UNIVERSITY OF TWENTE.

Usability testing for practical use:

Monitoring the process and detecting false positives face to face and via internet

Master thesis

Author:	Johannes Terwort
Student number:	s1224735
Date:	July 25, 2016
Study:	Master of Psychology
Specialization:	Human Factors Engineering Psychology
Institution:	University of Twente

Graduate Committee:

Dr. M. (Martin) Schmettow

Faculty: BMS - CPE

Prof. Dr. J.M.C. (Jan Maarten) Schraagen

Faculty: BMS - CPE

Abstract

A high usability has a positive influence on the usefulness and safety of systems (Vieritz, Yazdi, Schilberg, Ghöner, & Jeschke, 2011; Harty 2011). To ensure a systems usability, it has to be tested (Gould, Boied & Lewis, 1991). In practical use and real world settings testing usability is a question of the cost/benefit ratio and it is necessary to find the optimal method and required sample size for testing usability with the highest benefit and the lowest costs (Nielsen & Landauer, 1993; Pape, 2012; Schmettow, Vos & Schraagen, 2013). Sullivan (2013) gives lead to the idea that testing usability via a communication tool could be a cheaper but equally good approach as testing face to face. Schmettow, Vos and Schraagen (2013) described an approach called *zero-truncated logit-normal binomial model* (*LNBzt*) that is able to predict the required sample size and to monitor the process of data collection making it able to estimate a completion rate and a number of undiscovered problems. This method suffers from impacts of false positives, making it necessary to identify them. Schmettow, Vos and Schraagen (2013) used an approach to identify false positives after data collection was finished, but an approach being able to identify false positives parallel to the collection of data would be more suitable for the late control strategy. To investigate whether testing face to face works differently for testing usability than testing via a communication tool a sample of 29 participants was divided in two groups, testing one face to face and the other one via a communication tool. Additional data was gathered using semi-structured interviews and questionnaires. The process of data collection was monitored using the late control strategy while a new protocol called *False-positive Identification Protocol* (*FIP*), developed in this study, was used to identify false positives during the process. The results showed that FIP performed well for identifying false positives parallel to the data collection, improving the results of the late control strategy. The late control strategy performed more effective than older approaches to estimate required sample sizes like the magic number approach or the early control strategy. While performing good at identifying false positives, FIP failed at the point of additional data collection and an additional expert assessment was suggested to be a good addition for FIP. Furthermore, both methods, testing usability face to face and via a communication tool resulted in similar numbers of problems, yet being different problems. The question whether one method performs better for testing usability could not be answered finally because they perform differently. It is concluded that both methods, testing face to face and via a communication tool can be of benefit to the results, but more research is needed. Furthermore, in combination with the new developed protocol FIP for identifying false positives the late control strategy provides more precise results than without using FIP, estimating the required sample size and monitoring the process of data collection.

Index

List of figures and tables

1. Introduction

This section describes the reasons why this study is done, the involved parties, the theoretical background and the research question. First a general overview on the problem is given, later the theories are described in detail.

1.1 Usability and samples

Usability describes whether a product can be used by the estimated user to fulfill the product's purpose or not (Vieritz, Yazdi, Schilberg, Ghöner, & Jeschke, 2011). Shackel (2009) states that usability depends upon the design of the tool in relation to the users, the tasks, the environments and the success of the user support provided (for example: training, manuals). This means a tool has a high usability when the interaction of all those components work without big failures. Furthermore, tools have to fit the specific needs of their purpose, the user and the situation. Not meeting the usability goals of a product is one of the most occurring reasons for problems with a product like economic failure, common or even hazardous errors (Harty, 2011). To ensure that the usability of a product is given, Gould, Boies and Lewis (1991) describe four different principles, which commonly share the strong inclusion of the enduser in the design process. One of the most important parts of these principles is the testing of usability on the enduser. For this it is concluded that usability is an important precondition for a successful product and it is necessary to involve the real enduser in the design. Further it is concluded that based on the findings of Gould, Boied and Lewis (1991) the usability of a product has to be tested involving the enduser of the product.

To test a product with the aim of improving usability Gould, Boied and Lewis (1991) suggest the observation of real world interactions between the enduser and the product. As usability testing with real endusers is expensive the cost/benefit ratio of usability testing has to be optimal for a company (Nielsen & Landauer, 1993). All costs arising while testing usability are part of this cost/benefit ratio, like the execution of the test or the size of the sample needed in combination with the costs per participant. Two different approaches on testing usability on real endusers are described in greater detail later. Each participant leads to an increase of costs, making it necessary to test usability in an efficient and cost-saving way with an optimal sample size. Different approaches to find the required sample size are described later in the introduction focusing on the *zero-truncated logit-normal binomial* *model* (*LNBzt*) described by Schmettow (2009). This model is a late control model which is able to estimate the required sample size during the testing phase. This model is prone to false positives, problems that are no real problems but occur only once for different reasons (Schmettow, Vos & Schraagen, 2013). These single occurrence problems increase the estimated required sample size because the model estimates the sample size based on the number of new findings. An approach how Schmettow, Vos and Schraagen (2013) identified false positives is described in a later section, too.

1.2 Testing usability

To improve the usability of a product, it is necessary to find the problems related to usability and fix them. Accordingly, the testing of a product is required. Different factors influencing the choice for the most appropriate method on finding usability related problems are described in this section.

The first factor is described by the question: What is searched for and in which setting is this possible? Usability testing is an evaluation method to identify user requirements and usability problems involving the expected enduser (Nielsen, 1993). While testing usability it is possible to find user requirements that have not been taken into account until now. Furthermore, testing usability can decrease the number of usability related problems, sometimes also called *errors*, hidden in a product (Gould, Boied and Lewis, 1991). Errors are occasions in which the intended outcome of a planned mental or physical activity is not achieved whilst the failure is not a result of intervention or chance (Dain, 2002). Reason (1990) and Dain (2002) describe two types of errors which have to be distinguished: *Active* and *latent errors*. Active errors are slips, mistakes and lapses having an immediate effect (Reason, 1990) while latent errors occur only in combination with other factors, lying dormant and invisible within the system (Dain, 2002). Mostly, latent errors are hard to uncover because they occur only in combination with other factors and are therefore most of the time ignored assuming they will probably never occur. This kind of handling latent errors can lead to catastrophic failure of the system even though it may be prevented. Latent errors provide a larger window of opportunity for identification and intervention to prevent catastrophic failures. Both active and latent errors can be uncovered while testing the product in real world settings involving the expected enduser (Gould, Boied and Lewis, 1991). For active errors it is more likely to emerge when a normal task is done by an enduser because they have an expected similar chance to occur like in every day work. This represents the regular chance of an event in everyday work. This applies for latent errors as well. Latent errors need the interaction of different factors to occur, but in isolated test scenarios the interaction of different factors is limited due to a controlled situation. In a real world setting the interaction of multiple different factors is possible and has a similar chance to take place as in every day work. To conclude, testing in real world settings has a similar chance to find active and latent errors than daily work. Especially the possibility to identify latent errors makes it worth to test in real world settings because it is a chance to identify errors that would not be found otherwise and might possibly lead to catastrophic failures.

The second factor is described by the question: How can it be tested? To test a product with the aim of improving usability Gould, Boied and Lewis (1991) suggest the observation of real world interaction between an enduser and the product. Nielsen (1993) supports this suggestion while stating that verbal protocols and complementary interviews and questionnaires are also important tools. Verbal protocols can be concurrent or retrospective and are useful tools for uncovering problems while user opinions about improvements can be collected via complementary interviews and questionnaires. One type of verbal protocol that can be used to identify existing problems is the *Think Aloud protocol* (*TA*) used in research by Obradovich and Woods (1996), Lin et al. (1998) and Schmettow, Vos and Schraagen (2013).

Think Aloud protocols are used to understand how a user interacts with a product (Ericsson & Simon, 1993; Guan, Lee, Cuddihy & Ramey, 2006) TA protocols are a method often used in Human Computer Interaction (HCI) and are considered to provide a high face validity. There are two versions of Think Aloud protocols: concurrent (CTA) and retrospective (RTA) Think Aloud protocols (Guan et al., 2006; Haak, & Jong, 2003; Haak, Jong & Schellens, 2003); Nielsen, 1993). In CTA users work on typical tasks interacting with the product simultaneously verbalizing their thoughts and actions aloud (Nielsen, 1993). This version of the TA protocol is the most commonly practiced version but has some constraints. Nielsen (1993) describes three main problems that may occur by using CTA: First the cognitive workload could be higher due to the verbalization, resulting in affecting the task performance, and second resulting in distracting the user in his attention and concentration. The third problem is that the verbalization might lead to changes on how the user attends to components of the task. These problems have a lesser influence in RTA (Guan et al., 2006). While users perform verbalization and interaction at the same time during an CTA, the user completes the given task of product interaction first and then verbalizes his thoughts and actions afterwards in RTA. In some cases the verbalization is supported by reviewing a video-recording of the user while performing the task. Because the verbalization is performed afterward the task, the workload and attention of task components is not changed during the task. This makes RTA the preferred method for analysis where these factors are important. While compensating constraints of CTA, RTA has its own restrictions (Guan et al., 2006). The most important one is the delayed verbalization which interferes with the validity of the outcome. This is because users may give biased accounts of thoughts which they state to have had while performing the task. In addition to that users may conceal, change or even make up thoughts they report due to reasons of self-representation, social desirability, anticipation or personal opinions, resulting in biased verbalization and lower validity of the outcome. Regardless of these differences, both CTA and RTA have proven useful for usability testing in different settings and are both considered to result in comparable sets of reported usability problems (Haak, & Jong, 2003; Haak, Jong & Schellens, 2003).

The third factor is described by the question: How are problems matched and described? The data gathered by usability testing methods like TA have to be encoded and matched (Hornbæk, & Frøkjær, 2008). Coding is describing observations, while matching is a method to determine whether different descriptions of usability problems are about the same underlying design flaw or not. By using matching methods not only similarities of problem statements can be identified, but also false positives and the evaluator effect. In case of false positives the method used for matching has a large impact on whether an event is considered a false positive or not. This is because some methods use a more generalized level of matching with a greater variety of possible matching of similarities while other methods use stricter rules whether two events can be matched as similar or not. Hornbæk and Frøkjær (2008) state a comparison of two studies also has to consider differences in the matching method, as the method used for matching has a significant influence on the results of the usability test. A study using a more generalized level of matching is hard to compare to another study using a stricter level of matching. The results will most likely be different, based on the different level of generalization. The level of generalization may even differ between raters in the same study, making the description of the method and generalization level of matching even more important. Furthermore, this illustrates the importance of a method with a good comparability between the raters.

To conclude, a set of different methods has to be used to test usability. The combination of observation, Think Aloud protocols and complementary interviews and questionnaires has been proven useful (Nielsen, 1993). The choice to use either CTA or RTA depends on the type of task and cannot be set generally. In every day settings different other factors also have an influence on how usability can be tested and which methods have to be chosen (Nielsen & Landauer, 1993; Pape, 2012). Furthermore, to find active and latent errors it is necessary to test in real world settings. One of the most important factors influencing the choice of the methods and approaches is their efficiency.

1.3 Efficiency

Efficiency is an important topic in usability testing for different reasons (Woolrych, Cockton & Hindmarch, 2004). First usability itself has influence on how efficient a software or tool can be used (Bevan, 2000). Second the cost/benefit ratio of usability testing has to be optimal (Nielsen & Landauer, 1993). This means the benefits have to be greater than the costs on an optimal level. This has to be taken into account especially for companies which have to pay for improvements of their product (Pape, 2012). Companies want to improve their product as much as possible while investing as little energy and money as possible. Usability testing with real participants can be more reliable, but it is very labour intensive, time consuming and expensive (Woolrych, Cockton & Hindmarch, 2004). One part of usability testing featuring the chance of great impact on the testing efficiency is the sample needed to identify most of the usability problems (Schmettow, Vos & Schraagen, 2013). One factor of the sample which may influence the cost/benefit ratio is the amount of costs to get a sufficient user sample (Pape, 2012; Schmettow, Vos, Schraagen, 2013). For example, in some company the employees are widely spread over the whole land and some other countries in the world. Contacting these users at their original workplace or transporting them to a usability-test-lab is assumed to be an expensive and time consuming task. Sullivan (2013) suggests Skype and similar tools for technical supported communication can be an appropriate method of data collection for qualitative interviews. If a usability analysis could also be done via a software tool it would lower the costs of data gathering. Furthermore, Nielsen and Landauer (1993) state it is important for usability testing's efficiency to use the optimal size of sample needed in order to find most of the usability related problems. This assumption is based on the Pareto principle. The theory states twenty percent of energy is needed to identify eighty percent of the existing usability problems, leaving the remaining eighty percent to find the last twenty percent of the usability problems (Nielsen & Landauer, 1993; Rizwan & Iqbal, 2011). In this case the optimal sample size would be the one which is needed to identify eighty percent of the usability problems while using twenty percent of the energy, as the cost/benefit ratio is best at this point. It is furthermore necessary to know the approximate number of residing problems still hiding in the system to determine whether the optimal goal is met (Schmettow, Vos & Schraagen, 2013).

