How many participants are needed when usability testing physical products?

An analysis of data collected from usability tests conducted on physical products

Pontus Henstam

March 6, 2018
Master's Thesis in Interaction Technology and Design, 30 credits
Supervisor at Umeå University: Shafiq Urréhman
Supervisor at RISE: Mattias Widerstedt

Examiner: Thomas Mejtoft

Umeå University
Department of Applied Physics and Electronics
SE-901 87 UMEÅ
SWEDEN

Abstract

Testing a product on users before releasing it on the market can be very rewarding but also costly for companies. Therefore testing products on just the right number of users, that will be enough to include the benefits of the tests while keeping down the costs, would be most beneficial. A common advice means that five participants are enough to include in such tests. This advice is based on research mainly from testing computer-based interfaces on users. Though, how well this advice can be applied when testing physical products on users is less investigated.

This thesis has investigated how many participants that are needed when testing physical products on users. A literature study and an analysis of data collected from physical products tested on users were conducted. The results show that using five participants when testing physical products on users cannot be counted on to be enough. The results also show that the number of participants to use, when testing physical products on users, vary.

Contents

Chapter 1

Introduction

If a product is easy to use, easy to understand, and effectively performs what is expected of it, people using the product will have a bigger chance of continuing to do so. A product with such qualities benefits the users since they will more likely have a good experience using the product, but also the companies developing the product since it will make their product increase in demand on the market.

Testing a product on users, to increase the chance that the users will have a good experience using it before release can be vital in how popular the product will become on the market. Many companies are of course aware of this and therefore perform such tests on their products. Though performing tests on users can be done in different ways, and there are different opinions on how it is best done.

An important decision that has to be made when testing products on users, that there are different opinions about, is to decide the number of users to test the product on. A common advice [15] means that five users are enough to test products on and that after testing on five users 85% of the product's problems will be found. This advice is based on research mainly from testing computer-based interfaces on users [18]. Though, how well this advice can be applied when testing physical products on users is less investigated.

During this thesis, conducted at the research institute RISE, an investigation of the number of users needed when testing physical products on users was performed. The investigation was performed through a literature study and through a statistical analysis of data collected from physical products tested on users.

1.1 Usability

Usability is a word for describing how well something can be used. There are slightly different definitions of usability, and in this thesis the ISO 9241-11 definition of usability is used.

> *"Usability: Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."*
>
> [6]

For this definition to make sense, the definitions of effectiveness, efficiency, satisfaction and context of use have to be defined as well. The following terms are defined in ISO 9241-11.

> *"Effectiveness: Accuracy and completeness with which users achieve specified goals."*
>
> *"Efficiency: Resources expended in relation to the accuracy and completeness with which users achieve goals."*
>
> *"Satisfaction: Freedom from discomfort, and positive attitudes towards the use of the product."*
>
> *"Context of use: Users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a product is used."*
>
> [6]

1.2 Usability testing

Usability testing is a method, involving users, for evaluating a product's, or a system's, usability. Following are two definitions of usability testing.

> *"Usability testing is the practice of testing how easy a design is to use on a group of representative users. It usually involves observing users as they attempt to complete tasks and can be done for different types of designs, from user interfaces to physical products. It is often conducted repeatedly, from early development until a product's release."*
>
> [4]

> *"While there can be wide variations in where and how you conduct a usability test, every usability test shares these five characteristics:*
>
> *1. The primary goal is to improve the usability of a product. For each test, you also have more specific goals and concerns that you articulate when planning the test.*
>
> *2. The participants represent real users.*
>
> *3. The participants do real tasks.*
>
> *4. You observe and record what participants do and say.*
>
> *5. You analyze the data, diagnose the real problems, and recommend changes to fix those problems."*
>
> [1]

1.2.1 Usability testing or user testing

User testing is an expression that sometimes has the same meaning as usability testing, and sometimes not. In this thesis user testing and usability testing will both refer to the above definitions of usability testing.

1.2.2 A common method used in usability testing

A common method that can be, and often is, applied in usability testing is called the think aloud method or the thinking aloud test. Following is a description of this method.

> *"In a thinking aloud test, you ask test participants to use the system while continuously thinking out loud — that is, simply verbalizing their thoughts as they move through the user interface."*
>
> [17]

1.2.3 Other methods for evaluating usability

Other common methods for evaluating usability besides usability testing, that will be mentioned in this thesis report, are heuristic evaluation and cognitive walk-through.

Following is a description of the method heuristic evaluation.

> *"Heuristic evaluation is a usability engineering method for finding the usability problems in a user interface design so that they can be attended to as part of an iterative design process. Heuristic evaluation involves having a small set of evaluators examine the interface and judge its compliance with recognized usability principles (the "heuristics")."*
>
> [14]

Following is a description of the method cognitive walk-through.

> *"An inspection method for evaluating the design of a user interface, with special attention to how well the interface supports exploratory learning, i.e., first-time use without formal training. The evaluation is done by having a group of evaluators go step-by-step through commonly used tasks. It can be performed by evaluators in the early stages of design, before performance testing is possible."*
>
> [21]

1.3 RISE

RISE stands for Research Institutes of Sweden and is a network of research institutes that collaborates with academia, industry, and the public sector. RISE has 2,200 employees that support and promotes all manner of innovative processes and has roughly 100 testbeds and demonstration facilities that are instrumental in developing the future-proofing of products, technologies, and services. RISE is fully owned by the Swedish state.

1.3.1 Usability testing at RISE

Usability researchers at RISE perform usability tests in a usability test lab in Sweden's capital Stockholm. The usability test lab is adapted for testing and evaluating the usability of products through usability testing. The purpose of the lab is to help companies, that want to develop functional and easy-to-use products, by performing usability tests as a service. The usability tests can be performed on concepts, products under development, a product already on the market or a competitor's product.

The usability researchers working at the usability test lab has access to materials and furniture that can be used to create a context realistic to the environment the tested products usually are used in. The lab is also equipped with cameras, screens, microphones and two-way mirrors to create optimal possibilities for documenting the usability tests.

1.4 Problem statement

The decision of what number of participants to use in a usability test is a decision that has to be made by anyone who decides to do a usability test. Since the decision has a considerable impact on finding problems with the product and also will affect the cost in terms of time and resources, this is an important decision. Too few participants used can lack identifying usability problems, and too many participants used can mean high costs in the form of time and resources. An optimal number of participants, making sure to discover what parts of the product that can be improved while keeping down the costs, would be most beneficial.

A recent article [16] has re-raised an argument that five participants will be enough to use, and will find 85% of the problems, in usability tests. The research [18] behind this argument of using five participants is based on data collected mainly from usability tests performed on computer-based interfaces. Though, the usability researchers at RISE perform usability tests on a high number of physical products. Since physical products can be quite different from computer-based interfaces, a question was raised regarding what number of participants that is needed particularly when usability testing psychical products.

1.5 Objectives

This thesis will investigate following questions:

- Does the research indicating that five participants are enough to find 85% of the problems in usability tests apply to usability tests on physical products?
- How many participants are needed when usability testing physical products?

1.6 Limitations

- All the data analyzed during this thesis, collected through usability tests on physical products, were not collected on all types of physical products, only home appliances.
- All the collected data is from usability tests conducted in Sweden, with Swedish citizens as the main participants.
- All the collected data is from usability tests conducted on either 8, 9 or 10 test participants.
- All the collected data is from 16 usability tests.

Chapter 2

Theory

This section investigates previous research regarding the number of participants needed in usability tests.

A common advice that has been around for 25 years means that using five test participants in a usability test will discover 85% of the usability problems that there are to find. This advice is based on studies by Nielsen & Landauer [18]. A slightly older advice means that using four or five test participants in a usability test will discover more than 80% of the usability problems that there is to find. This advice is based on studies by Virzi [23]. Though, later studies [20] [24] [2] [5] has come to results that do not support the advice of using five test participants, and means that there are risks with following the advice. Later studies [9] also argue that the number of participants needed in usability tests will vary from case to case.

2.1 Research that supports using 5 participants

2.1.1 Virzi

In 1992 Robert A. Virzi reported findings of the use of test participants in usability tests [23]. His three main findings were that:

- 80% of the usability problems are detected with four or five participants.
- Additional participants are less and less likely to reveal new information.
- The most severe usability problems are likely to have been detected in the first few participants.

He came to these findings by conducting and evaluating three usability testing experiments, where the numbers of test participants being used were compared to the proportion of usability problems being identified.

Virzi proposed that the rate of discovered usability problems follows a geometric series described by the formula

$d = 1 - (1 - p)^n$

where d is the rate of discovered problems, p is the probability of detecting a problem, and n is the number of test participants being used.

So for example, if the probability of detecting a problem is 30% (p=0.3) and the number of test participants being used is five (n=5), then approximately 83% (d≈0.83) of the problems would have been found, according to the formula.

In Virzi's experiments, the probability of detecting a problem (p) was obtained by measuring the mean probability of detecting a problem between the participants. The mean probability of detecting a problem between the participants was, in Virzi's three experiments, measured to be 32%, 36%, and 42% respectively. Using these values in the formula as the probability of detecting a problem (p) indicates that more than 80% of the problems are found after four or five test participants. This, according to Virzi, also indicates that, in usability testing, over 80% of usability problems will be detected after four or five test participants.

The tree experiments analyzed by Virzi

In the first experiment, a manual to a voice mail system was evaluated through a thinking aloud test [7]. 12 participants with no previous experience, aged 18 to 65, were recruited through a local newspaper advertisement. With only the manual available the participants got to describe how the system would respond in three different scenarios. Totally 13 different problems were identified by the 12 participants. The number of problems encountered by each participant varied between one and eight. The mean chance of encountering a problem by the participants was 32%.

