - (Task 10) P1 tries preparing the model in order to check if support is applied, but he is not aware that he needs to press 'Show preview' in order to show the sliced version of the model.

10. Expects something to happen when selecting 'Support' in the object list.
    - (Task 10) P1 notices the 'Support' field in the list of objects, which he clicks on, expecting something to happen. But nothing does.
11. Expects support to be applied when clicking on the model with the paint on function.
    - (Task 10) P1 hovers the model and the cursor shows the paint on bubble for applying/removing support, but nothing happens when he clicks on the model, because the remove paint on function is selected by default and there is no support to remove. He does not find out that he can switch to the paint on function that applies support.
12. Does not expects printing settings to reset when clicking close window button.
    - (Task 10) P1 then proceeds to enter the printing settings menu once again where he finds the overhang angle settings and changes this value to 1. He then closes the settings window by clicking the close button in the top right corner, expecting the new overhang angle value to be applied, but in reality this value is reset because he does not press 'OK'.
13. Expects confirmation and closing of typing window when pressing enter after renaming object.
    - (Task 11) After renaming the model, P1 presses enter expecting the name to be confirmed by leaving the text field. However nothing happens. He later states during debriefing: 00:03:14: *"When you usually press Enter, it also usually clicks out of the text field, so you no longer continue typing when you press the keyboard."*
14. Unable to preview sliced model due to function causing the program to crash.
    - (Task 12) The task at this point included the instruction to 'show preview' which crashed the program.

Usability issues identified for P2

15 Difficulty figuring out how to import an STL file.

- (Task 2) P2 expects to be able to import files under the 'File' drop-down menu, stating the following: 00:01:27: *"It should be under the file menu."* He looks under the other drop-down menus.

16 Misinterprets meaning of 'Add model'.

- (Task 2) P2 then proceeds to hover the 'Add models' button which is the intended way to import objects. However he then says: 00:03:33: *"Is this a model? No, this must be for adding a new printer."* Thus he misinterprets the function of the button.

17 Initially unable to recognize that the model is selected.

- (Task 3) P2 is trying to figure out how to select the model, despite it already being selected.

18 Initially thought that 'Reset workspace' function was used to rotate model due to icon.

- (Task 3) P2 initially thought that reset workspace function was related to rotation, before reading the pop-up information box.

19 Initially selects the wrong side as the front of the printbed.

- (Task 5) P2 is unsure of which side of the printbed is the front. Later on in the test he realizes that the model is not placed in the front side of the printbed, thus attempting to correct this mistake. Due to the two factors described at usability issue 5, clicking 'Front' within the 'View' drop-down menu, shows the right side of the model, thus confusing P2. However he manages to move the model to the actual front of the printbed.

20 Misinterprets 'Prepare' button. Does not expect it to slice the model.

- (Task 6) P2 believes that in order to change infill pattern he needs to press prepare. He is surprised when the program starts to slice the object instead, stating: 00:09:40: *"It starts slicing right away, I didn't quite expect that."*

21 Difficulty locating 'Printing settings'.

- (Task 6) P2 then starts looking through the model manipulation functions (Support, Flip), then 'Edit'. He continues to look for it in the left section of the interface. He says that he is just looking around while hovering components in order to read the descriptions. After 4 and a half minute spend on the task he decides to click the gear icon which opens printing settings. In response to which aspects were the least easy to use (item 1.a from the questionnaire) he states *"Navigating to specific settings"*.

22 Unable to figure out what printing profile icon means due to bug not showing pop-up information box.

- (Task 6) P2 hovers the icon for printing profiles, but a bug occurs where no pop-up information box appears, despite being intended.

23 Unable to make the shell thickness 5mm.

- (Task 7) P2 tries to use the arrows to change the shell thickness to 5mm, but could not hit exactly 5mm as the thickness of each contour was set to .4mm (the goal of the task was changed from 6 to 5mm as the settings were not reset and therefore was already at 6mm). He then types 5.0, but the value is rounded down to 4.8 when he exits the text field.