To summarize, the sample is an important factor influencing the efficiency of usability testing. One important part of the sample is the optimal sample size needed to identify most of the usability problems in a software. For this purpose there are different approaches developed to determine the sample size needed (Schmettow, Vos & Schraagen, 2013).

1.4 Estimating the required sample size

In the last years three different approaches have been developed to estimate the optimal sample size needed to find most of the existing usability related problems (Schmettow, Vos & Schraagen, 2013). The term 'Most of the problems' is often described as 85% of the existing problems. These three approaches have different results and are still a topic under discussion. However, these different approaches have two purposes being related to each other: The first purpose is to identify the required sample size while the second purpose is to evaluate whether most of the existing usability problems have been found at a given point of research. A strategy that claims to be useful has to be effective on both accounts. To determine the efficiency of these approaches a variety of research has been carried out.

In general there are two strategies are used when planning a usability evaluation (Vos, 2011). Both strategies are based on the *geometric model* developed by Virzi (1992) stating the evaluation process follows a geometric series. The most important assumption in this model is that the function of new found problems follows a $D = 1\text{-}(1\text{-}p)\text{\^{}}n$ function (D = detection rate, p = chance of error occurrence, n = sample size) resulting in less additional events when increasing the number of samples. The first approach based on this model is the *Magic Number Approach* (Nielsen, 2000; Schmettow, Vos & Schraagen, 2013). This approach assumes that all studies, regardless of their setup, are similar in how fast completeness is reached with the increasing of the sample size and users based on this a fixed sample size defined before (Schmettow, Vos & Schraagen, 2013). This sample size is called the *'Magic Number'*, which is based on the evaluation of the necessary sample size in *N* past studies (Nielsen & Landauer, 1993). The results of these past studies are evaluated while results of the present study are not involved in this evaluation. There is a debate over this approach as it results in the statement 'five users are enough to elicit an 85% defect detection rate (*D*)', which not everyone agrees to. The reason for disagreement lies dormant in the Magic Number itself, because it is based on the assumption the chance of error occurrence is $p = .31$, with a considerable standard deviation of $sd = .12$. The fact that in most studies the chance of error occurrence is not known before performing the test and the high standard deviation, compared to the total value, makes assumptions about the optimal sample size considerably uncertain (Lewis, 2001; Schmettow, Vos & Schraagen, 2013). In a recent review Schmettow (2012) concluded that Magic Numbers are simply meaningless. This debate resulted in a subsequent strategy described by Lewis (2001) named *early control strategy*. This strategy is based on initial trials ($N = 2$-4) which are used based on Virzi's geometrical model to estimate the ultimate sample size. It is assumed that at this stage of a given project resources can be still assigned to it if necessary (Schmettow, Vos & Schraagen, 2013). However, the estimates resulting out of this approach are too uncertain to be of practical value in small samples (Schmettow, 2009). The third strategy is a *late control strategy*, accounting for both incompleteness and visibly variance (Schmettow, 2012). This approach is named *zero-truncated logit-normal binomial model* (*LNBzt*) (Schmettow, 2009). In this strategy data of trial runs of the present research are used to estimate the number of discovered and remaining undiscovered defects. This can be repeated during the research process to keep track of the progress. The results are compared to pretest goals for the detection rate (*D*). Based on this, decisions on whether the gathered data are sufficient or whether the research has to be continued are made. By using this model three important calculations are possible: First, the estimation of the proportion of usability problems being undiscovered at a given point of time. Second, the extrapolation of the evaluation process and the prediction of the required sample size for a given discovery target of usability problems. Third, the determination of the accuracy of predictions by constructing confidence intervals. The estimation of the proportion of usability problems being undiscovered in a product is an important task because diversity of users can affect the detection rate (Schmettow, Vos & Schraagen, 2013). Many different factors can influence the expectations, interaction style and performance of a user while operating a device (Caulton, 2001). Different groups of users have different backgrounds, tasks and working conditions. Previous experience may have positive or negative effects on the performance while interacting with devices (Finstad, 2008). For example, Carrol and Rosson (1987) state domain expertise can prevent users from making certain mistakes, while experience with legacy devices might cause a negative transfer. Based on this, Caulton (2001) explains the discovery of usability problems is likely to be incomplete if not all possible groups of users are evenly included in the sample. In these cases the usability researcher is at risk to miss possibly hazardous usability problems. The late control strategy accounts for this variation in defect visibility and prevents overoptimistic estimates of the problem detection rate (*D*), which may result in a premature termination of the search process while the goal for the detection rate is not reached (Schmettow, Vos & Schraagen, 2013). Another important advantage of the late control strategy is that it uses confidence intervals to estimate the required number of samples. Thereby the accuracy of the estimation is more trustworthy, accounting for variances in the sample and leading to a better estimation of the sample size (Schmettow, 2009; Vos, 2011). In small samples a larger confidence interval would lead to a wider range of the estimated sample size which would lead in the early control strategy to inaccurate estimations (Schmettow, 2009). In the late control strategy however the confidence intervals are used to narrow the estimated number in a smaller range of the confidence interval resulting in a more trustworthy estimation of the optimal sample size. First evaluations of the late control strategy described by Schmettow, Vos and Schraagen (2013) prove its value. The estimated sample size and the real sample size are strongly related and more accurate than the other two strategies.

In conclusion, different studies indicate the late control strategy is a promising approach accounting for several drawbacks of the other two strategies. However, since it uses the rates of found problems to estimate the required sample size it is prone to influences of false data like false positives (Schmettow, Vos & Schraagen, 2013).

1.5 False positives

Data gathered in usability tests can contain false positives (Schmettow, Vos & Schraagen, 2013). False positives are observations which falsely appear to be a hit. In usability tests this results in problems being reported even though they do not have any harmful effect on the usability of the test object (Sears, 1997). Furthermore, false positives reduce the validity of the usability test (Woolrych, Cockton & Hindmarch, 2004). Schmettow (2009) reported a high variance and large numbers of problems named only once in previous studies. These problems have the potential to be false positives (Schmettow, 2009; Schmettow, 2013). However, an event which is only *observed* once while interacting with the system in the testing phase should still be considered a real usability problem since it happened in an actual use case (Woolrych, Cockton & Hindmarch, 2004).

False positives have an influence on the function of the late control strategy (Schmettow, Vos & Schraagen, 2013; Vos, 2011). The late control strategy uses the rate of observations to calculate an estimated completeness. Each additional observation, independent on how often a problem is reported, will increase the required sample-size estimated and lower the rate of completeness. Therefore each false positive contained in the data will change the correctness of the results to a higher estimated required sample-size. This means false positives have a negative influence on the efficiency and correctness of the late control strategy and should be identified and excluded from the data.

Schmettow, Vos and Schraagen (2013) used a method called '*triage*' to separate false positives from usability problems. The problems that were directly observed during interaction were taken as usability problems, while the remaining problems were individually mapped to the matching questions of a post-test questionnaire. Problems related to at least one negative rating were taken as valid while problems related to unambiguously positive satisfaction ratings were rated as potential false positives. Those potential false positives were reviewed in an expert screening. In this study of Schmettow, Vos and Schraagen (2013) 13% of the results were identified as false positives by this method.

In conversation Schmettow suggested to change the questionnaire of the triage and the expert screening to be more dynamic. The triage had to be used after gathering the full set of data, making it difficult to already use in combination of the late control strategy while collecting data. Hence the late control strategy is not able to present a good prediction while gathering data due to influences of false positives. Schmettow suggested to simply ask other participants about their thoughts on a given problem which occurred only once, because they are the users who are the real experts concerning the test-object. Based on this suggestion a protocol called *False-positive Identification Protocol* (FIP) was developed and tested in the current research. The full description of this protocol is given in section 2.8.

1.6 Aim of this study

The goal of this study is to evaluate the late control strategy in a case study involving a real world scenario with different difficulties like the accessibility of the sample and the need for efficiency. In addition to that a second goal is it to develop and test a method for identification of false positives during the test period.

From the literature above different subgoals of the current study are extracted:

<u>1. The evaluation of the late control strategy in a case study concerning a software interface</u>
Schmettow, Vos and Schraagen (2013) used the late control strategy to evaluate the process of their usability test on an infusion pump interface. Their results indicated that the late control strategy is a more precise strategy than the magic number and the early control strategy. Furthermore it handles problems the other strategies could not deal with. Both the study by Schmettow, Vos and Schraagen (2013) and the current study perform a usability test on a user interface. However, an infusion pump has a different purpose and another user group than a data access program interface. From information known about the users of the current study it has to be assumed they are occupationally used to the computer interfaces. Furthermore, occurring errors generally affect the company's profit instead of creating life-threatening consequences. There is no prior knowledge on the impact of this on the required sample size. For this reason a subgoal of this study is to evaluate whether the late control strategy provides similar results in the current study as in the study of Schmettow, Vos and Schraagen (2013) or if they differ essentially from these results. This comparison will be based on a confidence interval of 80% and a given discovery goal of 85%. A possible drawback that is to be considered during the evaluation and testing is the effect of diversity of the user group. Prior to this study there was information about the possibility of two different groups of users because there are two different styles of controlling the software: Via mouse and via keyboard shortcuts.

<u>2. The identification of as many usability related problems as possible</u>
To support the sub goals it is inevitable to find as many of the existing usability problems as possible. For this a usability test has to be executed on a sample of end users.

3. To develop and test a procedure to identify false positives in the found usability related problems

To support subgoal 2 the quality of the results has to be secured. Hence it is necessary to identify false positives. For this purpose a protocol will be developed and tested in the current study. The results will be compared to the results found by Schmettow, Vos & Schraagen (2013).

4. The influence of face to face communication and communication via a technical communication interface

In the study by Schmettow, Vos and Schraagen (2013) all participants were tested face to face in a laboratory setting controlling many variables. For economical and efficiency reasons the use of technical communication interfaces like Sullivan (2013) suggests has to be investigated. The focus point lies on the efficiency of the late control strategy and the results of the usability test, evaluating the appliance of these communication interfaces as an alternative possibility.

2. Method

This section describes the methods and materials used in the current research. First general information is explained before the sample, procedures and methods are described in detail.

2.1 Usability Evaluation method and background of the research

This study was done in cooperation with a German company working in the financial sector. For different reasons, like privacy protection agreements, the identity of the company is not revealed. In this study the interface of a data access tool of this company is evaluated (See Appendix A for examples). Developed by the department for software development the interface was introduced in 2006, and afterwards only maintained regarding the logical and interface level. However, no usability test was performed ever since its introduction. The evaluation for usability related problems in the current study was done based on the principles explained under paragraph 1: A representative sample of end-users was used, representative tasks were observed during actual use, a collection of qualitative and quantitative data was gathered and evaluated for usability problems.

2.1.1 Focus of this study

This study focused on detecting usability problems manifesting in the graphical user interface. Other problems arising out of the software or related soft- or hardware have not been evaluated, including connectivity problems, deeper algorithms and logic of the software, work place layout, management problems and hardware problems like performance problems related to hardware. Furthermore, only the visual interface was evaluated, auditory parts of the interface like alarms or earcons were excluded.

2.2 Tools and materials

For this research different tools and materials were used. First the software of the company was used for evaluation. In the experiments the user-test-environment of the company was used due to security restrictions like privacy protection and avoidance of data compromise. The test-environment differs only slightly from the actual production environment. It contains no actual production data in the database to prevent data corruption in case of software failure during a test. However, the interface is identical in the production- and test-environment. To record the face to face tests a DELL-Notebook and a Logitec Quickcam E3500 webcam were used. While video recording was performed using the native software of the webcam, the audio-data were recorded using the application Voice recorder on a Samsung Galaxy Ace2 Smartphone. For the tests via a communication tool Screen Connect was utilized for video submission as well as video recording. The communication was planned via Screen Connect but done via telephone because of insufficient bandwidth of the internet-connection. The audio-data was recorded with the Smartphone the same way as the face to face interviews. The analysis was done by using Microsoft Word, Microsoft Excel, IBM SPSS for statistics, Atlas.ti for coding and mapping and R for advanced statistics and calculations for the late control strategy. The transcription of the interviews was performed with ListeNwrite.