The second experiment was evaluated through a thinking aloud test of a computer-based appointment calendar [22]. The calendar was evaluated by students, under Virzi's supervision, as a part of a design course at Tufts University. The test participants were 20 students that were recruited from an introductory psychology course. Only those with little or no computer experience and no experience with computer-based appointment calendars were included in the study. Each participant performed 21 tasks in a specific order. Totally 40 different problems were identified by the 20 participants. The number of problems encountered by each participant varied between 8 and 19. The mean chance of encountering a problem by the participant was 36%. Each problem's severity was rated from

one to seven by the experimenters. The results suggested that the more severe a problem is, the more likely it is to find it within the first few participants.

In the third experiment a voice response system, together with the design of the supporting documents, was evaluated. The system was interacted with by pressing the keys on a phone which generated prerecorded messages with information. 20 test participants were recruited through an advertisement in a local newspaper. Each participant performed seven tasks. Totally 17 different problems were identified by the 20 participants. The number of problems encountered by each participant varied between 3 and 12. The mean chance of encountering a problem by the participants was 42%. The problems were rated either low, medium or high by their impact on usability. As in the second experiment, the results suggested that the more severe a problem is the more likely it is to find it within the first few participants.

2.1.2 Nielsen & Landauer

In 1993 Nielsen & Landauer published a paper [18] where they established that a mathematical Poisson model quite well describes the finding of usability problems in user testing and heuristic evaluation. Data from 11 projects, containing five user tests and six heuristic evaluations, were evaluated in the study. The same data was later referred to in an article from 2000 by Nielsen [15], where he writes about why you only need to test with five users in user tests. Nielsen writes that the number of usability problems found in a user test is given by the formula

Number of problems found = $N(1 - (1 - L)^n)$

where N is the total number of usability problems there is to find in the design, L is the proportion of usability problems discovered while testing a single user, and n is the number of test participants being used.

So for example, if the proportion of usability problems discovered while testing a single user is 30% (L=0.3), and let us say that the amount of problems there is to find is 100 (N=100), and that the number of test participants being used is five (n=5), then approximately 83 problems should have been found, according to the formula.

Nielsen also states that the typical value, for the proportion of usability problems discovered in user tests while testing a single user, is 31% (L=0.31), since this is the averaged number Nielsen & Landauer found across the 11 projects they studied. Using the value L=0.31% in the formula indicates that approximately 85% of the problems are found after five test participants. This, according to Nielsen, indicates that, in user testing, 85% of usability problems will be detected after five test participants.

The 11 projects evaluated by Nielsen & Landauer through user testing and heuristic evaluation

The first user testing project evaluated a computer-based office system containing a word processor, a mail application, a calendar application and a spreadsheet [8]. 15 participants, with different background and previous experience using a mouse, were recruited. Each participant performed ten scenarios in mixed orders, except for the first scenario. Totally 145 problems were observed. The mean chance of encountering a problem by the participants was 16%.

The second user testing project evaluated is the same project that Virzi evaluated in his second experiment in his paper from 1992 [22]. The system evaluated was a computer-based appointment calendar [22]. 20 students were recruited as participants from an introductory psychology course. Only those with little or no computer experience and no experience with computer-based appointment calendars were included in the study. Each participant performed 21 tasks in a specific order. Totally 40 problems were identified by the 20 participants. The mean chance of encountering a problem by the participants was 36%.

The third user testing project evaluated was a computer-based commercial word processor [13]. The test was performed through a think aloud method by computers science students taking a class in interface design. The computer science students were given three hours of lectures of the think aloud method before they would run an experiment on their own. 24 students were recruited as participants. Totally nine problems were observed. The mean chance of encountering a problem by the participants was 30%.

The fourth user testing project evaluated was a computer-based outliner [13]. It is an application for manipulating and reorganizing nested hierarchical outlines of, for example, articles and books. The test was performed at the same time as the third project, by the same class of computers science students taking a class in interface design. 30 students were recruited as participants. Totally 14 problems were identified. The mean chance of encountering a problem by the participants was 28%.

The fifth user testing project evaluated was a computer-based bibliographic database [25]. The database was evaluated through a think aloud method. The evaluators were postgraduates in the first term of a conversion course, training them to be software engineers. The evaluators had no human factors training before participating in the study. The evaluators were divided into 13 groups of two, except for one group that was three. Each team was provided with a fifth-year undergraduate psychologist who acted as their user. Each user performed six specific tasks. Totally 29 problems were identified. The mean chance of encountering a problem by the participants was 31%.

The first heuristic evaluation was performed on ten screen dumps of a videotex system [19]. 37 participants, that were no usability experts, got to perform heuristic evaluations on the system. To measure the performance of the par-

ticipants, the number of problems they found were measured to the number of problems found by the authors that had performed a heuristic evaluation of the system as well. Totally 52 usability problems were known in the system. The mean percentage of the known problems found by the participants was 51%.

The second heuristic evaluation was performed on a system designed and constructed for the purpose of the test [11]. By dialing into the system, the system provided names and addresses of subscribers given the telephone numbers of the subscribers. 77 participants, that were no usability experts, got to perform heuristic evaluations on the system. To measure the performance of the participants, the number of problems they found were measured to the number of problems found by the authors that had performed a heuristic evaluation of the system as well. Totally 30 usability problems were known in the system. The mean percentage of the known problems found by the participants was 38%.

The third heuristic evaluation was performed on a voice response system [12] that were used for handling banking accounts. Heuristic evaluations were performed by three groups divided by their usability expertise. One group contained 31 "novice" evaluators that had no usability expertise, one group contained 19 "regular" usability specialists, and one group contained 14 "double" usability specialists who also had experience with the particular kind of interface being evaluated. Totally 16 problems were found. The mean percentage of problems found by "novice" evaluators was 22%. The mean percentage of problems found by "regular" usability specialists was 41%. The mean percentage of problems found by "double" usability specialists was 60%.

The fourth heuristic evaluation was performed on a voice response system [19] that could give information about the customer's bank account balances, current foreign currency exchange rates, etc. 34 participants, that were students in a user interface design course, got to perform heuristic evaluations on the system. To measure the performance of the participants, the number of problems they found were measured to the number of problems found by the authors that had performed a heuristic evaluation of the system as well. Totally 48 usability problems were known in the system. The mean percentage of the known problems found by the participants was 26%.

The fifth heuristic evaluation was performed on a voice response system [19] that could provide commuters in Copenhagen with information about bus routes. 34 participants, that were the same interface design students as in the fourth heuristic evaluation, got to perform heuristic evaluations on the system. To measure the performance of the participants, the number of problems they found were measured to the number of problems found by the authors that had performed a heuristic evaluation of the system as well. Totally 34 usability problems were known in the system. The mean percentage of the known problems found by the participants was 19%.

The sixth heuristic evaluation was performed on a prototype system for internal telephone company use [10]. 11 participants, that were usability specialists but that had not been involved with the design of the interface, got to perform heuristic evaluations on the system. Totally 40 usability problems were found. The mean percentage of the problems found by the participants was 29%.

2.2 Research that does not support using 5 participants

2.2.1 Spool & Schroeder

In 2001 Jared Spool & Will Schroeder reported results [20], by evaluating four websites through usability testing, using the same type of formula as Nielsen & Landauer, that they would only find 35% of the problems if they would use five test participants. Unlike Nielsen & Landauer, that got the mean probability of finding a problem to be 31%, Spool & Schroeder got the mean probability to be never higher than 16%.

The usability tests conducted by Spool & Schroeder

The usability tests were conducted on four websites. Three of the websites primarily sold music CDs, movie videos and DVDs. The fourth website sold electronic gadgets. The test participants all had previous experience buying these sorts of products online. 18 tests were performed on the first website, 18 tests were performed on the second website, 7 tests were performed on the third website, and 6 tests were performed on the fourth website. Each participant performed the same one task, to describe an item the participant wanted and then buy it from the website. The only difference between the tests were the objects each participant attempted to buy. The first website had totally 64 different problems identified, the second website had totally 59 different problems identified, the third website had totally 20 different problems identified, and the fourth website had totally 23 different problems identified.

2.2.2 Woolrych & Cockton

In 2001 Woolrych & Cockton reported findings [24] critical to Nielsen's claim that five users are enough to use in user testing. By re-examining some of their own data, Woolrych & Cockton showed that there is a risk with using Nielsen & Landauer's formula and with assuming that the participants mean probability of encountering a problem is 31%. Woolrych & Cockton found the mean probability of encountering a problem in their own data to be 43%, which would mean, according to Nielsen & Landauer's formula, that 81% of the problems

would be found after three participants. Though, Woolrych & Cockton also found that the probability of encountering a problem for each participant varied between 25% and 62,5%. If all of the participants would have had the same probability of encountering a problem as the one with the highest, then two participants would be needed to find approximately 86% of the problems. But if all of the participants would have had the same probability of encountering a problem as the one with the lowest, then seven participants would be needed to find approximately 86% of the problems. Also, by looking at the severity ratings of the problems of their own data, Woolrych & Cockton could see that depending on which group of, in this case five, test participants they would use the severity of the problems would have got different ratings.

The user test conducted by Woolrych & Cockton

The data were collected through user testing in a study that evaluated the scope and accuracy of the method heuristic evaluation. The user test evaluated a computer-based drawing editor contained in the computer program Microsoft PowerPoint from 1995. Twelve participants with varying computing experience were recruited. Totally 16 problems were identified.

2.2.3 Faulkner

In 2003 Laura Faulkner reported findings [2] about the risk with assuming that five participants are enough to use when conducting usability tests. By conducting and evaluating usability tests of a web-based application on 60 users, 45 problems were found. Faulkner found that the average percentage of problems found when using five participants was close to what Virzi and Nielsen found. From 100 trials of five participants, the average percentage of problems found was 85%, though, with a standard deviation of 9.3 and a 95% confidence interval of ±18.5%. By using a program that could pick out group combinations of the participants, Faulkner found that group combinations of five participants could find significantly less percentage of the problems than the average 85%. After picking out 100 trials of groups of five participants, the percentage of problems found ranged from 55% to 99%. Faulkner could also see that by increasing the size of the groups picked out, the variance of the percentage of problems found reduced and the odds of finding the problems increased markedly.