24 Layer thickness does not show in search result.

- (Task 8) P2 tries to locate the layer thickness settings by using the search bar. This works but the only result was 'Custom layer thickness', which he then clicked. This caused minor confusion as it was the overall layer thickness that he was looking for.

25 Confusion due to search bar not automatically clearing search result.

- (Task 9) P2 wants to search for support infill density but the search bar is already filled with the search word from the previous task. He does however later figure out that he has to click the bar, which then highlights the words and then write the new search word.

26 Unable to locate 'support infill density' setting using the search feature.

- (Task 9) P2 searches for 'support infill density' without getting a result. The setting is only available for 'Advanced' or 'Expert' level settings, and he is currently on the 'Basic' mode. However even if he would have been on either of the two higher levels, he would not have gotten a result, as writing 'Support' excludes any search result, even though the search result is named: 'Infill density - Support parameters'.

27 Unable to locate 'support infill density' which is hidden inside 'Advanced' mode.

- (Task 9) P2 is not able to locate infill density for the support structure on his own. He receives a hint to click on the 'Advanced' button, which results in him seeing the now visible

'Support Structures' section.

28 Does not understand the term 'Contour' and is unable to find information about it.

- (Task 9) P2 does not know what contour means in the context of 3D printing. Furthermore there is no description of contour in the tooltip, unless one enters expert settings and hovers one of the contour related settings.

29 Difficulty understanding the tooltip for 'Overhang angle'.

- (Task 10) P2 does not think that the overhang angle setting is explained very clearly. He states the following: 00:22:01: *"It's all about finding what you set the angle relative to, which hasn't been explained very clearly here."*

30 Difficulty understanding the 'Printing setting' section UI.

- (Task 10) P2 clicks twice on the 'Support Structures' section but nothing happens as he is already in that section.

31 Unable to correctly use the rotate menu.

- (Task 10) P2 diverts from the current task, as he is not satisfied with the placement and orientation of the model in relation to the front of the printbed. He manages to move the model to the actual front of the printbed. He then attempts to rotate the model in order to face it towards the front of the printbed, but does so by typing the amount of degrees he wishes to turn it into the text field and presses enter. This does nothing.

32 Expects something to happen when clicking (already selected) 'Customizable support' button.

- (Task 10) P2 notices 'Customizable support' and clicks on it. Nothing happens though as the menu is already open. However he expected something to happen according to his statement: 00:27:20: *"There is something called customizable support, perhaps there is a chance that something is included underneath. Well, not much happened."*

33 Does not understand the highlighting on the model, which indicates that support is applied to the model.

- (Task 10) P2 figures out the button for erasing support from the model, but he does not believe that there is currently added any support to the model, and does therefore not try to erase the support from underneath the eyes of the model. He says the following: 00:27:20: *"Well, currently there hasn't been any support made. Otherwise, I would assume that I need* *to use this one* ***(erase button)*** *to remove them."* He then proceeds to prepare the preview in order for him to check if there is applied any support. This crashes the program.

34 Does not figure out how to orbit the camera angle.

- When RvP had to be restarted, P2 noticed that the moderator was able to orbit by right-clicking and dragging, stating: 00:32:57: *"Now you figured out something about rotating the screen that I hadn't quite figured out."*

35 Expects the model renaming option to be located in the 'Edit' drop-down menu or the menu which appears when right-clicking the model in the object list.

- (Task 11) P2 looks under 'Edit' in order to rename the model. He then right-clicks the model name in the object list.

36 Does not know how to show the collapsed object list.

- In an attempt to close the pop-up box from the right-click he accidentally collapses the menu with the list of objects, without noticing what happened. He is then told how to reopen the menu, and figures out how to rename the object.

37 Confusion regarding the type of preview shown when model is sliced.

- (Task 12) When P2 has sliced the model he states the following: 00:35:46: *"So it looks like it is showing some sort of preview. Some sort of... because it has placed a skirt around it* ***(the model)****, it seems."* It could seem as if P2 is slightly confused as to why the skirt/brim is shown but not the rest of the preview.

38 Unable to hide contour due to 'Show preview' function causing crash.

- (Task 13) P2 wants to be able to hide the shell in order to check the infill pattern, but is not able to due to the 'Show preview' feature not working.