2.3 Participants

A total of 29 end-users were recruited as a convenience sample. 15 participants were evaluated face to face (Group G1) and 14 were evaluated by using a communication tool (Group G2). In detail, the total distribution consisted of eight workers from a near workplace A (G1), seven from another near workplace B (G1) for face to face testing, eight from a distant workplace C (G2), three from another distant workplace D (G2) and finally three from a distant workplace E (G2). In group G1 (*n* = 15) six males with an average age of 47.16 years (*SD* = 7.75) and nine females with an average age of 39.55 years (*SD* = 13.52) participated. 26.66% of the participants of G1 graduated either from a University or a University of Applied Science, while 66.66% did their apprenticeship as a skilled worker. One person (6.66%) had no further education beyond secondary school. In this group the average experience using the software was 9.06 years (*SD = 1.53*; *min* = 5; *max* = 10). Four of the participants in G1 used the mouse to interact with the interface, 11 participants used both keyboard and mouse. None of the attendees limited oneself to keyboard shortcuts. In Group G2 (*n* = 14) one male at the age of 60 years and 13 females with an average age of 47.61 years (*SD* = 9.40) took part. In group G1 all participants were trained as skilled workers, no one had a higher level of education. In this group the average experience using the software was 7.71 years (*SD* = 2.43; *min* = 2; *max* = 10). Three of the participants in G2 used the mouse to interact with the interface, 11 used a combination of both. Just as in G1 no one confined him- or herself to keyboard shortcuts.

All participants had normal or corrected vision and hearing. The study was approved by the ethics committee of the University of Twente. Furthermore, participants got approval of their department to participate in the experiment and were freed from regular tasks for the duration of the experiment. All participants handed their written consent prior to the test trial in and were fully informed about the goals and procedure of the experiment. Participants had the option to receive a short overview of the results as a reward for participation after the study was finished. Demographics of all participants were recorded prior to the test.

All members of the sample were living and working in Germany. The complete data was gathered in German since it is the native language of the participants and the company. This reduces impacts of participants who are not fluent in the test-language.

2.4 Procedure

All test trials were executed at the regular workplace of the participant for economical reasons. All participants had a computer screen, keyboard and mouse placed in front of them on a table. Group G1 was tested first while G2 was tested later. In each group first a set of five participants was tested, afterwards a set of three. In G1 then a set of four and a set of three participants was examined. In G2 two sets of three participants were tested. After each set the data was transcribed and analyzed.

2.4.1 Procedure G1

Group G1 was tested face to face. The researcher visited the participant at his or her workplace for an appointment arranged by the company. The participant was informed about the procedure and the written consent of the experiment. According to the written consent an audio and video-recording of the display was set up by placing a camera behind the participant. The pre-questionnaires (Appendix B) were completed by the participant on a laptop, after which he or she was asked to connect the software-client to the test-environment. Finally, he or she was instructed to complete the tasks as well as he or she was able to. If he or she was not able to complete the task, he or she should name and explain the steps he or she would normally do. The first task was read aloud by the test-leader and executed by the participant as good as possible while giving a CTA. After the first task the participant was requested to reflect less than one minute on how he or she thought he had done in the task in a RTA. This was done for the reason that learning-effects from the tasks could bias the RTA if only given afterward finishing all the tests. Afterwards the second task was given, repeating the procedure until all tasks were performed. No hints on the tasks were given to the participant. If the task could not be completed the trial was terminated after five minutes or by request from the participant. In this case the participant had one minute to reflect on the task before starting the next one. In the case of problems with the think aloud protocols additional information and instructions were given to the participant. After finishing the tasks participants got the post-questionnaires (Appendix C). After completing the post-questionnaires the participants took part in an interview (Questions in appendix D). Next the interview participants were asked to answer the questions of FIP which was developed based on the possible usability problems found in the previous tests. Only the first five participants of each group (G1 and G2) were excluded from FIP because there were no sufficient data for setting up the questionnaire. After this the participant had time to add information to the questionnaires and propose their own opinion about what should be done to improve the usability of the interface.

2.4.2 Procedure G2

Group G2 was tested via the tool Screenconnect. The researcher mailed an invitation for participation in a conference session via e-Mail to the participant at the moment of a previously arranged appointment. The participant was welcomed and informed about the procedure and the written consent of the experiment. The written consent was attached to the mail and the signature was gathered via an online questionnaire. According to the written consent a video-transfer of the desktop of the participant was set up through Screenconnect, recording the screen simultaneously. To tape audio the application Voicerecorder came into action. All questionnaires and the written consent were given in a digital version via a link attached to the e-Mail. The pre-questionnaires (Appendix B) were given to the participant in digital form and completed. The participant was asked to connect the software-client to the test-environment and instructed to complete the tasks as well as possible. The procedure to complete the tasks was identical to the procedure of group G1. After completing the post questionnaires the participants took part in an interview (List of questions in appendix D). After the interview the participants got a set of questions of FIP which were developed based on the possible usability problems found in the previous tests of both groups (G1 & G2). After this the participant had time to add information to the questionnaires and propose their own opinion about what should be done to improve the usability of the interface.

2.5 Tasks

Because of the total number of more than thousand different possible actions a preselection of tasks had to be made. For this study two sets of three predetermined, representative tasks were tested. Set 1 had a longer duration than set 2 as the participants were in the need of typing more data. The set included a starting task and two follow ups. It was not necessary to fully complete the starting task to be able to participate in tasks two and three. The choice of tasks being convenient for this research and requiring usage of most of the graphical interface was done in cooperation with the company. The employees being involved in the preselection of the tasks did not participate in this study. The visual appearance of the interface did not change while performing different tasks. The first five participants of each group got the larger set of tasks (set 1) to make the results comparable. A list of individual task assignments can be found in appendix E. The time for completion was minimal 2:35 minutes and maximal 35:15 minutes.

2.5.1 Description of tasks

In set 1, first the attendee had to create a new data-file using the application. In this procedure he had to fill in a great number of different fields and options, taking at least about eight minutes to finish this task. The second task in this set was to find and open a list being located in the application for each user as part of the interface. The third task was to search for the newly created data-file in the software and open it. If the user would not be able to complete the first task, a data-file was already prepared. This never occurred.

The second set of tasks started with the task to find and open a prepared data-file. The second task was to edit a part of the data stored in the file. The third task was to create a new entry in the data-file or edit the entry if already preexistent.

2.6 Questionnaires

During this study the participants had to complete different pre- and post-test questionnaires. The pretest questionnaire (Appendix B) was used to collect demographic data like age, gender and educational level and work related data like the years of experience using the software, period of company membership and preferred style to interact with the software. After completing the total set of tasks and reflections a post-questionnaire was given to the participant (Appendix C), dealing with the personal feelings they had while giving the CTA. The answers had to be given on a five point Likert-scale, with five being the most positive and one the most negative option. All questions were formulated positive to avoid confusion of participants.

2.7 Interview

The semi-structured interview contained a set of eight different questions regarding the visual appearance of the interface, the interaction while working with the software and the personal feelings of the participant towards the software. Each participant took part in the same interview-survey, being able to individually add further thoughts about other topics. The interviews had a duration of minimal 8:33 minutes and maximal 45:38 minutes (G1: *Min* = 10:34minutes; *Max* = 45:38minutes; *Mean* = 24:38minutes; G2: *Min* = 8:33minutes; *Max* = 29:27; *Mean* = 17:37 minutes).

2.8 False-positive Identification Protocol

The False-positive Identification Protocol (FIP) (Appendix F) was developed for this study to identify false positives in the gathered data while performing the study. Possible usability related problems found in the data are listed and formulated to hypothetical statements or questions. Possible usability problems are problems that are mentioned only once by one participant. Problems that could be verified by a participant demonstrating it to be a problem in the actual software were regarded as a real problem even if they were found only one time. Following participants were informed that the questions are hypothetical. They were asked to give their opinion about the statement and whether they agree with the statement or disagree, also asking for reasons for their choice. Except for the first five attendees of each group every participant got a small set of different questions of FIP. In case no one agreed to a single statement or question, the entry was removed from the list after three participants and flagged as being probably no problem. Those problems were added to the list of false positives. If the problem was mentioned or occurred in a later test it was returned to the list of problems. If, in total, two or more participants agreed to the problem it was added to the list of problems found. If a problem turned out having the need of being reformulated, this was done and validated on the following participants again.

2.8.1 Analysis for false positives

All codes were entered into an error-matrix for each participant’s combined data (See section 2.10 for the explanation of the used coding method). A count was made to monitor the number of observations being found for the specific code. Problems found by observations like slips and lapses in the video data were marked as definite usability problems even if they had only one finding. Codes with two or more observations were flagged as real usability problems as more than one participant had problems with it. Codes with only one observation were formulated into a hypothetical question or statement and added to FIP. If after three participants at least one additional participant approved the problem, the problem had an observation-score of two or more and was regarded as a true problem. If at least one person had falsified the problem and no approval was found the problem was flagged as most likely a false positive. If a participant demonstrated the problem was no real one or the description had to be different it was revised or regarded as a false positive based on a problem different than usability like communication problems. If the problems description was revised or no participant had an opinion on the hypothetical statement or question an additional set of three participants was asked based on the same rules. After this second set the hypothetical question or statement was dropped as a false positive if not confirmed. All false positives were added to the error-matrix with the tag ZSFP in front of the name of the code. Problems that could not be evaluated at the end of the data-collection were flagged as ZMFP. The letter “Z” was selected to put these problems at the bottom of an alphabetical list, while “S” was used to flag definite problems (German: “sicher”). Respectivly, “M” intended to flag possible problems (German: “möglich”). “FP” is a shortcut for false positive.

2.9 Analysis

For analysis of the progress of problem detection *D* the late control strategy suggested by Schmettow (2009), Vos (2011) and Schmettow, Vos and Schraagen (2013), already described in section 1.4, was used. The variance of defect visibility was taken into consideration and a confidence interval of 80% was used. The number of undetected problems was monitored using this method, too. The sample was tested on the type of distribution (normal versus not normal distributed) by using graphical and statistical methods. The Kologomorov-Smirnov-test was used for statistical analysis of the distribution. The groups G1 and G2 were compared by using descriptive statistics on differences in the detection rate. Therefore only the data of the video-analysis, the CTA, the reflections, the interviews and FIP were used. The data gathered from the questionnaires were not included in this comparison.

2.10 Coding and mapping

As described in section 1.2, coding and mapping are important to mention if data are compared. Coding and mapping are the base on which the data are analyzed. It is supposed to be identical for all data. In this section the coding and mapping of the different data is described for each part of the data separately.

2.10.1 Coding and mapping of audio- and video-data

After the first tests were gathered the audio files were checked for completeness and transcribed for further analysis. The data gathered with FIP were transcribed and added to the interview-data. The analysis was done by marking each observation in the transcript or video data that could be a signal to a possible usability problem in atlas.ti using an open template like Cassel and Simon (2004) describe. This template is constructed based on the observations found in the data. Single occurred or uncertain data were added to the FIP if needed. Other problems with more than one occurrence were gathered in a list of problems found. Problems within the state of possible usability problems not confirmed by FIP or dismissed were added to a separate list in order to not mixing up problems with possible false positives. This process was iterated for the whole set of 29 participants.

Problems were mapped based on the approach of similar solutions (Hornbæk & Frøkjær, 2008). Problems being solved by identical changes in design were mapped to each other. This was done on a very high level of detail to be able to give specific solutions to problems.

2.10.2 Coding of questionnaires

The first questionnaire (Appendix B) was analyzed using box plots and descriptive statistics. Due to the problem of users not understanding parts of the second questionnaire (Appendix C), those data were not used and analyzed any further.

2.11 Analysis of completeness/discovery rate

To analyze the number of problems discovered and, respectively, undiscovered at given point in the research within an 80% confidence interval a matrix similar to the one used by Vos (2011) was developed. The matrix contained a full set of all problems found until the specific point of research. For each participant a score was added into the matrix, with 0 indicating the problem was not found by this participant and 1 contributing to the problem being found. The negative value -1 was used in case of a participant identifying a problem as a false positive. This way the matrix provided a display of which problem was found by whom. Furthermore, using this matrix the frequency of how often a problem was reported may be estimated from the total population of participants.

This matrix was imported into a program for R developed by Martin Schmettow and based on the LNBzt model described by Schmettow (2009) (See paragraph 1.4). This model is able to account for the variance of defect visibility and the undiscovered problems simultaneously within the used confidence interval of 80%. To receive the full set of data for all subgroups G1 and G2 as well as the total group Ges, each group was analyzed individually, starting with the subgroups. Both subgroups could be the initial value for the analyses and were able to provide a perspective on the further development of the other subgroup if combined to group Ges. This way the progress of problem detection *D* became visible over the full study and sample.

2.12 Additional analysis

From the gathered data an additional sample was created containing only the first half of participants (n = 16) of each test-location excluding false positives. This sample was analyzed the same way as the other groups using the late control strategy to explore whether a combination of the two methods would lead to the same results as each group on its own. A second additional analysis was done on the first four participants of group G1 to investigate how the early control strategy performs compared to the late control strategy.