> *"Merely by chance, a practitioner could encounter a 5-user sample that would reveal only 55% of the problems or perhaps fewer, but, on the basis of the 5-user assumption, still believe that the users found 85%"*

The usability test conducted by Faulkner

The usability test was conducted on a web-based employee timesheet application. 60 participants, with different levels of experience with computers and with the application, were given one task, to complete a weekly timesheet. To complete the task the participants were handled the same predetermined data to be entered into the application. All usability tests were conducted in the same location using the same computer. Totally 45 different problems were identified.

2.2.4 Hwang & Salvendy

In 2010 Wonil Hwang & Gavriel Salvendy reported findings [5] about how many participants one should test with while evaluating usability. By Collecting data from usability studies from different sources and evaluating the data, with focus on the tests overall discovery rate, they came up with that using 10 ± 2 participants will find 80% of the problems. The data collected were from 27 different usability studies. The usability studies were conducted either through think aloud methods, heuristic evaluations or cognitive walkthrough. Hwang & Salvendy used a linear regression analysis based on 36 data points from the 27 experiments. According to Hwang & Salvendy, the results of their analysis indicates that, when think aloud methods is used 9 participants will find 80% of the problems, when heuristic evaluations is used 8 evaluators will find 80% of the problems, and when cognitive walkthrough is used 11 evaluators will find 80% of the problems. Though, Hwang & Salvendy found deviations from three points of the 36 data points used in their regression analysis. One of the points, from a heuristic evaluation, showed that 68.3% of the problems were found after two users which were seen as a too high overall discovery rate compared to the number of evaluators. Hwang's & Salvendy's explanation of this case was that the usability specialists evaluated relatively simple interfaces for enough duration of evaluation. Two of the points, one from a heuristic evaluation and one from a cognitive walkthrough, showed that 8% of the problems were found after three users which were seen as a too low overall discovery rate compared to the number of evaluators. Hwang's & Salvendy's explanation of these cases was that non-experts evaluated interfaces for a relatively short duration and reported usability problems in special formats, such as automated report, structured forms, and diary.

2.2.5 Macefield

In 2009 Ritch Macefield published a paper [9], using secondary literature, for helping practitioners to specify the number of participants to use in usability studies. The paper was also written to help practitioners understand and articulate the bases, risks, and implications of any specification.

> *"There are different types of usability studies and, similarly, studies take place in a wide variety of contexts. This means that we must be careful when applying any particular research based advice."*

> *"In summary, problems with interfaces are often fuzzy and subjective in nature. Indeed, these properties of problems are one reason why there is so much controversy as to what statistical methods and thinking best applies to these studies."*

Macefield writes that depending on how critical it is to find the problems with a system, the more participants should be used. Examples of these situations, given by Macefield, are:

- *"work in highly secure environments e.g., the military"*
- *"work involving safety critical applications e.g., air traffic control and the emergency services"*
- *"where the socio-economic or political stakes are high e.g., with governmental applications"*
- *"work with enterprise critical applications where the financial stakes are high e.g., on-line banking and major e-commerce systems"*
- *"when a previous study, using a small(er) participant group size, has yielded suspect or inconclusive results"*

According to Macefield, the complexity of a study is also something that decides how many participants will be necessary to test with. Macefield writes that the variation of complexity in different studies is a key reason to why researchers have come to different conclusions about how many participants to use. For example, Macefield means that Nielsen's argument that five participants are optimal is based on relatively simple studies using quite closed and specific tasks, while Spool & Schroeder, who found that five participants were not nearly enough, conducted more complex studies using very open tasks. Macefield writes that there are factors that can aid us in the challenge of assessing a study's complexity and that an increase in the following factors typically increases a study's complexity.

- *"scope of the system(s) being used"*
- *"complexity of the system(s) being used"*
- *"(potential) pervasiveness of the system"*
- *"scope, complexity, and openness of the tasks(s) being performed"*

- *"number and complexity of the metrics being used"*
- *"degree of diversity across the facilitators being used"*
- *"(potential) degree of diversity across the target user group"*
- *"degree of diversity across the study participants"*
- *"degree of potential for contaminating experimental effects in the study"*
- *"degree to which the study participants reflect the target user group, particularly in terms of what relevant knowledge they will bring to the interactions"*

Another factor Macefield points out that can affect a study's complexity is any particular training given to the participants before the test. Depending on how consistent and well reflecting the training is for the target users the complexity of the system can both increase or decrease.

Another thing to take into consideration when deciding how many participants to use, according to Macefield, is how much of an early conceptual prototype the system is. Macefield means that it is easy to argue that early prototypes are more likely to contain more severe problems and that this significantly increases the likelihood that fewer participants will be required. Macefield also writes that another factor that points to using fewer participants in early prototypes is that early prototypes typically are low fidelity, which typically makes them capable of only supporting simple constrained tasks. Macefield means that this makes it easy to argue that early prototypes are less complex, which makes it reasonable to why Nielsen's advice, based on relatively simple studies, supports using fewer participants.

By going through the different span of recommendations from earlier research Macefield means that it is easy to argue that for most studies it is valid to use a group of 3-20 participants, with 5-10 participants being a sensible baseline range. Macefield also means that the number of participants used should be increased with the study's criticality and the study's complexity. In cases when the study is related to early conceptual prototype testing, Macefield means that there will be typical factors that point the optimal group size towards the lower end of the range. Macefield also writes that specification of the participant group size for a usability study remains a matter of opinion and debate.

Chapter 3

The data collected from usability tests conducted by RISE

This section will describe how the data, used for analysis in this thesis, were collected by usability testing physical products at the usability test lab by the usability researchers at RISE.

3.1 How the data were collected

The data were collected by usability testing physical products using the think aloud method. Each usability test was performed on either 8, 9 or 10 participants. The products were tested by one participant at a time.

3.1.1 The procedure of the usability tests

Each test session starts with an introduction of the test room where the product will be tested. In the test room a usability researcher from RISE verbally gives the participant predetermined tasks to perform while encouraging the participant to explain his or her thoughts while interacting with the product.

The usability test is observed by a usability researcher who takes notes while watching through a two-way mirror window and on live streamed recordings on monitors. Several cameras are being used at the same time to catch both facial expressions on the participants and interaction with the product. Two or sometimes three camera angles are being displayed on the monitors. All

streamed video materials are being recorded, both sound and picture, to enable the usability researchers to go back and observe specific parts again.

When all participants have performed the usability test, the usability researchers sum up the collected materials from the tests in spreadsheets called observation files. These files contain information of which participants encountered which problems, and calculations of each problem's severity rating, together with additional information.

3.2 The products tested

The data were collected by usability testing physical products, more specifically home appliances, only. These products were either washing machines, dishwashers, dryers, vacuum cleaners, inductions hobs, ovens, fridges or freezers.

3.3 The participants

The participants recruited for the usability tests were picked carefully to make sure they would be within the target group for the product tested.

3.4 The usability researchers

The usability researchers working at the usability test lab are well educated and experienced in the methods used for usability testing. Their educational background is from either engineering focused on design, industrial design or behavioral science.

3.5 Severity ratings of the problems

Each problem identified in the usability tests are rated, by the usability researchers, by its degree of severity using a specific method. The method cannot be explained in this thesis report due to confidentiality agreements, but what can be said is that the method focuses on effectiveness, efficiency, and satisfaction, which are defined in the introduction of this thesis report. Each identified problem is rated either low, medium or high by its severity. The higher the severity a problem is rated, the more important it is to solve that problem by adjusting or redesigning the tested product.

Chapter 4

Methods

This section will describe the methods used, and why they were decided to be used, in this thesis.

4.1 Study of the usability testing methods at the usability test lab

To get an insight in how the data from the usability tests at the usability test lab were collected, two study days at the usability test lab, during usability testing sessions, were conducted. During the two days at the lab a vacuum cleaner was usability tested on nine participants.

To get more specific information in the procedure of performing usability tests at the usability test lab a study of documents, describing the procedures and methods used at the lab, were conducted. During the study of the documents usability researchers working at the lab were available to help with answering any questions about the information in the documents.

The observation files, which are the files where the observations during the usability tests are documented, was also studied to understand the format of the collected data.

4.2 Literature Study

To find a method for measuring how many participants that are needed when usability testing physical products, based on the data collected at the usability test lab, a literature study was performed. The literature consisted of scientific articles regarding the number of participants to use in usability testing. The

literature studied were scientific articles found on Google Scholar and on the website of Umeå University Library.

4.3 Calculating method

Based on the literature study and the study of the data collected from the usability tests at the usability test lab, a method for calculating how many participants that would have been needed to find the identified problems in the usability tests were developed. This calculating method is inspired by the method used by Faulkner [2] when showing the risk with assuming that five participants will identify 85% of the problems in usability tests.

4.3.1 What the calculating method does

The calculating method calculates what the chance would have been to find a decided percentage of the problems identified at a conducted usability test if a fewer number of the participants used in the usability would have been used.

To explain the calculating method with an example, let us say that a usability test was performed on 10 participants and that totally 15 problems were identified. Now we would like to know what the chance would have been to find at least 85%, which is at least 13, of the problems if, for example, only 6 of the participants would have been used.

First, the percentage of the identified problems that would have been found by each possible combination of 6 of the 10 participants has to be calculated. Then, by looking at the distribution of those calculations, the chance that at least 85% of the problems would have been found after 6 participants can be calculated.

To calculate the number of possible ways to combine 6 participants from 10 participants we use the formula

z = n!/((n-k)!*k!)

where n is the number of participants used in the usability test, k is the number of participants we want to combine from n, and z is the number of ways we can combine k from n.

So continuing the example, we calculate that there are 210 possible ways to combine 6 participants from the 10 participants. This means that if we would have decided to use 6 of the 10 participants, we could have picked out 6 participants in 210 different ways. Now, by looking at the percentage of the identified problems that would have been found by each of these 210 different groups of 6 participants, let us say that 190 of the groups would have found over 85% of

the problems and the remaining 20 groups would have not. Since 190/210 = 90,47%, we can say that if we would have picked out to use 6 of the 10 participants we would have with a 90,47% certainty found at least 85% of the problems that were found after using the 10 participants.