39 Initially unsure what the 2D viewer is.

- (Task 13) P2 is initially unsure what the 2D preview screen in the right side of the screen is, as only a small triangle of the top layers of the model is showing, it is difficult to make out what it is.

40 Expects to be able to click and jump to the middle of the 2D preview slider.

- (Task 13) P2 tries to click in the middle of the 2D preview layer selection slider. This would indicate that he tries to preview some of the middle layers, however the slider only steps down one layer at a time.

41 Does not realize that the orange material in the 2D preview represents support material.

- (Task 13) While P2 is looking at the 2D preview (where support is showed) he states the following: 00:37:41: *"We haven't seen any sign of any support or anything."* This shows that he does not recognize the orange material visualized in the 2D preview as support. The screen size of the laptop reduces the size of the 2D preview window, so that the checkboxes with the different material types are hidden. However there would still be no color coding, indicating that the orange material depicted should represent support.

42 Expects the initial name of the G-Code to be the same as the renamed model name.

- (Task 14) P2 is confused as to why the initial name of the G-Code is the old file name, and not the name he has given the file during task 11. He states the following: 00:38:07: *Well, now it's called 'test model', after I've just given it a new name.*

Usability issues identified for P3

43 Expects clearer indication of the printer model drop-down menu.

- (Task 1) P3 states: 00:00:21: *Actually, there's nothing that indicates **(printer models)** other than, yes, I can indeed see that it **(the logo)** resembles a printer.*

44 Difficulty figuring out how to import an STL file.

- (Task 2) In order to import a model P3 starts by going to 'File'. Then he clicks 'Import settings' (where STL files are not visible). He then starts looking around different directories, but is then told that it is located on the desktop. He then closes the window and tries 'Open workspace file' instead (where STL files are not visible). He then minimizes RvP and locates the file on the actual desktop. He double clicks the file, which then imports the file to the program. During debriefing he states: 00:01:53: *"The thing about if you need to find the file, you should just be able to go to "File" and then "Open." And then there shouldn't be... then it should say "Open File" instead of "Add Model" because then it means... it can get a little confusing."*

45 Difficulty figuring out which axis to correctly turn the model on.

- (Task 3) P3 is unsure of which axis to turn the model on, stating: 00:04:40: *And then I need to rotate it on... maybe it's on the x-axis or maybe it's on the y-axis..*

46 Difficulty changing camera perspective.

- (Task 3) After rotating the model correct, P3 wants to change the viewing perspective in order to check that it is oriented correctly. He struggles to find out how to change viewing angle. He states: 00:04:40: *"Can I grab hold here?"* while trying to change perspective by left-clicking on the workspace and dragging, which does nothing. If he right-clicked and dragged instead he would have discovered the feature he was looking for. He then discovers the 'View' drop-down menu and figures out how to change perspective. Afterwards he says that he would prefer the selection of viewing angles to be located on the main interface instead of inside a drop-down menu. He later during debriefing states: 00:00:15: *"...the thing about having to go up in 'View' to see from the top/bottom could possibly have been available outside **(of the 'View' drop-down menu)**, and then there were maybe some other things out there, which maybe didn't need to be there unless you needed them, so it's like they popped up."*

47 Thought that 'Reset workspace' function was used to rotate model due to icon.

- (Task 3) P3 believes that the workspace reset function is used to rotate the model. He states the following: 00:04:40: *If I turn here, then I believe I'm turning the actual figure itself.* while hovering the reset workspace button.

48 Insinuates that easier ways of scaling exists, than those in RvP.

- (Task 4) P3 initially discovers how to actually scale the model, but then decides to look for an easier way of scaling the model. He states the following: 00:07:40: *"I'm just trying to make sure there isn't an easier way."* By saying this he is likely insinuating that the current way of scaling is more difficult than what he is used to.

49 Identifies the wrong side as the front of the printbed.

- (Task 5) P3 successfully moves the object, but to the right side of the printbed, as the model is facing in that direction. This is due to the two factors described at usability issue 5.