2.13 Priorization

Problems that were reported more than 7 times were expected to be high priority. This threshold was chosen because at this point approximately one quarter of the participants mentioned the problem.

3. Results

In this section the results of the current study will be stated. This section is divided in seven parts: First general methodical findings of this research will be explained shortly before the general statistics of the usability tests are presented. In the third part the findings for the two groups G1 and G2 will be compared to each other while in the fourth part the results of the late control strategy are described. Afterwards the combined and additional findings are described and in the end the results for the usage of FIP will be stated. The detailed results of the usability test are not included in this section. For examples a selected list of discovered problems can be found in appendix G. This list includes some examples for false positives found by FIP as well as translated transcripts.

3.1 General methodical findings

In this research different questionnaires were used. One questionnaire - the Post-questionnaire - was designed to find influences of the testing-situation on the results. Many users got confused by the Likert-scale. They did not know whether the 'normal' value was '1' or '3'. This resulted in a faulty completion of the questionnaire. Based on this the questionnaire was gathered on all participants but not analyzed for further usage.

In the test two different scenarios were used to investigate the usability of the software. The effect on the results was tested using the correlation coefficient by Pearson on the set of tasks and the total findings per user. For the findings per user including false positives the Pearson correlation coefficient was -.10, just as when excluding false positives, meaning that the effect can be ignored because it is only small or not existing

To test the functionality of FIP, one observation definitely not being a problem was first gathered along with the data. This observation was flagged 'Hinweise: inflationäre Warnhinweise' and later excluded from the data.

3.2 General statistics of findings in the usability test

In total, 90 usability problems were found in the usability tests. After deleting all false positives using FIP a total of 75 problems remained, resulting in 15 false positives. In this tests false positives made up 16.66% of the total problems. All in all, users found between eight and 22 problems per head, including false positives (*mean* = 15.41; *SD* = 4.42). After deletion of false positives users discovered minimal six and maximal 21 problems (*mean* = 14.68; *SD* = 4.12). Including false positives users found at least zero and at most 15 problems being undetected at this point of research (*mean* = 3.06; *SD* = 3.46). When false positives are excluded each user found between zero and 14 new problems (*mean* = 2.58; *SD* = 3.01). Each problem was reported by at least one and at most 22 individual users (*mean* = 4.96; *SD* = 4.17). Of the 75 problems found 22 were categorized as high priority since they were mentioned by seven or more participants. This results in 29.33% of the total problems having a high priority.

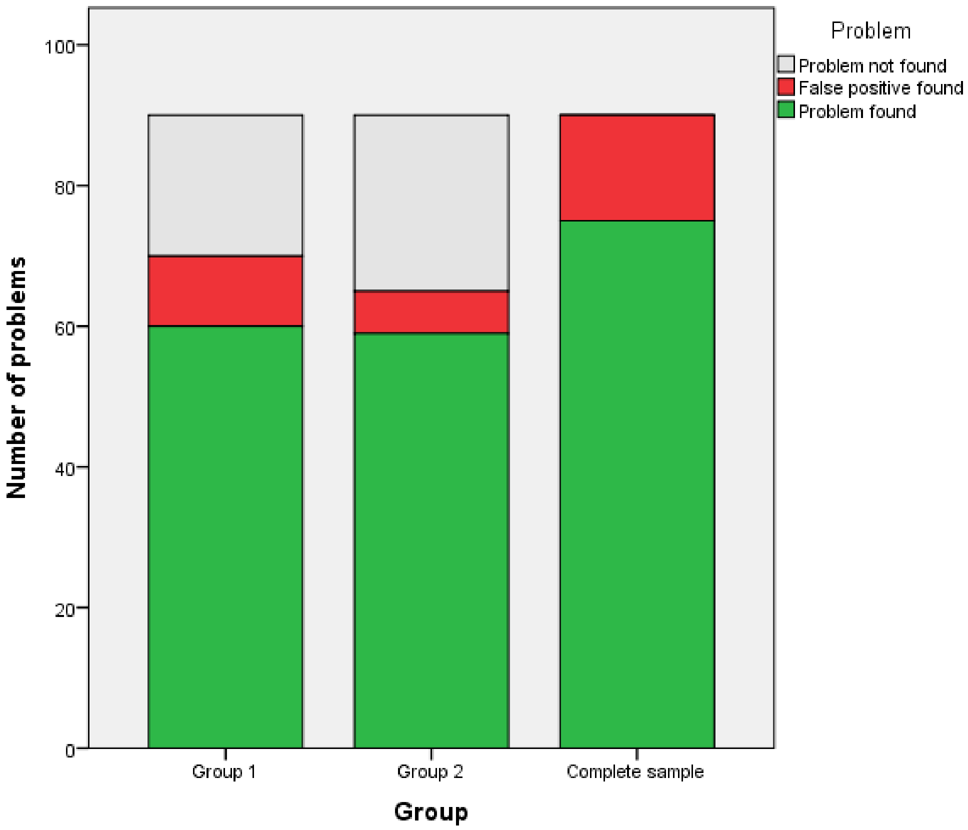


Figure 3.1. Figure of problems found for each group and in total are displayed in green. The false positives per group and in total are displayed in red. The undetected problems are displayed for each group in gray.

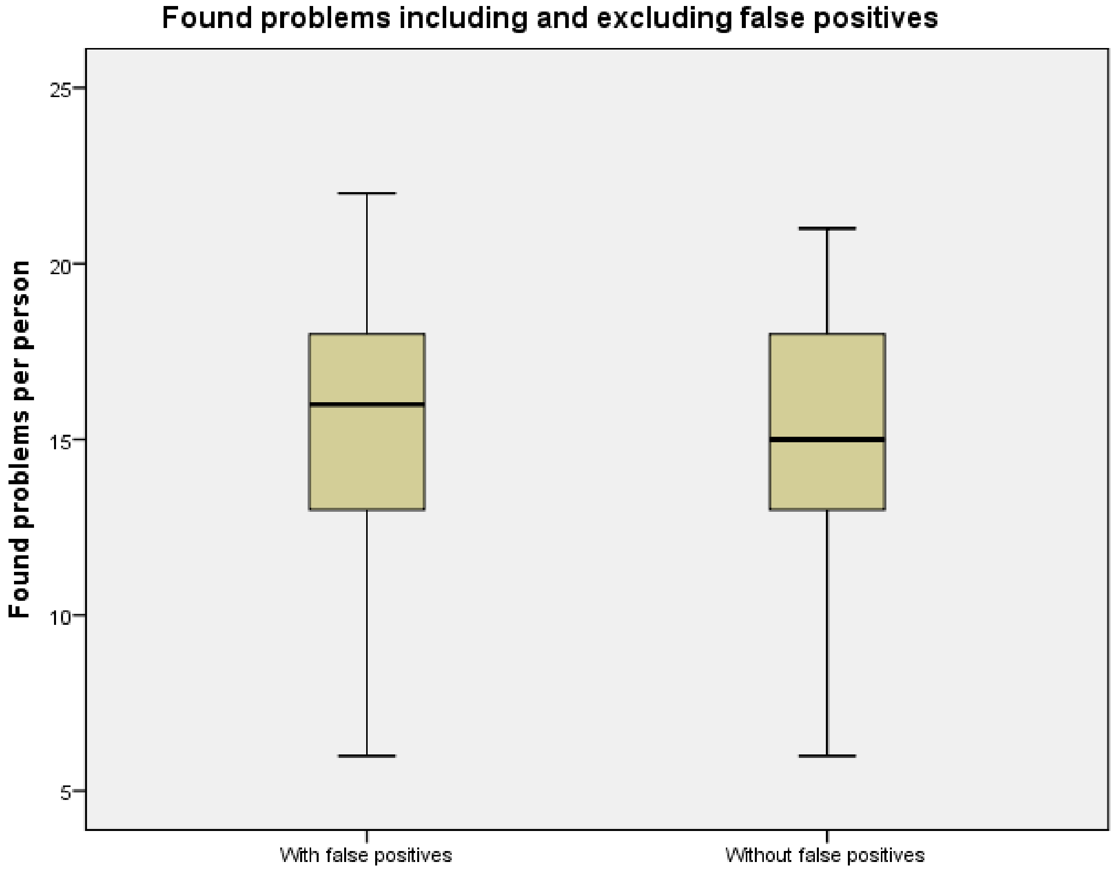


Figure 3.2. Figure of the distribution of problems found per person in Group Ges on the left side including false positives and on the right side excluding false positives.

3.3 Comparison of group G1 and group G2

Group G1 found 70 problems in total, implying 10 false positives, resulting in 60 real problems. Therefore group G1 found 77.77% of the problems including false positives and 80% of the real usability problems without false positives. The false positives turned out to be 14.28% of the findings in group G1. In group G2 a total of 65 problems including six false positives was discovered, resulting in 59 real usability problems. Group G2 found 72.22% of all problems including false positives and 78.66% of all problems excluding false positives. False positives made up 9.23% of the findings in group G2.

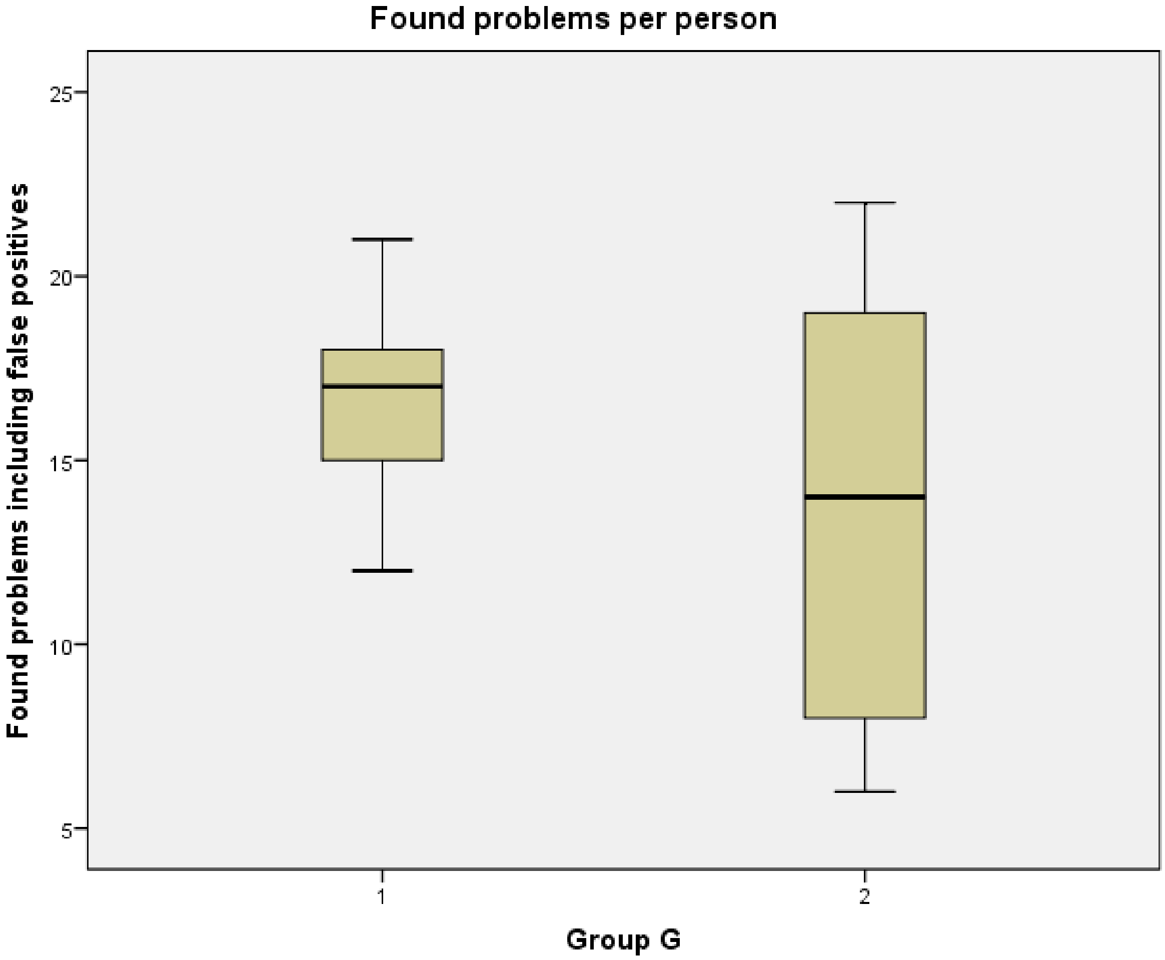


Figure 3.3. Figure of the distribution of the number of problems found per person including false positives per group.