If a larger number of the participants would have been picked out, there would be a bigger chance that the specific percentage of problems searched for would have been found. If a smaller number of the participants would have been picked out, there would be a smaller chance that the specific percentage of problems searched for would have been found. Using this calculating method picking out different numbers of participants makes it possible to see what number of participants that would have been needed to find the specific percentage of problems searched for.

4.3.2 Why the calculating method were decided to be used

The reason for using this calculating method was that it made it possible to make exact calculations about if a lesser number of participants than what was used in a usability test would have been enough to find the problems in that usability test. Also since data from 16 usability tests were available, this method combined with the usability tests were suited to tell if there would be a recurring pattern in how many participants that would have been needed.

Since the data from the usability tests contained information only from usability testing between 8-10 participants, information about if a higher number of participants than that would encounter any additional problems would not be possible to get. No method, for predicting if a higher number of participants would encounter any additional problems with high certainty, could be found either. Therefore, it was a natural alternative to create a method that calculates if a lower number of participants than what was used would be enough to find the identified problems.

4.3.3 Programming the calculating method

To make the calculating method's calculations possible to execute in an efficient way, a calculating program was written in the programming language, and computer program, Matlab. To execute the calculations on the data from the usability tests, the calculating program reads the data from the usability tests in the form of adapted spreadsheets that describes which participants encountered which problems. The output of the calculating program is the chance in percentage, that would have been, for each number of participants lesser than what was used to encounter a decided percentage of the problems.

4.4 Sorting data

Since the calculating program reads the data in the form of adapted spreadsheets, the data in the observation files from each usability tests had to be sorted into such new adapted spreadsheets. This was done by manually transferring the information from the observation files about which participants encountered which problems, into such spreadsheets. During this phase usability researchers from RISE were able to help out with any questions about the observation files and how they should be interpreted.

Following is an example of a spreadsheet that can be read by the calculating program. The rows show which problems that were encountered by each participant, and the columns show which participants that encountered each problem. The problems marked in red has been rated high by their severity, the problems marked in orange has been rated medium by their severity, and the problems marked in blue has been rated low by their severity. The product that was usability tested, that this spreadsheet represents, was an oven.

	A	B	C	D	E	F	G	H	I
1		problem1	problem2	problem3	problem4	problem5	problem6	problem7	problem8
2	participant1	1	2	3	4	5			
3	participant2		2			5	6		
4	participant3		2			5			
5	participant4	1	2	3					
6	participant5	1	2	3		5			
7	participant6		2					7	
8	participant7	1	2	3					
9	participant8		2			5			
10	participant9	1	2		4	5		7	8

4.4.1 Decision of what a ”problem” would refer to

The decision of what a ”problem” would refer to, in the observation files, were a matter of definition and choice.

The observation files contained notes of specific problematic behaviours of the participants that can be called ”problem behaviours”. These problem behaviours had been divided up by what area of the product the problem behaviours referred to, that can be called ”problem areas”. Several different problem behaviours could refer to the same problem area.

For example, a problem area could be that the start button placed on top of a vacuum cleaner was hard to find, and two problem behaviours both referring to this problem area could be that one participant looked for the start button on the side of the vacuum cleaner, and that another participant looked for the start button on the handle of the vacuum cleaner.

A choice had to be made about if a problem, in this study, would refer to a problem behaviour or a problem area. If a problem would be decided to refer to a problem area, then any problem behaviour associated to a problem area would be enough to encounter for a problem to be counted as found. This would mean that other valuable problem behaviours associated with the problem areas could be left out. For this reason a problem in this study was decided to be referred to, what in the observation files could be called, a problem behaviour.

4.4.2 How the data needed to be sorted to be able to use the calculating program on different severity rated problems

The analysis of the collected data was decided to be applied in different ways by including different problems depending on the problems' severity ratings. One way was to analyze all problems with all kinds of severity ratings, and another way was to analyze only the problems that had been severity rated either high or medium. To be able to do this in the calculating program, separate adapted spreadsheets had to be created depending on which kind of severity rated problems would be included.

For example, if we would like to use the calculating program to know what the chance would have been to find all kinds of severity rated problems, a spreadsheet containing problems with all kinds of severity ratings had to be created. If we instead would like to use the calculating program to know what the chance would have been to find specifically the high- and medium-severity rated problems, a spreadsheet containing only the high- and medium-severity rated problems had to be created.

Following is an example of two spreadsheets, one including all kinds of severity rated problems, and one including only high- and medium-severity rated problems.

	A	B	C	D	E	F	G	H	I
1		problem1	problem2	problem3	problem4	problem5	problem6	problem7	problem8
2	participant1	1	2	3	4	5			
3	participant2		2			5	6		
4	participant3		2			5			
5	participant4	1	2	3					
6	participant5	1	2	3		5			
7	participant6		2					7	
8	participant7	1	2	3					
9	participant8		2			5			
10	participant9	1	2		4	5		7	8

	A	B	C	D	E	F	G	H	I
1		problem1	problem2	problem3	problem4	problem5			
2	participant1	1	2	3	4	5			
3	participant2		2			5			
4	participant3		2			5			
5	participant4	1	2	3					
6	participant5	1	2	3		5			
7	participant6		2						
8	participant7	1	2	3					
9	participant8		2			5			
10	participant9	1	2		4	5			

4.5 Analysis of data

The collected data from the usability tests were analyzed, using the calculating program created in Matlab, and then visualized in graphs.

4.5.1 The three questions that were focused on using the calculating method

The data, collected through usability testing at the usability test lab, were analyzed in three different ways with a focus on answering three different questions. The three questions are the following.

1. How many participants would have been needed to find 85% of the problems identified in the usability tests?
2. How many participants would have been needed to find 100% of the high- and medium-severity rated problems identified in the usability tests?
3. How many participants would have been needed to find 85% of the problems, plus 100% of the high- and medium-severity rated problems, identified in the usability tests?

Why the analysis focused on answering the three questions

The reason for analyzing the data with a focus on answering these three questions was to get information from different angles of what the outcome would have been for each usability test, depending on the number of the participants used.

The reason for analyzing the number of participants that would have been needed to find 85% of the problems identified in the usability tests was to be able to easily compare the results to Nielsen's indication that five participants will find 85% of the problems.

The reason for analyzing the number of participants that would have been needed to find 100% of the high- and medium-severity rated problems identified in the usability tests, was that by studying the work of the usability researchers at RISE the high importance of finding the high-and medium severity rated problems became very clear. Analyzing the high- and medium-severity rated problems separately from the other problems would also show if a lower number of participants are needed to find higher severity rated problems, as meant by earlier research [23].

The reason for analyzing the number of participants that would have been needed to find 85% of the problems, plus 100% of the high- and medium-severity rated problems, identified in the usability test combined, was since both these sorts of information can be of interest when performing usability tests.

4.5.2 Analyzing the data with different levels of certainty

Each of the three ways of analyzing the data was analyzed with four levels of certainty that the searched for problems would have been found. With a lower level of certainty to find the problems, the chance increases that a lower number of the participants would have been needed. With a higher level of certainty to find the problems, the chance increases that a higher number of the participants would have been needed. The different levels of certainty were 100%, 95%, 90% and 80%.

To explain this with an example, let us say that a usability test were performed on 10 participants, and we wanted to know the number of participants that would have been needed to find some percentage of the identified problems. If we would calculate the number of participants that would have been needed to find the problems with a 100% certainty, let us say, we calculate that nine participants would have been needed. But, if we would calculate the number of participants that would have been needed to find the problems with an 80% certainty, let us say, we calculate that five participants would have been needed.

Why the data were analyzed with different levels of certainty

The reason for why the three ways of analyzing the data were analyzed with different levels of certainty was to give different perspectives on what the outcome could be depending on how certain one would accept to be to find the problems searched for. How certain one can accept to be to find the problems searched for when performing a usability test can differ for several reasons. Some reasons could, for example, be about how much time and resources that can be afforded in the usability test, or how important it is that the product tested will work without flaws.

The reason for picking out 100%, 95%, 90% and 80% as the different levels of certainty was that these are the levels of certainty that usually are used in research associated to usability [3].

Following is written, by the Interaction Design Foundation, about the different percentage of certainty when using confidence intervals in UX-research (User Experience-research) [3].

> *"Near Certainty (Confidence Value: 99% or greater)*
>
> *There are industries in which close enough is simply not good enough. These are the industries in which products may kill or maim or expose companies to unacceptable levels of risk. Pharmacy is one industry like that. Autopilot manufacturers for planes would be another."*

In these instances you need to be as close to 100% certain in your research as possible. That comes at a pretty hefty price-tag and takes a lot of time. Which is why new drug research, for example, can run at billions of dollars even after the new chemical has been identified and it can take years for that drug to come to market."

"Publishable in Journals (Confidence Value: 95% of [sic] *greater)*

Journals are not as exacting as customers of pharmaceutical companies but nor do they tend to allow you to publish data without a very high-degree of certainty in its accuracy. The benchmark for most academic peer-reviewed journals tends to sit at 95% confidence in the results.

There are other industries which require this kind of accuracy too. Political polling organizations are expected to deliver this kind of accuracy; which explains why their sample sizes for polls tend to be in the hundreds or thousands.

Again, this kind of research takes time and is costly. In most cases, it's simply too high a degree of certainty to be attained by a UX researcher except in very specific circumstances."

"Good Enough to Make Commercial Decisions (Confidence Value: 90% or greater)

Much of corporate life is about compromises and the general compromise for companies engaged in UX research is that they like to be around 90% certain that results are valid before they base major user-facing decisions on them.

For small UX teams or solo researchers – this can be a big challenge, to reach this degree of confidence and it requires a lot of attention to detail in research design. The payoff is that it becomes harder to blame the UX team if something goes wrong – there's a 10% chance that whatever your research results show that they weren't correct.

"Good Enough to Justify More Research/Development (Confidence Value: 80% or greater)

If we had to wait for 90% certainty in all our research – it would still take too long to develop products. So before they are pushed in front of users, most business will accept a lower level of certainty, usually around the 80% mark.