50 Initially unable to recognize that the model is selected.

- (Task 5) P3 mentions that: 00:09:21: *"The only thing missing is a indication that the figure is selected and ready to be moved."* He then proceeds to explain how other programs have an x, y and z axis depicted inside the object and that those can be used to move the object on the respective axis.

51 Difficulty locating 'Printing settings'.

- (Task 6) When looking for infill pattern settings, P3 begins by checking the different drop-down lists. Then he looks through the rest of the left side menu, and proceeds to check through the drop-down menus in the menu bar. He then discovers the 'Printer settings' window by clicking the top gear icon. He then closes this window and clicks the gear icon below, opening 'Printing settings'. As a response to which aspects were the least easy to use (item 1.a from the questionnaire) he states *"Finding settings"* as one of those.

52 Expects a click-and-hold feature to be available for setting arrows.

- (Task 10) P3 uses the bottom arrow to lower the overhang angle value. He starts by clicking multiple times and then clicks and holds, expecting the value to continue to drop when holding down the button. However nothing happens, so he continues to click on the button.

53 Difficulty understanding the tooltip for 'Overhang angle'.

- (Task 10) P3 contemplates for 35 seconds on what the value should be in order to apply support to all overhangs.

54 Expects something to happen when clicking (already selected) 'Customizable support' button.

- (Task 10) P3 clicks on 'Customizable support' expecting something to happen, however nothing happens as it is already selected.

55 Does not initially figure out how to orbit the camera angle.

- (Task 10) P3 accidentally discovers how to orbit the camera angle by right-clicking and dragging, when he was trying to remove support from the eyes of the model. He was not aware that this feature existed, stating: 00:17:25: *"Oh, well now I figured out how to move it. You just need to know that you have to right-click."*

56 Expects a single click on the object (in the object list) to activate renaming.

- (Task 11) P3 looks in the 'File' drop-down menu. After not finding anything there, he finds the object list, and clicks once on the model. Later in the following debriefing he explains how he expects a single click to open up the renaming option: 00:04:29: *"Normally, you're supposed to be able to click once and the you should be able to do it, but you have to double-click. That's probably what caused the confusion."*

57 Does not notice active renaming indicator (blinking line).

- (Task 11) Afterwards P3 actually double-clicks but does not notice that it is then possible to change the name.

58 Expects the model renaming option to be located in the menu which appears when right-clicking the model in the object list.

- (Task 11) P1 then tries to right-click on the model in the object list, in order to rename it.

59 Finds the term 'Prepare' inappropriate for a slicing button.

- (Task 12) When trying to slice the model P3 starts by looking in the left-side menu. Then he notices the 'Prepare' button and states the following: *"One would almost assume that it was the one called 'Prepare'. So I press on it. Then it slices. They could have written "prepare slicing" or "slicing" there."* In the debriefing he states: 00:01:00: *"...a button like the one that says "Prepare" should say "Slice." Because what am I preparing it for? I'm* *not quite sure."*

60 Is not aware of the 2D viewer.

- (Task 13) P3 starts looking for the infill settings in order to check that it is set correct. However this is a slight misunderstanding of the task as he has to assure that the sliced version of the model has the correct infill pattern. He then gets informed that he has to check the infill pattern on the model. He continues to look for the 'Printing settings' window and opens it. Here he checks that the infill pattern is still set to 'Triangular'. He does not discover the 2D viewer feature.

61 Difficulty understanding the 'Printing setting' section UI.

- (Task 13) When P3 was checking the infill pattern he mentions that: *If I go to the model or material properties, it says 'infill', but it says 'triangular', so that's probably how I can check it.*. This could indicate that he does not know which of the two sections he is currently in.

Annex C

The annex contains the following folders and files.

- Study1
  - Study1_questionnaire_responses.pdf
- Study2
  - Study2_questionnaire_responses.pdf
  - P1_thinkaloud_transcription.pdf
  - P2_thinkaloud_transcription.pdf
  - P3_thinkaloud_transcription.pdf
  - P1_debriefing_transcription.pdf
  - P3_debriefing_transcription.pdf