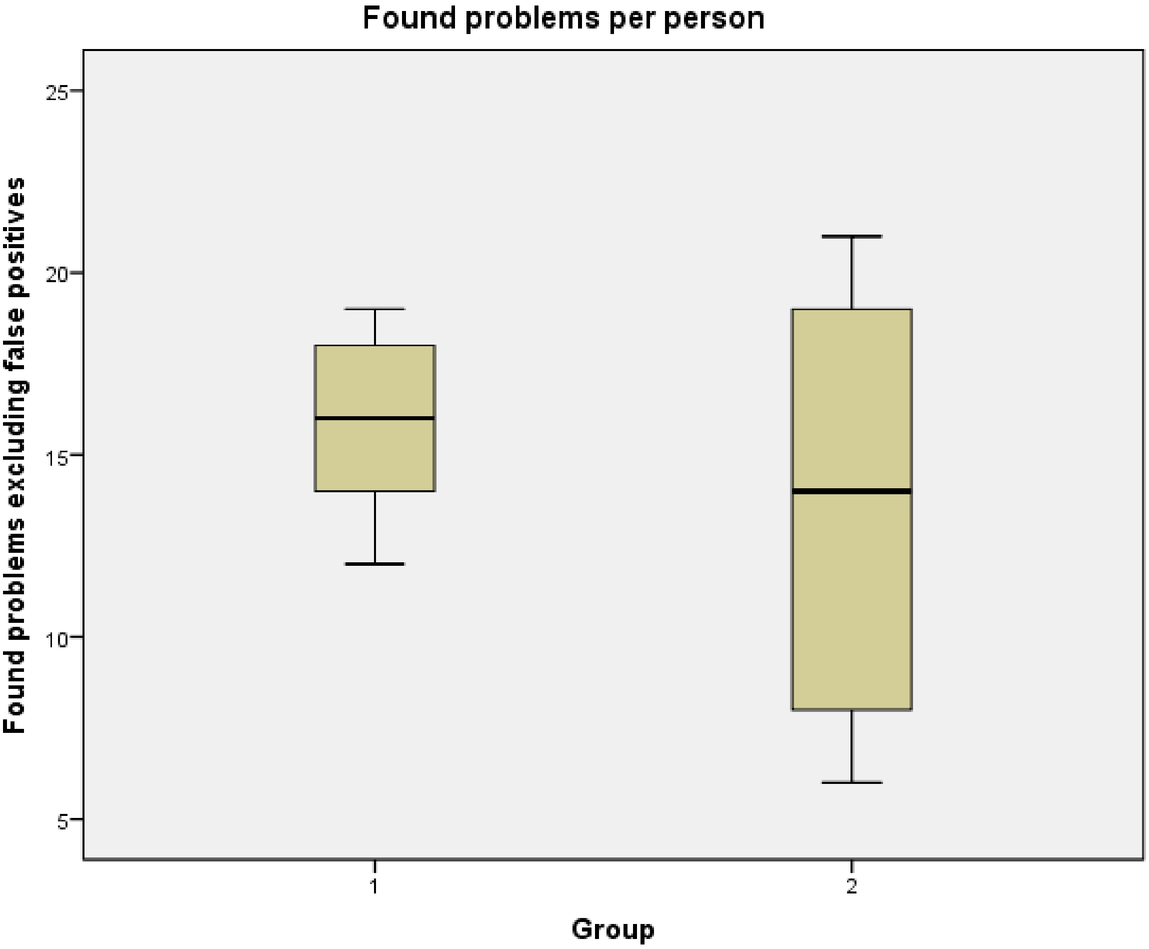


Figure 3.4. Figure of the distribution of the number of problems found per person excluding false positives per group.

Each user of group G1 found minimally 12 and maximally 21 problems including false positives (*mean* = 16.86; *SD* = 2.50). If false positives are excluded each user of G1 found between 12 and 19 problems (*mean* = 15.86; *SD* = 2.32). In this group each user found at least one and maximal 15 new problems of the set of problems discovered in G1 including false positives (*mean* = 4.66; *SD* = 3.75). Excluding false positives each user discovered between one and 14 new problems (*mean* = 4; *SD* = 3.27).

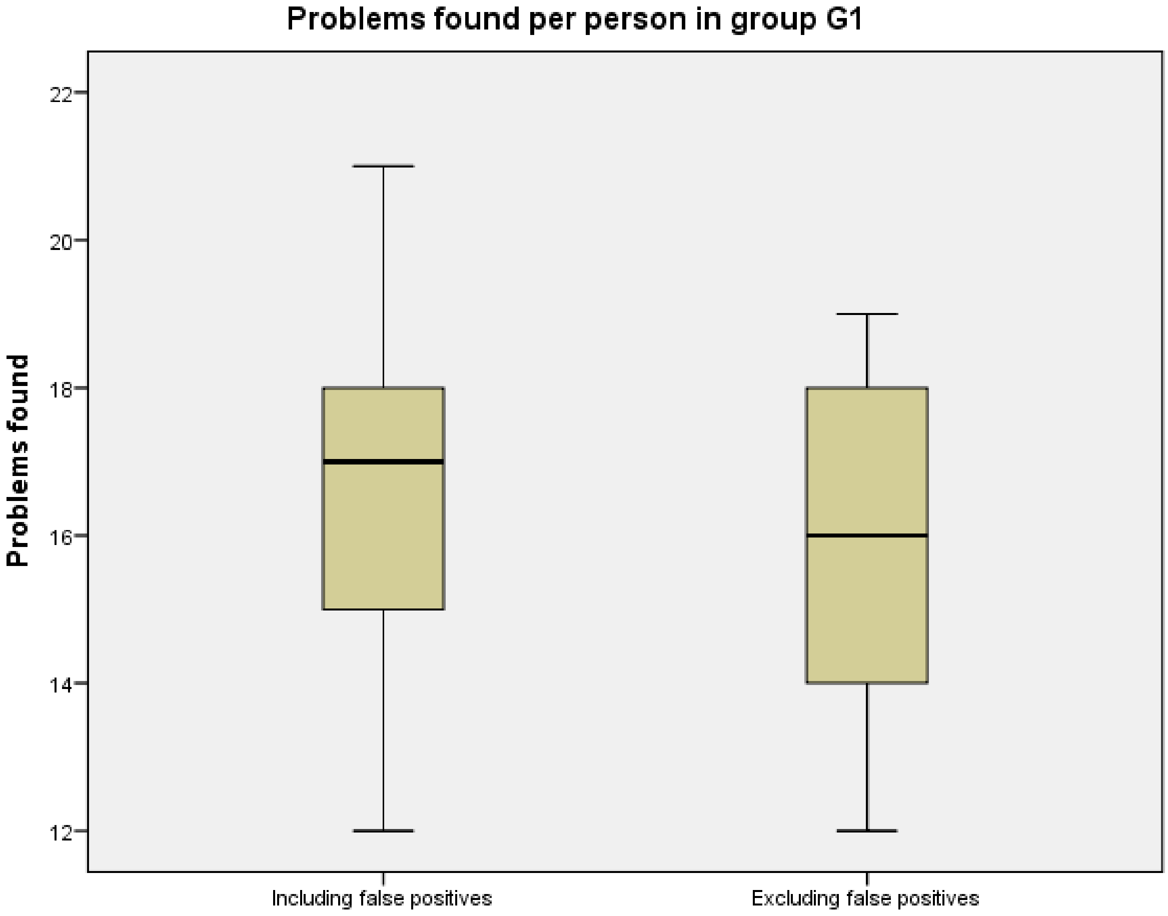


Figure 3.5. Figure of the distribution of the number of problems found per person in- and excluding false positives in group G1.

In group G2 each user found minimally six and maximally 22 new problems including false positives (*mean* = 13.85; *SD* = 5.51) and between six and 21 problems when false positives are excluded (*mean* = 13.42; *SD* = 5.24). In group G2 each user found minimal one and maximal 14 new problems of the problems found in G2 including false positives (*mean* = 4.64; *SD* = 3.99) and between one and 14 new problems of the problems of this group if the false positives are excluded (*mean* = 4.21; *SD* = 3.90).

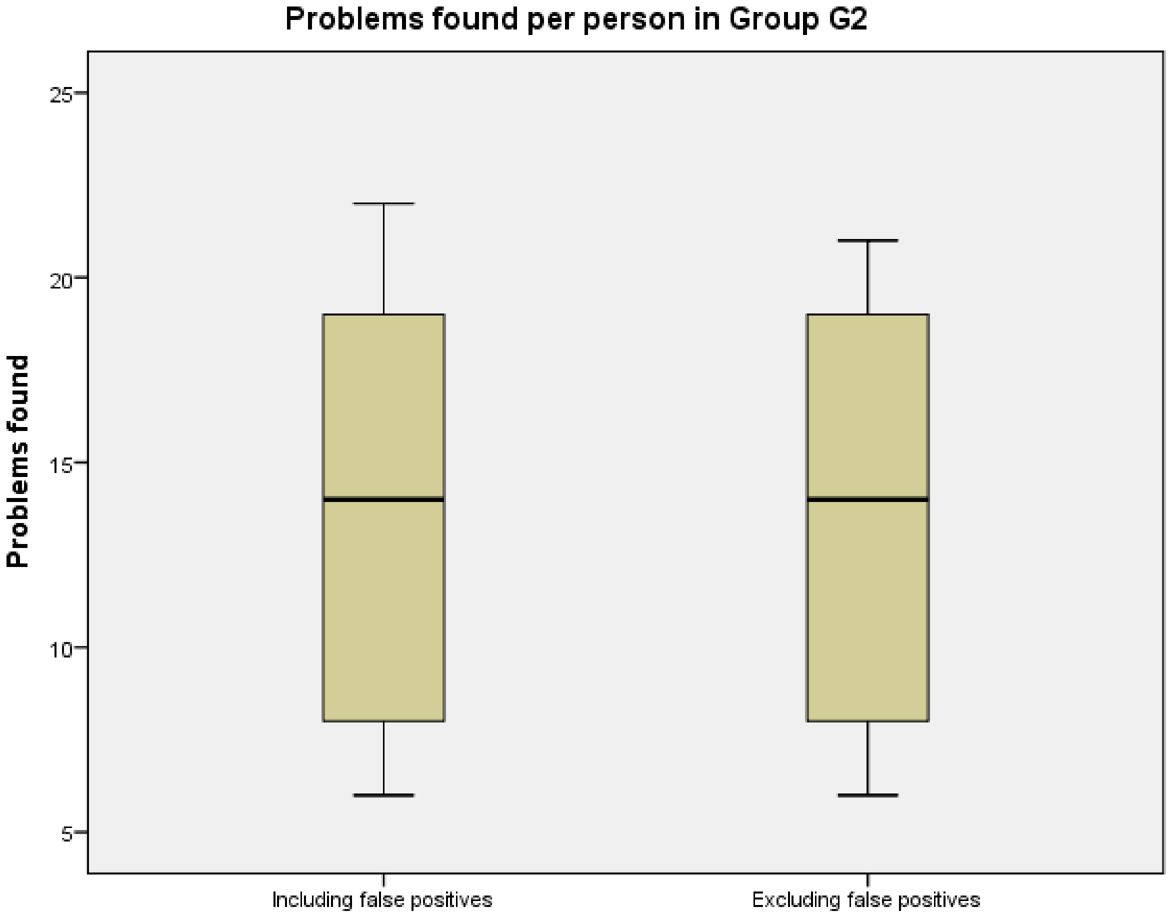


Figure 3.6. Figure of the distribution of the number of problems found per person in- and excluding false positives in group G2.

Group G1 found 25 problems group G2 did not find. Of these 25 problems nine were false positives. Group G2 found 20 problems group G1 did not find, inheriting five false positives. Resulting, group G1 and G2 found a similar number of problems, yet they discovered different problems. Furthermore, most of the problems found in only one of the groups were found between one and four times. Each group detected one problem with a priority-score of five the other group did not find. Only group G1 found one problem with a score of seven, remaining undiscovered in group G2. Overall, the difference between the problems found mainly emerges in low priority problems.

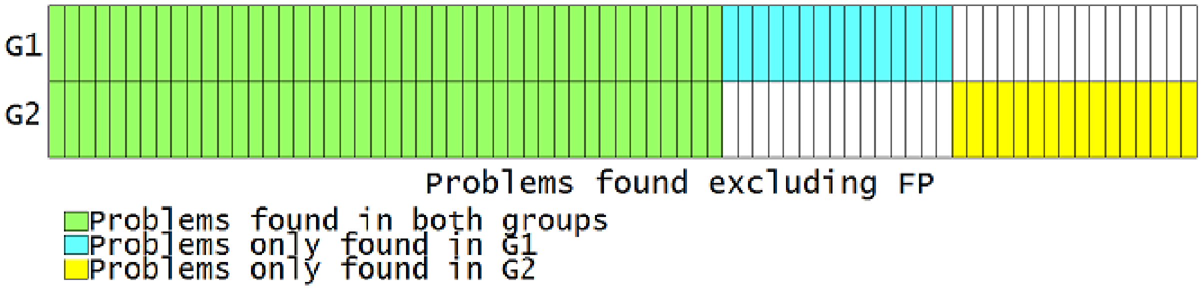


Figure 3.7. Number of detected and undetected problems per group.