That's a 4 in 5 chance of being right and makes it worth pursuing a research or development avenue. It's also much cheaper and easier to achieve and can be done with relatively small sample sizes (of

course this varies depending on technique selected and the size of the user pool).

If getting it wrong won't cost a fortune or destroy a company's reputation – this can be a healthy place to conduct most of your research. It allows for rapid iteration and ideation without descending into a guessing game about user needs either."

4.5.3 Analyzing the data divided up by the numbers of participants used in the usability tests

The analysis of the data from the usability tests was separately divided up by the number of participants that were used in the usability tests. The data from the usability tests that used 8 participants were analyzed together, the data from the usability tests that used 9 participants were analyzed together, and the data from the usability tests that used 10 participants were analyzed together.

Why the analysis of the data was divided up by the number of participants used in the usability tests

The reason for analyzing the data from the usability tests that used 8, 9 and 10 participants separately was that it otherwise would have caused the usability tests to have unequal conditions in the analysis.

To explain this with an example, let us say that we analyze the data from a usability test that used 8 participants and from a usability test that used 10 participants, and let us say that the results showed that 8 participants would have been needed to find the searched for problems for both these usability tests. The results combined would show two usability tests where 8 participants would have been needed, but, since the usability tests had unequal conditions, in case of different numbers of participants used, the results would not necessarily say anything about if eight would be the suitable number of participants to use. Let us say that the usability test that used 8 participants instead would have used 10 participants, then the number of participants that would have been needed could be something else than 8, for example, 10. This also applies to the usability test that used 10 participants if it would have used 8 participants instead.

4.5.4 Visualizing the results of the analysis

To get an overview of any recurring pattern about which number of participants to use, the results were put in table graphs to illustrate how many participants

that would have been needed for the different usability tests in the different cases.

Chapter 5

Results

In this section, the results from calculating what number of participants that would have been needed to find the identified problems are presented in chart tables.

This section presents results from three different ways of analyzing the collected data, each way with a focus on answering a different question. The different questions are the following.

1. How many participants would have been needed to find 85% of the problems identified in the usability tests?
2. How many participants would have been needed to find 100% of the high- and medium-severity rated problems identified in the usability tests?
3. How many participants would have been needed to find 85% of the problems, plus 100% of the high- and medium-severity rated problems, identified in the usability tests?

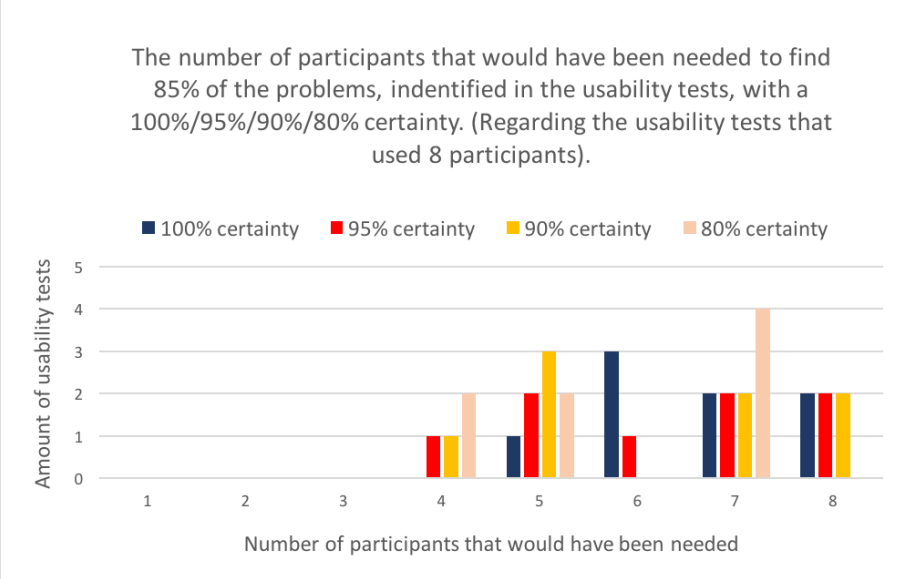
Each of these three ways of analyzing the data was analyzed with four levels of certainty that the searched for problems would have been found. The different levels of certainty were 100%, 95%, 90% and 80%.

Each of these three ways of analyzing the data were separately analyzed by the numbers of participants used in the usability tests. So, each way of analyzing the data separately analyzed the data from the usability tests depending on if the usability test used 8, 9 or 10 participants.

5.0.1 Number of participants that would have been needed to find 85% of the problems identified in the usability tests

The following three graphs show what numbers of participants that would have been needed to find 85% of the problems identified in the usability tests, with different levels of certainty.

Usability tests that used 8 participants



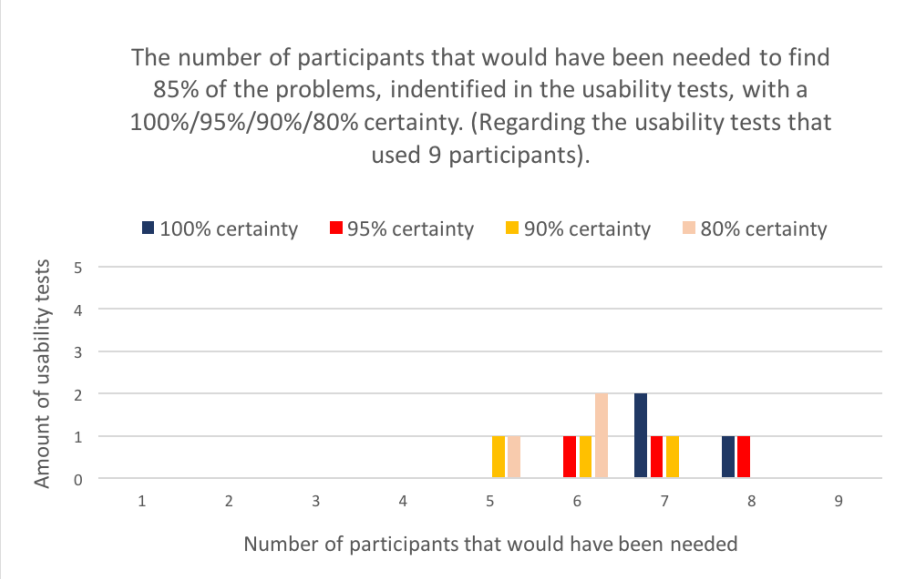
With a 100% certainty that the 85% of the identified problems would have been found, two usability tests would have needed 8 participants, two usability tests would have needed 7 participants, three usability tests would have needed 6 participants, and one usability test would have needed 5 participants.

With at least a 95% certainty that the 85% of the identified problems would have been found, two usability tests would have needed 8 participants, two usability tests would have needed 7 participants, one usability tests would have needed 6 participants, two usability tests would have needed 5 participants, and one usability test would have needed 4 participants.

With at least a 90% certainty that the 85% of the identified problems would have been found, two usability tests would have needed 8 participants, two usability tests would have needed 7 participants, three usability tests would have needed 5 participants, and one usability tests would have needed 4 participants.

With at least an 80% certainty that the 85% of the identified problems would have been found, four usability tests would have needed 7 participants, two usability tests would have needed 5 participants, and two usability tests would have needed 4 participants.

Usability tests that used 9 participants



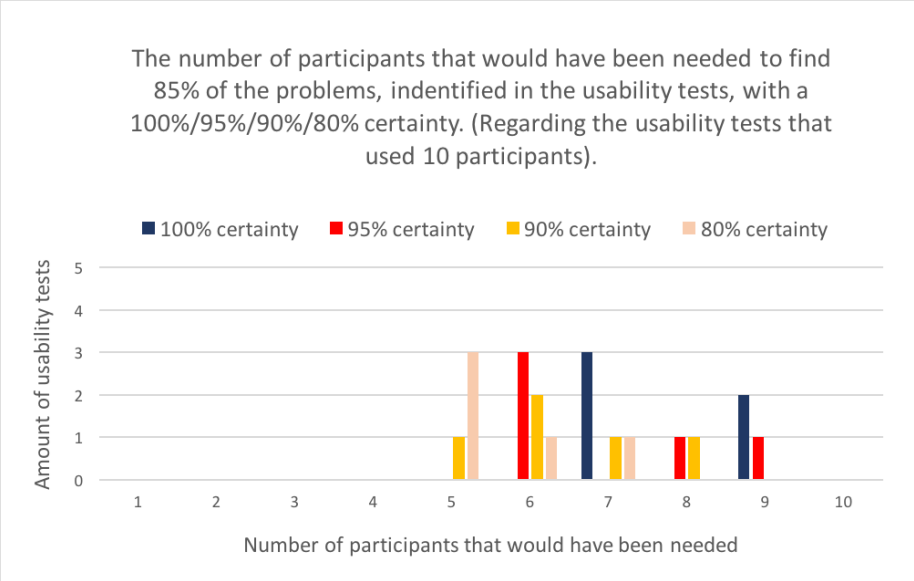
With a 100% certainty that the 85% of the identified problems would have been found, one usability tests would have needed 8 participants, and two usability tests would have needed 7 participants.

With at least a 95% certainty that the 85% of the identified problems would have been found, one usability tests would have needed 8 participants, one usability tests would have needed 7 participants, and one usability tests would have needed 6 participants.

With at least a 90% certainty that the 85% of the identified problems would have been found, one usability tests would have needed 7 participants, one usability tests would have needed 6 participants, and one usability tests would have needed 5 participants.

With at least an 80% certainty that the 85% of the identified problems would have been found, two usability tests would have needed 6 participants, and one usability tests would have needed 5 participants.

Usability tests that used 10 participants



With a 100% certainty that the 85% of the identified problems would have been found, two usability tests would have needed 9 participants, and three usability tests would have needed 7 participants.

With at least a 95% certainty that the 85% of the identified problems would have been found, one usability tests would have needed 9 participants, one usability tests would have needed 8 participants, and three usability tests would have needed 6 participants.

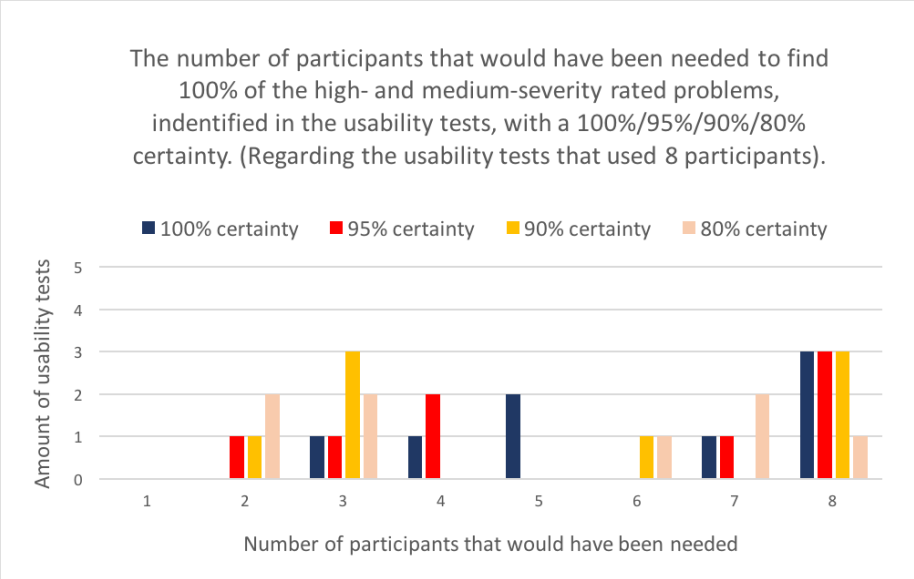
With at least a 90% certainty that the 85% of the identified problems would have been found, one usability tests would have needed 8 participants, one usability tests would have needed 7 participants, two usability tests would have needed 6 participants, and one usability tests would have needed 5 participants.

With at least an 80% certainty that the 85% of the identified problems would have been found, one usability tests would have needed 7 participants, one usability tests would have needed 6 participants, and three usability tests would have needed 5 participants.

5.0.2 Number of participants that would have been needed to find 100% of the high- and medium-severity rated problems identified in the usability tests

The following three graphs show what numbers of participants that would have been needed to find 100% of the high- and medium-severity rated problems identified in the usability tests, with different levels of certainty.

Usability tests that used 8 participants



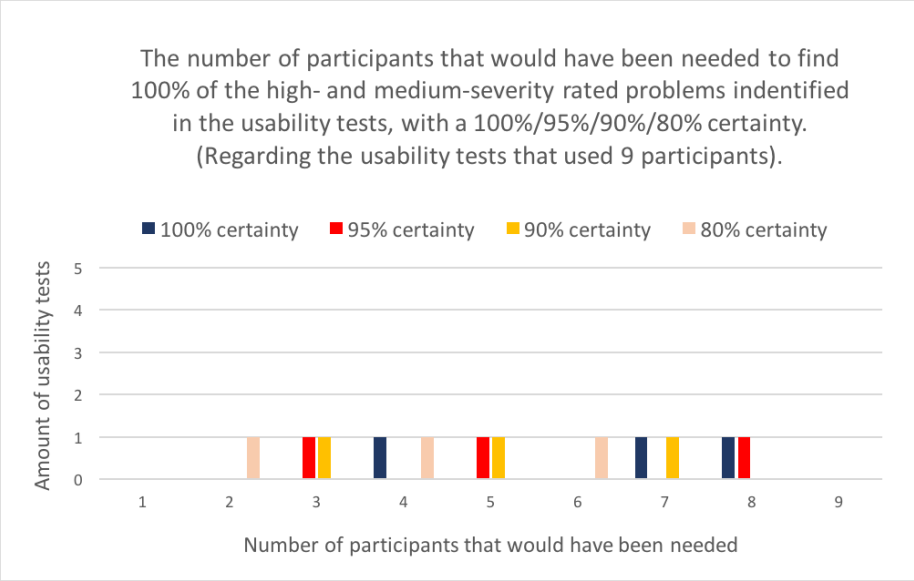
With a 100% certainty that the 100% of the high- and medium-severity rated problems would have been found, three usability tests would have needed 8 participants, one usability tests would have needed 7 participants, two usability tests would have needed 5 participants, one usability tests would have needed 4 participants, and one usability tests would have needed 3 participants.

With at least a 95% certainty that the 100% of the high- and medium-severity rated problems would have been found, three usability tests would have needed 8 participants, one usability tests would have needed 7 participants, two usability tests would have needed 4 participants, one usability tests would have needed 3 participants, and one usability tests would have needed 2 participants.

With at least a 90% certainty that the 100% of the high- and medium-severity rated problems would have been found, three usability tests would have needed 8 participants, one usability tests would have needed 6 participants, three usability tests would have needed 3 participants, and one usability tests would have needed 2 participants.

With at least an 80% certainty that the 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 8 participants, two usability tests would have needed 7 participants, one usability tests would have needed 6 participants, two usability tests would have needed 3 participants, and two usability tests would have needed 2 participants.

Usability tests that used 9 participants



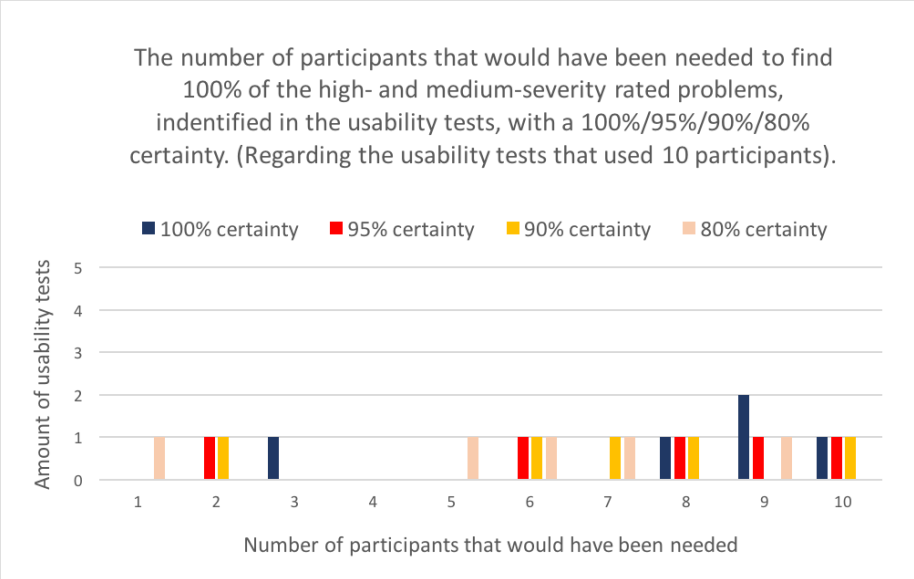
With a 100% certainty that the 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 8 participants, one usability tests would have needed 7 participants, and one usability tests would have needed 4 participants.

With at least a 95% certainty that the 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 8 participants, one usability tests would have needed 5 participants, and one usability tests would have needed 3 participants.

With at least a 90% certainty that the 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 7 participants, one usability tests would have needed 5 participants, and one usability tests would have needed 3 participants.

With at least an 80% certainty that the 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 6 participants, one usability tests would have needed 4 participants, and one usability tests would have needed 2 participants.

Usability tests that used 10 participants



With a 100% certainty that the 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 10 participants, two usability tests would have needed 9 participants, one usability tests would have needed 8 participants, and one usability tests would have needed 3 participants.

With at least a 95% certainty that the 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 10 participants, one usability tests would have needed 9 participants, one usability tests would have needed 8 participants, one usability tests would have needed 6 participants, and one usability tests would have needed 2 participants.

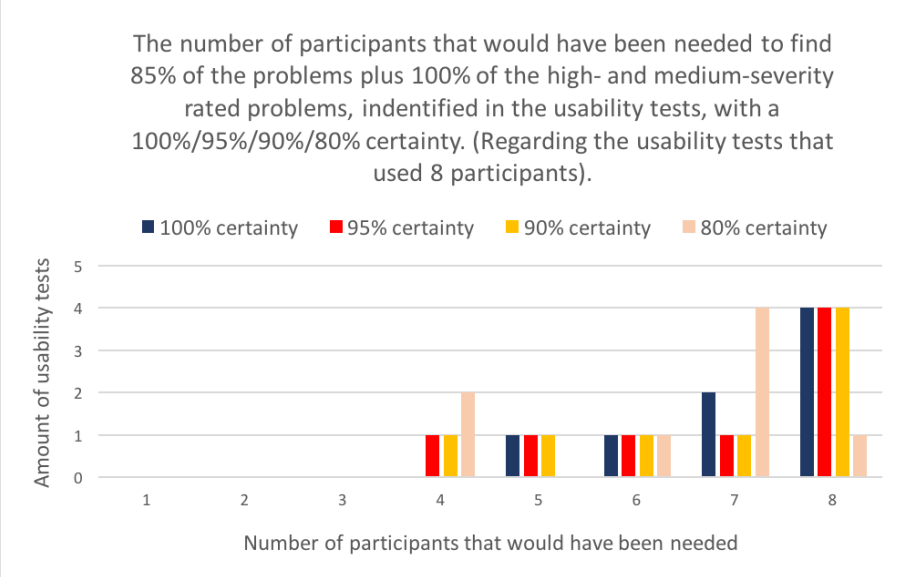
With at least a 90% certainty that the 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 10 participants, one usability tests would have needed 8 participants, one usability tests would have needed 7 participants, one usability tests would have needed 6 participants, and one usability tests would have needed 2 participants.

With at least an 80% certainty that the 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 9 participants, one usability tests would have needed 7 participants, one usability tests would have needed 6 participants, one usability tests would have needed 5 participants, and one usability tests would have needed 1 participant.