The correlation coefficient by Pearson for the correlation between group and the total findings per user including false positives is -.34 and for the findings excluding false positives -.30. Both indicate a weak negative linear relationship between group and findings per user.

The distribution of the samples was tested for normal distribution using graphical and statistical tests. The outputs of the statistics program SPSS for a 95% significance level of these tests are listed in Appendix H. The results of group G1 including false positives graphically follow a normal distribution. The results of the Kolomogorov-Smirnov-Test on 95% confidence support this assumption with an extremest deviation of .188 on *n* = 15 and an asymptotic significance of .162. An extremest deviation of .338 or higher would mark the threshold of not being a normal distribution when considering 15 attendees. Excluding false positives the optical test cannot verify a normal distribution. The Kolomogorov-Smirnov-Test found an extremest deviation of .256 on an asymptotic significance of .009. This cannot support the assumption of a normal distribution either. For group G2 including false positives the graphical test could not answer the question of distribution explicitly. The Kolomogorov-Smirnov-Test indicates a normal distribution with an extremest deviation of .142 on *n* = 14 and an asymptotic significance of .200, being limited by an extremest deviation of .349 considering 14 people to verify normal distribution. If false positives are excluded the optical test cannot answer the question either. In this case the Kolomogorov-Smirnov-Test supports the assumption of normal distribution with an extremest deviation of .142 and an asymptotic significance of .200, too.

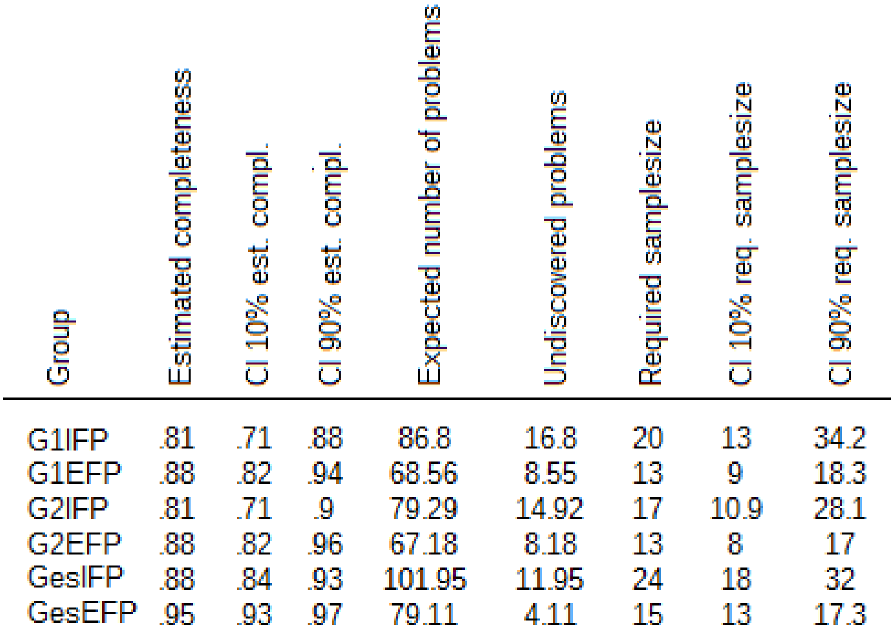
Concluding, based on the results of the descriptive statistics there is no difference in the final detection-rate of both groups. Both of them found a similar number of problems and a similar number of problems the other group did not find. However, the results are not always normally distributed and there is a difference between the individual detection-rate of participants in group G1 and G2. In G2 the difference between the number of problems found per participant is greater than in group G1. Furthermore, on average participants of group G1 found more problems per person than group G2 because of low performance of some participants of group G2.

3.4 Results of the late control strategy

The late control strategy was used for six different data-sets separately. First it was used for the total set of problems found, in each case including and excluding false positives, and then for both groups G1 and G2 separately, respectively including and excluding false positives. The results are listed below. (For a full list of the R-outputs see Appendix I.)

Table 3.1

Results of the late control strategy per group



Group	Estimated completeness	CI 10% est. compl.	CI 90% est. compl.	Expected number of problems	Undiscovered problems	Required samplesize	CI 10% req. samplesize	CI 90% req. samplesize
G1IFP	.81	.71	.88	86.8	16.8	20	13	34.2
G1EFP	.88	.82	.94	68.56	8.55	13	9	18.3
G2IFP	.81	.71	.9	79.29	14.92	17	10.9	28.1
G2EFP	.88	.82	.96	67.18	8.18	13	8	17
GesIFP	.88	.84	.93	101.95	11.95	24	18	32
GesEFP	.95	.93	.97	79.11	4.11	15	13	17.3

Note: Results of the late control strategy per group (G1 and G2) and total (Ges). IFP indicates that the results are including false positives, EFP indicates that the results are excluding false positive.

All samples had an estimated completeness of at least 81%. When excluding false positives, the completion rate was increased by seven percent in all cases. The confidence intervals were also smaller when false positives were excluded. When false positives were excluded, the expected total number of problems was smaller in each case. For the total number of expected problems in group Ges, the exclusion of false positives reduced the expected number by 22.84 problems from 101.95 to 79.11. In all groups the number of estimated problems being undiscovered dropped when excluding false positives. The estimated required sample size to find 85% of the problems was also reduced when false positives were excluded, holding true for the confidence intervals as well. In total, the exclusion of false positives led to a reduction of all values including the confidence interval.

The estimated completeness of groups G1 and G2 turned out to be identical with 81% for G1 including false positives and 81% for G2 including false positives. When false positives were excluded the estimated completeness was 88% for both groups. The confidence intervals were alike, too. In contrast to this, the required sample-size to find 85% of the problems differed between G1 and G2 when false positives were included with 20 people for G1 and 17 for G2. If false positives are excluded the required sample-size yields to 13 people for both groups.

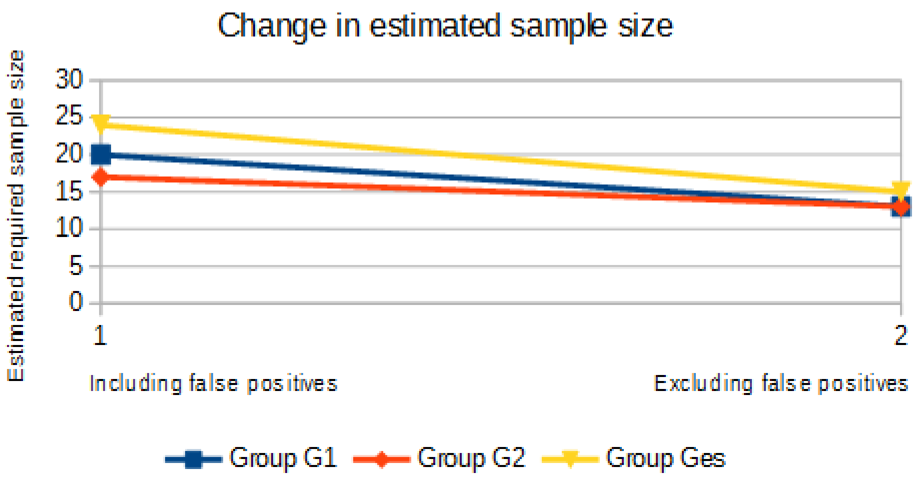


Figure 3.7. Graphic illustrating the change in the estimated required sample size when excluding false positives per group.

3.5 Combination of different results

The results of the descriptive statistics show that there is a great chance of not gaining any significant difference in the finding rate of groups G1 and G2. Regarding effectiveness, the late control strategy supports the suggestion that both groups have the same behavior if false positives are excluded and lead to overall similar results in this method. However, the fact that, even if they found a similar number of problems, they sometimes found different problems leads to the conclusion of both groups only having a statistical similarity, yet not practical. There is no difference in how much they found but in what they found. This can either imply the different approaches used for G1 and G2 somehow have an effect on the test, or the samples have different problems using the application. This has an effect on the total number of problems in the combined group Ges.

G1 and G2 are combined for the total sample Ges. Each group has nearly half of the participants of the total sample. Only the order in which they are tested defines which of the two groups is the 'first half' of the total sample. In the late control strategy it needs to be possible to predict the required sample size based on half of the sample. In both groups a sample size of 13 is predicted, while the predicted sample size for Ges is 15. Despite the differences between groups G1 and G2 the expected sample size for Ges only differs slightly being still covered by the confidence intervals of all groups for each other. However, Ges has a higher expectancy value of problems, being 10.55 problems more than G1 excluding false positives and 11.93 problems more than G2 predicted. Thereby the conclusion for the late control strategy not being able to fully compensate differences between sample-groups is drawn. Only the dimension of the sample size can be predicted likewise, but not the expected number of total problems.

3.6 Analysis of the additional groups

The analysis of the additional created group, containing only the first half of participants of each location excluding false positives, found a completion rate of 80% (CI 10% = .72; CI 90% = .88). The expected number of problems is 87.74 with 17.74 undiscovered problems. The required sample size is 22 (CI 10% = 31.1; CI 90% = 17.1). Those results are, even if false positives are excluded, more similar to the results of group G1 and G2 including false positives than to the groups excluding false positives.

The investigation on how the early control strategy performed estimating the required sample size based on the first four participants of group G1 failed. The estimated required sample size is 384 persons and an error being displayed regarding the integral indicates it to be probably divergent. This makes the results of the early control strategy unusable for practical use.

3.7 Results of FIP

In total the participants were asked 50 different questions, excluding the first five participants of group G1 and Group G2. Each participant answered at least one question, with the preference of not answering more than five if preventable. The complete list of questions can be found in appendix F, examples with transcripts are listed in Appendix G. 15 out of 50 questions were confirmed to be false positive. It should be mentioned that one of the 15 is only regarded as false positive because it was found at the end of the tests and could not be confirmed as an actual problem. Another one of the false positives contains all findings of bad mood and feelings of participants that were mentioned as problems but confirmed as a personal opinion or mood.

Different categories of false positives were identified. Following the most important categories identified by FIP are described, also providing an actual example found in the current study for each category. It should be mentioned that also false positives were identified that did not belong to this category, for example based on background-information and corresponding logic.

One category of false positives was confirmed to be dependent on the test scenario. For instance, a problem was found for users who were unable to find the correct button to open the application in their Windows task bar. This was confirmed as false positive by asking other users whether they knew this problem or not. Based on the answers the icon of the button turned out to be different in the test than the actual icon known to the attendees, resulting in participants searching for it.

Another category of false positives was based on faulty ideas of users on the functionality of the application. Users being informed about the working principle confirmed the status as false positive. For example, one user mentioned a dataset had to be reloaded first in order to see changes in the output. This was confirmed being a false concept by another user, demonstrating the opposite. The changes in the dataset had to be saved before they appeared in the output. Reloading the dataset implied saving, resulting in the misconception. These sorts of problems were regarded as false positives for usability problems, actually being real problems based on communication.

A third category of false positives was based on the misunderstanding of the researcher who, for example, misunderstood a text window popping up if a dataset was closed. Most participants closed it without noting anything in the textfield. Later users explained this field just was not used in the test, normally having the reasons for dataset changes being inserted in the window. Instead, the researcher had misunderstood this as a window falsely popping up.

The remaining 35 potential false positives were confirmed to be actual problems by using FIP. Some of them were simply confirmed by users finding it was an actual problem, while other problems were wrongly described in the first instance and turned out to be an actual problem with a slightly different description after asking other participants. In those cases the first description would have been a false positive. The enhanced description included the first explanation, but also stated the actual problem at hand.

A methodological problem found was the need for a proper description of the problem to formulate an appropriate question. Sometimes participants misunderstood the question and answered accordingly. Another problem regarding the methodology was the idea to request more information of participants if they confirmed a problem via FIP at first, like the impact and the frequency of occurrence. This procedure was canceled after a few tries because participants got irritated by the additional questions. This especially occurred if the participant did not spend much thought about the problem until the point of discussion, therefore not knowing an accurate answer and resulting in defensive reactions.

Using this protocol it was possible to identify reported problems that were wrongly described or not correct by asking other participants for their thoughts about the problem. Those false positives would have spoiled the list of problems and had great impact on the results of the method, as shown in the analysis of the late control strategy. In all cases the results became more defined when false positives identified by FIP were excluded from the analysis.

4. Discussion

This study attempted to answer the question of how the late control strategy performs in a case study involving a real world scenario with different difficulties like the accessibility of the sample and false positives. Therefore, two groups were tested, one face to face and an other one via a communication tool. Furthermore, a protocol to identify false positives in the data gathered was developed and tested. The results show that the two groups identified similar total numbers of problems, but both groups found different problems to a given degree. Group G1 found problems group G2 did not, and the other way around. The newly developed protocol FIP was able to identify different false positives originating in misconceptions, misunderstandings, personal opinions and problems in the test setup. Excluding those, false positives generally had a positive effect on the results of the late control strategy, narrowing the confidence intervals and lowering the extent of the estimated required sample size to a given degree. Furthermore, after excluding false positives from the data the late control strategy was able to give a more precise prediction of the experimental progress. Even if group G1 and group G2 differed in the qualitative results they found, the required sample sizes being predicted for each group were close to the required sample size estimated for the total sample.

4.1 False positive identification and the effects of it

In the current study the existence of false positives in the data was verified. The new developed *False-positive-identification protocol* (*FIP*) identified different false positives by asking users to verify or rather falsify the objective. Sears (2009) stated that false positives would be mainly detected in data where the evaluator is a novice. In the current study some false positives occurred only because the test-leader was a novice to the situation and the test-object as well, leading to misunderstandings. This supports the statement of Sears (2009). Yet, not all false positives originated from this. Furthermore, just as Sears (2009) described false positives were found when users were not used to a specific task. However, different false positives were identified on expert users as well, making it hard to give a general statement on the relation of false positives and the experience of the user.

Gray and Salzman (1998) described a tradeoff between declaring a feature to be no problem and the possibility to miss an important problem. When finding a problem, for example only reported once and thus not definitely a problem of the product but maybe of the person reporting it, it could be a false positive or an actual problem only occurring by chance. The different problems being confirmed while using FIP validate the assumption that it would be disputable to simply remove all problems only found once. This is because 35 of the total 50 checked FIP-questions were confirmed actual problems. Removing them would have led to 35 undiscovered problems potentially being harmful. This is in line with the description of

Woolrych, Cockton and Hindmarch (2004) that a true problem could be removed from the data wrongly when erasing all singletons without checking them carefully. Concluding, it is necessary to check carefully which problems occurring only once are false positives and which are not.

The present research attempted to improve the method for identification of false positives used by Schmettow, Vos and Schraagen (2013) to optimize matching with the late control strategy. While they used the '*triage'* after the data-gathering was finished, the FIP used in this research is a systematic method to identify false positives parallel to the collection of data. One characteristic of the late control strategy is the possibility to monitor the process of data collection at any moment and to terminate the data collection when a desired goal is reached (Schmettow, Vos & Schraagen, 2013). FIP is compatible with this characteristic of the late control strategy to be able to keep track of the sample size needed as well as the completion rate while gathering the data, while the '*triage*' could not support this characteristic since it was used after the data collection was finished. Schmettow, Vos and Schraagen (2013) identified 13% of their results as false positives while in this research around 16% of the full sample were identified as false positives. Despite the somewhat higher percentage it is not likely that the FIP is more effective in identifying false positives. Yet, only the sample and test differs. This assumption is based on the knowledge that some of the identified false positives in the current study emerge from the setup of the test.

An unexpected finding was FIP being able to identify different categories of false positives like personal misconceptions of participants, test related false positives and wrongfully described problems that had to be described differently. Finding more personal opinions was expected because of background information about the expected feelings of participants towards the tool, an assumption yet not applying. Especially the hits on wrongfully defined problem-statements were useful to avoid tagging real problems based on wrongful description as false positives. On the other hand a problem originating from the test setup would have been hard to identify if not asked for.

An unexpected error occurred while trying to get more information on a problem identified as a real problem. It was meant to ask additional questions if a participant confirmed a statement of FIP as a real problem. Those questions had the aim to get information about the frequency of occurrence and their appropriate impact. Participants who were asked those questions reacted insecurely and gave indefinite answers, mentioning they did not think about this problem properly and thus could not give a suitable answer. As the output of these additional questions were of no further value, they were excluded from the test after a few tries. A suggestion would be to include an expert assessment afterwards, like done in the '*triage*' used in the research by Schmettow, Vos and Schraagen (2013). In this expert assessment the expert is provided time to think properly and give answers on the questions still open like frequency of occurrence and possible impact.

Another result was that removing or adding participants from a group checked by FIP for false positives, like done in the additional group, strongly effects the results of the late control strategy. This can be explained logically. This is because FIP is an iterative, self-developing protocol which influences the future participants' data for each newly added dataset, depending on their order. If the order and set is disrupted new singletons appear in the data because connected data is removed, influencing the late control strategy. This means data gathered using FIP are hard or even impossible to change without affecting the general results. That is not only true for the data excluding false positives, but also for the full data set since FIP improved the counts of confirmed real problems in both data sets adding additional hits to the results when a FIP-question is confirmed. However, a problem arises from FIP changing the frequency of how many participants identified a specific problem. In the current research, problems were categorized by their frequency of occurrence to be high priority or not. Problems confirmed by FIP are potentially added to the list of high priority problems due to this. For a prioritization based on frequencies of occurrence, problems only found by FIP should be excluded from this prioritization. Because this is not done in the current research the prioritization is of low reliability.

Gray and Salzman (1998) point out the importance of validity. Whether FIP is valid or not cannot be answered completely, but different findings give rise to the assumption of validity. Woolrych, Cockton and Hindmarch (2004) stated that observations of actual interactions are always a reliable source for identification of real problems. This in turn means that if it is able to identify real problems, it has to be able to identify unreal problems – false positives – the same way. In the current research participants identified false positives either by mentioning not to think about it as problematic in real use or by showing that the actual application works in a different way than the error-statement described by actual interaction with the system. Based on the statement by Woolrych, Cockton and Hindmarch (2004) this supports the assumption FIP was able to identify false positives as intended. Furthermore, Woolrych, Cockton and Hindmarch (2004) state that falsification testing allows confident coding of false positives. FIP is a method being able to verify real problems and falsify false positives in the data and thus has to be able to code data with confidence. Differences between groups or participants do not endanger the validity of the results because of multiple participants getting the same questions. A low performance of one participant can be compensated by a better performance of one of the at least two other participants getting the same FIP-question.

The following conclusions are drawn from this section: Despite the failure of one part of the protocol the main part turned out to have positive effects on the results of the late control strategy. All results improved when false positives identified by FIP were removed from the data. While lowering the total expected number of problems it also reduced the estimated required sample size to an identical value for both subgroups. The required sample size for Ges turned out to be 15, which is close to the value of 13 as an estimated required sample size for both of the subgroups. Furthermore, the confidence intervals were narrowed by removing the false positives from the data making it more reliable for the estimated required sample size to be correct. FIP was able to compensate for one problem the late control strategy can suffer from, like false positives, in an easy way to apply while gathering data. Furthermore, FIP was able to check for correct descriptions of problems, making it more reliant for the problems being described correctly. However, because of the failure to collect additional data a change in the method is suggested. The additional questions for frequencies and impact of the problems should be replaced by an expert assessment at the end of the data collection to obtain these data. This change might provide the necessary data and the expert has the chance to think over the question or to gather additional information. Compared to the '*triage*' used by Schmettow, Vos and Schraagen (2013) the advantage of FIP is mainly the possibility to execute FIP parallel to the data collection, making it possible to identify false positives in the data and to monitor the progress of data collection using the late control strategy without effects of false positives. To research the validity and the effectiveness of FIP more systematically the execution of a usability test using both FIP and the '*triage*' could be used and the results could be compared to each other. When using both methods, the results of FIP have to be stored separated from the raw data to not influence the data the '*triage*' later analyzes.

4.2 The late control strategy

The use of the late control strategy in this research had different aspects that have to be mentioned: While positively influenced by FIP, the different methods of data collection in G1 and G2 had a negative impact on the results of the late control strategy for the combined sample Ges. This is described in section 4.3 and compared to the results found by Schmettow, Bach and Scapin in 2014. For each group itself the late control strategy was able to predict an estimated total of problems, undiscovered problems and an estimated required sample size. Regarding the single groups, this corresponds to the results found by Schmettow, Vos and Schraagen (2013). The data for the combined group Ges was different in this case. Compared to the subgroups, the predicted total of problems increased for a value of more than 10 problems (15.39%) and the estimated required sample size increased by two (15.38%). Because of this the prediction of either group G1 or G2 as the first half of the total sample could not predict the total sample Ges. Still, even if the prediction of the required sample size increased, it was only by two, which is still within the 80% confidence intervals of G1 and G2. Furthermore, the estimated required sample sizes of G1 and G2 are also in the range of the 80% confidence interval of Ges. Based on the confidence intervals it has to be concluded that while suffering from changes in the data on a later point of time, the impact of those changes still ranges inside the confidence intervals, making the prediction still reliable. In addition to that, based on the illustration of the outputs, the completion-rate calculated by the late control strategy is correct, leaving not much space for expecting many new problems given the observation-rates.

The estimated sample sizes of G1 and G2 excluding false positives were both 13, bringing a magic number such as described by Nielsen (2000) back in mind. As the estimated sample size for Ges was 15, a magic number of around 13 is likely to be unreliable for this group. A use of the magic number would lead in the total sample to underestimation of the residing errors and a likely completion rate below 85%. The magic number of five required participants described by Nielsen (2000) is also refuted by the results of the current research. A total sample of five participants would not have found most of the problems that were found in the current data. The first five participants of G1 found 38 problems including false positives which is 42.22% of the problems found in the current data including false positives. By this only a number of problems far below 80% of the total problems would have been identified. This is based on the fact that the data show that after five participants still a great number of problems was found. The updated magic number of 10 +/- 2 introduced by Hwang and Salvendy in 2009 suffers from similar problems. While 13 is close to 12, which would be in the range of the magic number, it would not be effective to estimate the required sample size. The estimated sample size of 15 in group Ges is even more distant to 12, leading to the conclusion that a sample size predicted on the magic number of 12 would not be able to identify 85% of the existing problems. This refutes the theory of magic numbers being useful in practical use for cases like in the current study and is in line with the results found by Schmettow (2012) and Schmettow, Vos and Schraagen (2013) who came to a similar conclusion. However, while not tested in this study, it could be possible that a magic number has value in the initial stage of planning for a study. For example, a test based on a magic number could be planned and edited and monitored by the late control strategy later. This provides a number to work with during basic planning.

The early control strategy introduced by Lewis in 2001 states that based on a set of four participants it should be possible to estimate the final sample size. Based on the finding it was not possible with a number of 15 participants in G1 to estimate the full required sample size of Ges, which was 15, it is hard to support this strategy. An attempt to estimate the required sample size based on the first four participants of group G1 failed with an estimated sample size of 384 participants and the output that the integral is probably divergent. An estimation only based on the first four participants would have drastically overestimated the required sample size in the current research rendering the early control strategy very inaccurate. Surprisingly the estimated sample size is by far overrated in this case, which is in contrast to the results by Schmettow (2009), who found mostly underrated sample sizes. A reason for this could be the error of the integral to be probably divergent, leading to wrong results. Another result limiting the effectivity of the early control strategy is the problem that it can never be ruled out that a different observation pattern occurs after a few participants, like happened in the current research. An estimation based on a part of a sample having a different frequency to detect problems than the rest of the sample is likely to lead to an inaccurate prediction. A method to estimate the sample size has at least to be able to compensate for such changes to some degree like the late control strategy did in the current research by adapting the estimation to the newly found problems and the detection-pattern of the whole sample. This is because an underestimation based on too few observations would possibly let important and possibly hazardous problems undiscovered while it could be prevented. By this it has to be concluded that the early control strategy cannot estimate the required sample size as precisely as the late control strategy which is generally in line with the results by Schmettow (2009) and Schmettow, Vos and Schraagen (2013). However, the early control strategy could be useful on tests including an expected smaller sample size where the late control strategy can not be used due to its small sample size. Another possible use case might be simple, routinized systems where no differences in usage patterns are expected. Yet this may lead to problems, as the try to estimate a sample size based on four participants failed in the current study, leaving the question open whether the early control strategy would work in such cases.

In conclusion, the current research was able to evaluate the late control strategy as being able to monitor the process but suffering to a given degree from changes in the observed patterns or different samples. Furthermore, the late control strategy outperforms other strategies like the magic number or the early control strategy in tests involving both complex systems and a high level of user diversity because of its ability to adapt to changes in the results found.