5.0.3 Number of participants that would have been needed to find 85% of the problems, plus 100% of the high- and medium-severity rated problems, identified in the usability tests

The following three graphs show what numbers of participants that would have been needed to find 85% of the problems plus all high- and medium-severity rated problems identified in the usability tests, with different levels of certainty.

Usability tests that used 8 participants



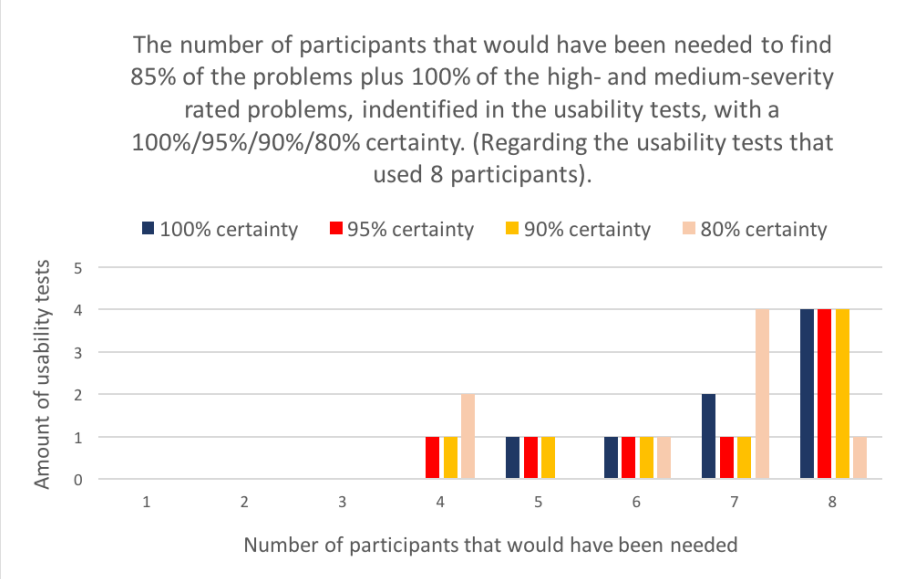
With a 100% certainty that the 85% of the problems plus 100% of the high- and medium-severity rated problems would have been found, four usability tests would have needed 8 participants, two usability tests would have needed 7 participants, one usability test would have needed 6 participants, and one usability test would have needed 5 participants.

With at least a 95% certainty that the 85% of the problems plus 100% of the high- and medium-severity rated problems would have been found, four usability tests would have needed 8 participants, one usability test would have needed 7 participants, one usability test would have needed 6 participants, one usability test would have needed 5 participants, and one usability test would have needed 4 participants.

With at least a 90% certainty that the 85% of the problems plus 100% of the high- and medium-severity rated problems would have been found, four usability tests would have needed 8 participants, one usability test would have needed 7 participants, one usability test would have needed 6 participants, one usability test would have needed 5 participants, and one usability test would have needed 4 participants.

With at least an 80% certainty that the 85% of the problems plus 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 8 participants, four usability test would have needed 7 participants, one usability test would have needed 6 participants, and two usability test would have needed 4 participants.

Usability tests that used 9 participants



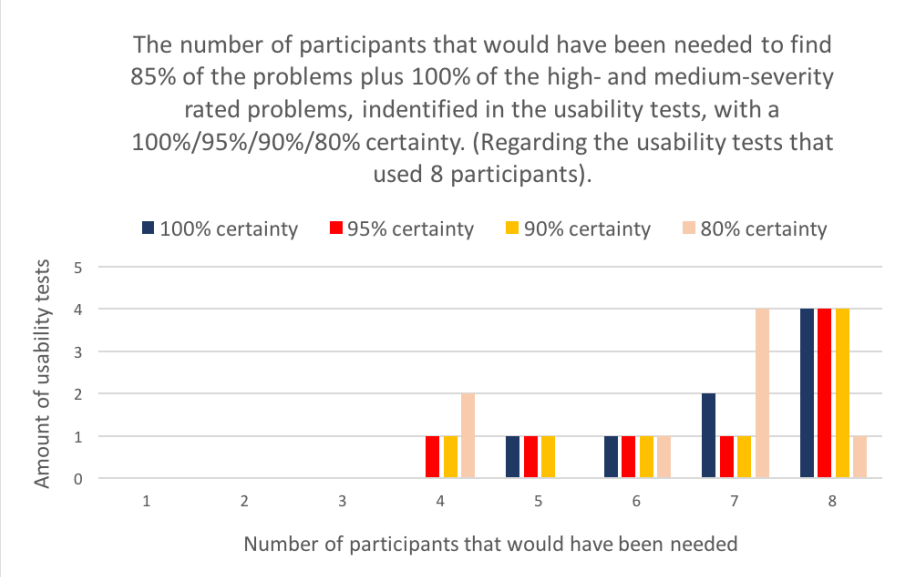
With a 100% certainty that the 85% of the problems plus 100% of the high- and medium-severity rated problems would have been found, two usability tests would have needed 8 participants, and one usability tests would have needed 7 participants.

With at least a 95% certainty that the 85% of the problems plus 100% of the high- and medium-severity rated problems would have been found, two usability tests would have needed 8 participants, and one usability test would have needed 6 participants.

With at least a 90% certainty that the 100% of the high- and medium-severity rated problems would have been found, two usability tests would have needed 7 participants, and one usability test would have needed 5 participants.

With at least an 80% certainty that the 85% of the problems plus 100% of the high- and medium-severity rated problems would have been found, two usability tests would have needed 6 participants, and one usability test would have needed 5 participants.

Usability tests that used 10 participants



With a 100% certainty that the 85% of the problems plus 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 10 participants, three usability tests would have needed 9 participants, and one usability test would have needed 8 participants.

With at least a 95% certainty that the 85% of the problems plus 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 10 participants, one usability test would have needed 9 participants, two usability tests would have needed 8 participants, and one usability test would have needed 6 participants.

With at least a 90% certainty that the 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 10 participants, one usability test would have needed 8 participants, two usability

tests would have needed 7 participants, and one usability tests would have needed 6 participants.

With at least an 80% certainty that the 85% of the problems plus 100% of the high- and medium-severity rated problems would have been found, one usability tests would have needed 9 participants, one usability test would have needed 7 participants, two usability tests would have needed 6 participants, and one usability tests would have needed 5 participants.

Chapter 6

Discussion

6.1 Participants that would have been needed to find 85% of the problems identified in the usability tests

When calculating the number of participants that would have been needed to find 85% of the problems identified in the usability tests, the results varied. With a 100% certainty that 85% of the identified problems would have been found, a variation between just a bit more than half of the participants used to, in some cases, all participants used, was what would have been needed. With at least an 80% certainty of finding the 85% of the identified problems, 5 participants would have been enough for half of the usability tests. Though, with a certainty of at least 80% of finding the problems the risk of not finding the problems could be as high as 1 in 5 times.

Comparing the results to Nielsen's indication that 85% of the problems will be found after five participants in usability tests shows that this is not the case for the usability tests analyzed in this thesis.

It is difficult to compare the results of this study to the research by Nielsen & Landauer since their research focused on the mean probability of finding the problems, without including the deviations of the number of problems found in their calculations. The deviations of the calculations of the mean probability were not included in their calculations either. One of their calculations from a user test, of the mean probability of finding the problems, from the 11 tests were as low as 16% [18]. That shows that there could be a risk of 1 in 11 times of getting a mean probability of finding the problems to be only 16%, which according to Nielsen & Landauer's formula would mean that 11 participants would be needed to find 85% of the problems. [18]. Also, 6 of the 11 tests, used

for calculating the mean probability values, were from heuristic evaluations. Since heuristic evaluations are not the same as user tests, the results of Nielsen's & Landauer's research can for that reason also be questioned.

6.2 Participants that would have been needed to find 100% of the high- and medium-severity rated problems identified in the usability tests

When calculating the number of participants that would have been needed to find 100% of the high- and medium-severity rated problems, we can see a wider variation in the number of participants that would have been needed in the usability tests. With a 100% certainty that 100% of the high- and medium-severity rated problems would have been found, a variation between 3 participants to in some cases all participants used, was what would have been needed. Lowering the level of certainty to 95% or 90% showed that in some usability tests a variation between 2 participants, to all participants used, was what would have been needed. When lowering the level of certainty to 80%, even 1 participant would be enough to use in one of the usability tests.

The high variation of participants needed to find 100% of the High- and medium severity rated problems in the usability tests, seem logical to why earlier research [23] has found that problems with a higher severity are found after fewer participants. Though the results also indicate that this does not has to be the case, since several usability tests required all, or nearly all, participants to find the high- or medium-severity rated problems.

6.3 Participants that would have been needed to find 85% of the problems, plus 100% of the high- and medium-severity rated problems, identified in the usability tests

When calculating the number of participants that would have been needed to find 85% of the problems plus 100% of the high- and medium-severity rated problems, we can see a smaller variation in how many participants that would have been needed, and that the usability tests would have needed a higher number of participants. With a 100% certainty of finding the problems, all or close to all participants would have been needed for the majority of the usability tests. When lowering the level of certainty, the variation of participants that would have been needed varies more between a bit more than half of the participants used to all participants used.

A reasonable explanation to why a higher number of participants would have been needed, when calculating the number of participants needed for finding 85% of the problems plus 100% of the high- and medium-severity problems, is that two demands had to be fulfilled. When only calculating the number of participants needed to find 85% of the problems or when only calculating the number of participants needed to find 100% of the high- and medium- severity rated problems, only one demand had to be fulfilled, and therefore it is logical that fewer participants would have been enough in some of the usability tests in those cases.

6.4 Limitations in the results

6.4.1 Limitations in the results regarding the number of participants used in the usability tests

If the data analyzed were collected through usability tests that used more than 8-10 participants the results could have shown a more detailed analysis of how many participants that would have been needed. If this analysis, of data collected through usability tests that used 8-10 participants, are enough to say something about what number of participants to use can always be discussed. Those who want to conduct usability tests on physical products and use significantly more than 8-10 participants may not get the information they are looking for from these results. Though, those who know that they will not use more than 8-10 participants when conducting usability tests on physical products can probably get more useful information from these results.

6.4.2 Limitations in the results regarding parameters affecting the usability tests

Another limitation in the results, whose effect on the results seems more obvious to us after this thesis study, is that each usability test is affected by different parameters. The parameters we can imagine has a strong affect on how many participants that will be needed in usability tests are the following.

- How advanced the tested product is?
- How comprehensive the usability test of the product is?
- How well the participants represent the target group?
- Who the people conducting the usability test are and what their experience of conducting usability tests are?
- Decisions of how to conduct and analyze the usability test?

In this study, a decision of what a problem would refer to had to be made, which were a parameter that definitely would affect the results. To explain how this decision can affect what number of participants that will be needed, imagine if it would be decided that a problematic behavior must have happened in exactly the same way by participants to be able to say that several participants encountered the same problem. That would probably mean that pretty much all participants would be needed, for all problems to be found, in that usability test. But, if it would be decided that a problematic behavior could have happened in a wide variety of ways by participants to be able to say that several participants encountered the same problem, then fewer participants would definitely be needed.

Macefield's research about what parameters that seem to affect the number of participants needed in usability tests seem to be in the right direction to be able to know how many participants to use.

Chapter 7

Conclusion

The results show that the number of participants needed to find 85% of the problems, identified in the usability tests, vary. Depending on the certainty of finding the problems, a variation between around half of the participants used, to around all participants used, was what would have been needed.

The results show that the number of participants needed to find 100% of the of the high- and medium-severity rated problems, identified in the usability tests, vary significantly. Depending on the certainty of finding the problems, a variation between the first few participants used, to around all participants used, was what would have been needed.

The results show that the number of participants needed to find 85% of the problems plus 100% of the of the high- and medium-severity rated problems, identified in the usability tests, vary less. Depending on the certainty of finding the problems, around half of the participants used, to around all participants used, was what would have been needed. For most usability tests in this case, around all participants used would have been needed.

From these results, we can conclude that five participants cannot be counted on to be enough to find 85% of the problems when usability testing physical products.

We can also conclude that the number of participants needed to find the higher severity rated problems, when usability testing physical products, varies significantly.

We can also conclude that the number of participants needed to find 85% of the problems, plus the higher severity rated problems, when usability testing physical products, vary. Though, the variation, in this case, points more to the higher numbers of participants used, compared to the numbers of participants needed to find 85% of the problems.

Chapter 8

Future work

Since the usability tests analyzed in this study were conducted on 8-10 participants, the analysis was restricted to calculating what lower number of participants that would be enough to use than that. Some usability tests may have needed even more participants for finding a significant number of additional problems, but we will never know that from this study. To get more detailed answers of what number of participants that would have been needed in usability tests, usability tests conducted on a higher number of participants should be analyzed.

When deciding how many participants to use in a usability test, it would be helpful to understand what affects how many participants that are needed. Therefore, research focusing on how different parameters affects the number of participants needed in usability tests should be conducted.

Chapter 9

Acknowledgments

The author would like to thank the usability researchers at RISE in Stockholm for their feedback, support and welcoming attitude throughout this thesis. A special thanks to Mattias Widerstedt, Lisa Jonsson, Elin Tybring, John Hedlund, Jesper Andersson and Johan Gretland.

A special thanks to the author's supervisor Mattias Widerstedt at RISE, for his feedback and support, which has been very helpful in the process of this thesis.

Also, a special thanks to the author's supervisor at Umeå University, Shafiq Urréhman, for his guidance and help.

Finally, a huge thanks to the peer reviewers Malin Jofjärd Lövgren and Stina Olofsson for their feedback and support throughout this thesis.

Bibliography

[1] Joseph S Dumas and Janice Redish. *A practical guide to usability testing*. Intellect books, 1999.

[2] Laura Faulkner. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3):379–383, 2003.

[3] Interaction Design Foundation. Confidence intervals and ux research. `https://www.interaction-design.org/literature/article/confidence-intervals-and-ux-research`. [Online website].

[4] Interaction Design Foundation. What is usability testing? `https://www.interaction-design.org/literature/topics/usability-testing`. [Online website].

[5] Wonil Hwang and Gavriel Salvendy. Number of people required for usability evaluation: the 10±2 rule. *Communications of the ACM*, 53(5):130–133, 2010.

[6] W Iso. 9241-11 ergonomic requirements for office work with visual display terminals (vdts). *The international organizaion for standardization*, 45, 1998.

[7] Clayton Lewis. *Using the" thinking-aloud" method in cognitive interface design*. IBM TJ Watson Research Center, 1982.

[8] James R Lewis, Suzanne C Henry, and Robert L Mack. Integrated office software benchmarks: A case study. In *INTERACT*, pages 337–343, 1990.

[9] Ritch Macefield. How to specify the participant group size for usability studies: a practitioner's guide. *Journal of Usability Studies*, 5(1):34–45, 2009.

[10] Robert L Mack and Jakob Nielsen. Usability inspection methods: Executive summary. In *Readings in Human–Computer Interaction*, pages 170–181. Elsevier, 1995.

[11] Rolf Molich and Jakob Nielsen. Improving a human-computer dialogue. *Communications of the ACM*, 33(3):338–348, 1990.

[12] Jakob Nielsen. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 373–380. ACM, 1992.

[13] Jakob Nielsen. Estimating the number of subjects needed for a thinking aloud test. *International journal of human-computer studies*, 41(3):385–397, 1994.

[14] Jakob Nielsen. How to conduct a heuristic evaluation. `https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/`, January 1995. [Online website].

[15] Jakob Nielsen. Why you only need to test with 5 users, 2000.

[16] Jakob Nielsen. How many test users in a usability study? `https://www.nngroup.com/articles/how-many-test-users/?lm=eyetracking-task-scenarios&pt=youtubevideo`, June 2012. [Online website].

[17] Jakob Nielsen. Thinking aloud: The 1 usability tool. `https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/`, January 2012. [Online website].

[18] Jakob Nielsen and Thomas K Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 206–213. ACM, 1993.

[19] Jakob Nielsen and Rolf Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256. ACM, 1990.

[20] Jared Spool and Will Schroeder. Testing web sites: Five users is nowhere near enough. In *CHI'01 extended abstracts on Human factors in computing systems*, pages 285–286. ACM, 2001.

[21] Usability.gov. Cognitive walkthrough. `https://www.usability.gov/what-and-why/glossary/cognitive-walkthrough.html`. [Online website].

[22] Robert A Virzi. Streamlining the design process: Running fewer subjects. In *Proceedings of the Human Factors Society Annual Meeting*, volume 34, pages 291–294. SAGE Publications Sage CA: Los Angeles, CA, 1990.

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

[24] Alan Woolrych and Gilbert Cockton. Why and when five test users aren't enough. In *Proceedings of IHM-HCI 2001 conference*, volume 2, pages 105–108. Eds)(Cépaduès Editions, Toulouse, FR, 2001), 2001.

[25] Peter C Wright and Andrew F Monk. A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies*, 35(6):891–912, 1991.

Appendix A

The code of the calculating program

```
%---------------------Fill these manually first---------------------
%The file name of the Excel-file
fileName = 'NameOfTheFile.xlsx';
%The number of test participants used in the usability test
NrOfTPsUsed=10;
%The number of problems identified in the usability test
NrOfProblems=20;
%The number of percentage of the identified problems you want to
%know if you would have found with fewer participants
NrOfPercentageSeachedFor=85;
%-------------------------------------------------------------------

%Reads data from spreadsheet in excel-file
tempData = xlsread(fileName,'Sheet1','A1:ZZ100');

y=[];

for i=1:1:NrOfTPsUsed

    tempNrOfTPs=i;
    problemsFoundCounter=0;
    foundPercentage=0;
    resultList = [];

    %problemAlreadyFound = 0 if a problem has not been found yet,
    %problemAlreadyFound = 1 if a problem already has been found
    problemAlreadyFound = zeros(1,NrOfProblems);

    %Calculates the different combinations of participants you can
    %pick from of the participants used
    participantCombinations = nchoosek(1:NrOfTPsUsed,tempNrOfTPs);

```

```
33      [rows,columns] = size(participantCombinations);
34
35      %Calculates the number of problems found by each combination
36      %of participants
37      for i=1:1:rows
38          for j=1:1:columns
39              for k=1:1:NrOfProblems
40
41                  check = tempData(participantCombinations(i,j),k);
42
43                  for q=1:1:NrOfProblems
44
45                      if check == q && problemAlreadyFound(q) == 0
46                          problemsFoundCounter=problemsFoundCounter+1;
47                          problemAlreadyFound(q) = 1;
48
49                      end
50                  end
51              end
52          end
53
54          %Calculates the percentage of problems found by each
55          %combination of participants
56          percentageFound = (problemsFoundCounter/NrOfProblems)*100;
57
58          for i=1:1:NrOfProblems
59              problemAlreadyFound(i) = 0;
60          end
61
62          %Calculates how many combinations of participnats that found
63          %at least the decided percentage of problems searched for
64          if percentageFound ≥ NrOfPercentageSeachedFor
65              foundPercentage=foundPercentage+1;
66          end
67
68          resultList(end+1) = percentageFound;
69
70          problemsFoundCounter=0;
71      end
72
73      %Calculating the certainty that the searched for problems would
74      %have been found
75      certaintyOfFindingProblems =...
76          ((foundPercentage/(numel(resultList))))*100;
77
78      y(end+1)=certaintyOfFindingProblems;
79
80      %Prints explanations of the calculations
81      explanation = ['If ',num2str(tempNrOfTPs),...
82          ' participants would have been used, '...
83          ,num2str(NrOfPercentageSeachedFor),...
84          '% of the identified problems would have been found with ...
                a certainty of '...
85          ,num2str(certaintyOfFindingProblems),'%'];
86
87      disp(explanation)
88
```

```
89  end
90
91  %Print bar chart
92  bar(y);
93  xlabel('Number of participants');
94  ylabel('certainty in % that the decided percentage of the ...
        identified problems would have been found');
```