4.3 Different results for different methods

Sullivan (2013) implied that testing via a communication tool could be an alternative approach to test usability more efficiently. Therefore in this research one group G1 was tested via the traditional face to face approach while the other one was tested via a new communication-tool-based approach. In the current study the groups G1 and G2 found a similar number of problems, even though they differed in the kind of problems they found. This may have two reasons: first the samples were different, second the approaches had an impact on the results. A difference between the groups cannot be ruled out completely, but the participants had to do the same tasks and the same jobs like in their everyday work. Consequently, they needed to follow the same steps as the participants in the other group. It is quite likely they would stumble upon the same problems the other group did. Differences resulting from the setup of the subgroups being built up from different locations are not likely because based on descriptive statistics no significant differences between the general discovery frequencies were found. Furthermore, both groups had similar numbers of false positives and problems found. However, the differences between the number of detected problems were larger in group G2 than in G1. Furthermore, testing different users at different locations in both groups compensated for social influences from participants to other attendees. As the identities of the participants were unknown they were unable to talk to each other in order to retrieve a specific, desired result. This made the results more individual and in general more neutral, giving an unaffected view on the tested system. However, the groups G1 and G2 differ in some demographical aspects like the level of education, age and distribution of gender. These differences possibly had an influence on the kind of problems they found. For instance, participants with a higher level of education might be expected to work in a different manner or do different tasks in general than participants with a lower level of education and thus struggle for other problems. The same would be possible for differences in age because it has to be expected that older participants are more experienced in life, having a possible influence on their style of thinking and problem solving. However, it cannot be determined whether this had an influence on the results or not. Maybe the second questionnaire, which was not analyzed in this study, could have given insight into this. As it is not possible to determine the influence of the samples, another likely result is for the approach to have some effect, potentially in the interaction with differences in person's characteristics. In general the interviews taken in group G2 were experienced more distant and it was harder to build up a personal relationship to the participant, resulting in shorter answers. In total the interviews in G2 were on average approximately seven minutes shorter than in group G1. This might be related to the problems mentioned by the participants in different ways. For example, they might have mentioned only problems they were able to explain in short statements instead of reporting problems being hard to explain. If the results would be related to the approach on how the tests were taken, it could be related to a study performed by Schmettow, Bach & Scapin (2014). They compared different methods for problem identification (for additional information see Schmettow, Bach & Scapin, 2014), identifying differences as a result. One of the methods performed generally poor, but the other two methods showed similar characteristics as found in the current study: The methods found equal numbers of problems, but different problems. Schmettow, Bach and Scapin (2014) stated that the two methods "do different things equally well" (Schmettow, Bach & Scapin, 2014, p. 7). This might be the same in the current study, but for confirmation more information is necessary. A study with a more controlled sample could help comparing both approaches to communicate with the user in a more differentiated way than was possible in the current research. Furthermore, Schmettow, Bach and Scapin (2014) stated both methods counterbalance each other's weakness and, in effect, they may find more problems with less effort when combined. If the general characteristic of two methods or approaches doing "different things equally well" (Schmettow, Bach & Scapin, 2014, p. 7) would hold true in the current study as well it could be possible that the total result of problems found can gain benefit from the combination of both methods. The fact that in the current research the estimated sample size for group Ges excluding false positives is still 15 persons leads to the assumption that a combination of both groups would find more problems with nearly the same or less effort. This result was unexpected, because the method of data gathering was the same for both groups, only the approach to contact the participants was different. A similarity to results found by generally different methods could not be expected. An additional analysis of the data gathered only including the first half of the participants for each location was done to investigate the possibility of a combined approach leading to the same results. The results of this additional group were more similar to the results from the other groups including false positives. An explanation for this might be that the data for the groups G1 and G2 has been cleared from problems occurring only once while the additional group was not, leading to single events in this group. To answer the question whether a combination of both approaches would be of benefit for the results additional research is needed, keeping this question in mind during the planning-phase. Furthermore, a similarity to the research by Schmettow, Bach and Scapin (2014) could be possible, yet not verified and needing further research.

In conclusion, the question of a communication-tool-based approach being as effective as a face to face based approach may not be answered completely yet. Both approaches performed equally in numbers, yet it is hard to answer whether they perform equally in quality due to the different problems they identified. This should be investigated in more detail. Another recommendation out of this is to investigate the influence of combined and separated approaches on the gathered data and the late control strategy. In this research an attempt on creating an additional group combining both methods was made, most likely failing because of influences of the usage of FIP. Therefore the hypothesis that a combined approach would lead to more efficient testing could not be tested.

4.4 Limitations

The current research has two limitations that should be mentioned. The first limitation is the single person who performed coding, mapping and analysis. The interview-leader was the person who coded, mapped and analyzed all data during the whole phase of data collection. This is a source of bias like Gray and Salzman (1998) described. The interview-leader could have changed unconsciously the way he coded the data. Furthermore, the personal line of thinking the interview-leader got during the interviews could have influenced the way he mapped problems to each other and interpreting them. For example, problems only based on the transcripts did not fit fully could be mapped despite the differences because of meta-effects between different samples. This problem was difficult to mitigate because of differences in time of data-collection and the need for anonymization of the data. Resulting from this anonymization some parts of the transcripts had to be reformulated in order of not revealing business secrets. Mistakes in coding have an effect on the final data (Woolrych, Cockton & Hindmarch, 2004). Each mistake in coding or mapping would mean a change in the number and priority of problems found. However, a small number of mistakes would not change the results drastically because of the setup of the late control strategy, which can compensate for it to a given degree by its mathematical setup based on the geometrical function. Furthermore, it was tried to code and map data on a high level of detail, not leaving much space for interpretation.

The second limitation is the unused post-questionnaire. Gray and Salzman (1998) write that the internal validity can be endangered if potential uncontrolled variables could have influenced the data. The unused post-questionnaire was meant to give insight in the effects of the think aloud protocols regarding the concentration and feelings of the user. Because of problems in the interpretation of the participants the questionnaire was filled in biased and would have provided wrong information. To fully rule out interactions regarding the test-situation it would be necessary to analyze the data that should be gathered by this questionnaire. This could not be done, so the results could be influenced by different factors resulting out of the test-situation. This would not render the results useless but would possibly give rise to the assumption that some of the problems found will only occur in specific situations and therefore have a different impact and frequency.

4.5 Conclusion

In the current study the two different methods of data collection – face to face and via a communication tool - proved to be equally effective to gather different data. More research is needed to find out which method is in practical setting of more use or if a combination would work best.

In general the late control strategy was verified as an effective method to estimate the required sample size while testing usability and monitor the completeness reached. The assumption that this method is greatly affected by observations only appearing once has been proven to be correct. Furthermore, it was somewhat affected by changes in the data-pattern. But despite those limitations it proved to be superior to other known strategies like the magic number or the early control strategy when testing complex systems with a high level of user diversity. For the elimination of false positives FIP was used. The newly developed protocol FIP had a positive effect on the results of the late control strategy, improving all results. A supposition for an additional step in FIP was made: An expert assessment to gather additional data on the problems found. Furthermore, FIP influences the data in a way making it hard to remove participants later from the sample. Those characteristics should be kept in mind when using FIP.

For practical and scientific use some recommendation can be formulated based on this thesis:

1. When testing usability and needing to estimate the required sample size or to monitor the process a combination of the late control strategy and an improved version of FIP is expected to perform well and enhance the efficiency and precision of the tests. It should be kept in mind that large changes in the observation-pattern influence the results of the late control strategy, so the confidence-intervals should also be checked not only relying on the output required sample size. Furthermore, the data should be processed as it was gathered when using FIP because it is likely to get undesired effects when changing the sample setup. Third, the effect of FIP on the frequencies of occurrence have to be kept in mind when analyzing the data.

2. Other approaches like the magic number and the early control strategy were outperformed and rendered useless in comparison with the late control strategy in cases involving complex systems and a high level of user diversity. Those two methods should not be used in such cases because they are not always able to predict the required sample size as accurately as required and, consequently, the results will not have the desired precision. Potential harmful problems could be not detected leading to several problems.

3. Testing usability face to face seems as efficient as testing usability via a communication tool. However, the results of the two methods are different. More research is needed to find a good solution on how to deal with these differences. When only testing for high priority problems it is likely that both methods perform equally well.

References:

Bevan, N. (2001). International standards for HCI and usability. *International Journal of Human-Computer Studies, 55*(4), 533-552. Doi: 10.1006/ijhc.2001.0483

Carrol, J. M., & Rosson, M. B. (1987). Paradox of the active user. In Carroll J.M. (Ed.), *Interfacing thought: cognitive aspects of human-computer interaction* (pp. 80-111). Cambridge, MA: MIT Press.

Cassel, C., & Symon, G. (2004). *Essential guide to qualitative methods in organizational research.* Thousand Oaks, CA: Sage Publications, Inc.

Caulton, D. A. (2001). Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology, 20*(1), 1-7.

Dain, S., (2002). Normal accidents: Human error and medical equipment design. *The Health Surgery Forum, 5,* 254-257.

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data. *Cambridge, MA, US: The MIT Press,* 443.

Finstad, K. (2008). Analogical problem solving in casual and experienced users: When interface consistency leads to inappropriate transfair. *Human-Computer Interaction, 23*(4), 381-405.

Gould, J. D., Boies, S. J., & Lewis, C. (1991, January). Making usable, useful, productivity-enhancing computer applications. *Communications of the ACM, 34*(1), 74-85. Doi:

10.1145/99977.99993

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*(3), 203-261. Doi: 10.1207/s15327051hci1303_2

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated Retrospective Think-Aloud Method as measured by eye tracking. *CHI 2006.*

Haak, M. J., & Jong, M. D. T. (2003). Exploring two methods of usability testing: Concurrent versus retrospective think-aloud protocols. *IEEE International Professional Communication Conference Proeedings.*

Haak, M. J., Jong, M. D. T., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology, 22*, 339-251.

Harty, J. (2011, February). Finding usability bugs with automated tests. *Communications of the ACM, 54*(2), 44-49. Doi: 10.1145/1897816.1897836

Hornbæk, K., & Frøkjær, E. (2008). Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers, 20*(6), 505–514.

Hwang, W., & Salvendy, G. (2009). Integration of usability evaluation studies via a novel meta-analytic approach: What are significant attributes for effective evaluation? *International Journal of Human-Computer Interaction, 25*(4), 282-306.

Lewis, J. R. (2001). Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International journal of human-computer interaction, 134*, 445-479.

Lin, L., Isla, R., Doniz, K., Harkness, H., Vicente, K., & Doyle, J. (1998) Applying human factors to the design of medical equipment: patient-controlled analgesia. *Journal of Clinical monitoring and Computing, 14,* 271-283.

Nielsen, J. (1993). Usability engineering. *Academic Press, New York.*

Nielsen, J. (2000). Why you only need to test with five users. Retrieved from http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '93* (206-213). Retrieved from http://portal.acm.org/citation.cfm?doid=169059.169166

Obradovich, J.H., & Woods, D. D. (1996). Users as designers: how people cope with poor HCI design in computer-based medical devices. *Human Factors, 38,* 574-592.

Pape, U. (2012). *Grundlagen der Finanzierung und Investition: Mit Fallbeispielen und Übungen* (3. Aufl.). Berlin: Oldenbourg Wissenschaftsverlag.

Reason, J. (1990). Human error. Cambridge University Press. In Liljegren, E., Osvalder, A., Dahlman, S. (2002). Setting the requirements for a user-friendly infusion pump. *In: Proceedings of the IEA 2000/HFES 2000 congress, 132.*

Rizwan, M., & Iqbal, M. (2011). Application of the 80/20 rule in software engineering rapid application development (rad) model. In J. M. Zain, W. M. b. W. Mohd, & E. El-Qawasmeh (Eds), *Software Engineering and Computer Systems: Second international conference, ICSECS* (pp. 518-532). Berlin, DE: Springer-Verlag Berlin Heidelberg

Schmettow, M. (2009, September 1-5). Controlling the usability evaluation process under varying defect visibility. In: BSC-HCI '09. *Proceedings of the 23rd British HCI group annual conference on people and computers: celebrating people and technology (UK),* Cambridge, (pp. 188-197). Swinton, UK: British Computer Society.

Schmettow, M. (2012, April). Sample size in usability studies. *Communications of the ACM, 55*(4), 64-70.

Schmettow, M., Bach, C., & Scapin, D. (2014, September 9-12). Optimizing usability studies by complementary Evaluation Methods. In: HCI2014. *28th British HCI Conference,* Southport. Southport, UK: BCS Learning and Development Ltd.

Schmettow, M., Vos, W., & Schraagen, J. M. (2013). With how many users should you test a medical infusion pump? Sampling strategies for usability tests on high-risk systems. *Journal of biomedical informatics*, *46*(4), 626-41. Retrieved from http://www.sciencedirect.com/science/article/pii/S1532046413000506

Sears, A. (1997). Heuristic walkthroughs: finding the problems without noise. *International Journal of human-computer-interaction*, 213-234.

Shackel, B. (2009). Usability – context, framework, definition, design and evaluation. *Interacting with Computers, 21*(5-6), 339-346. doi: 10.1016/j.intcom.2009.04.007

Sullivan, J. R. (2013). Skype: An appropriate method of data collection for qualitative interviews? *The Hilltop Review, 6*(1), 54-60. Retrieved from http://scholarworks.wmich.edu/cgi/viewcontent.cgiarticle=1074&context= hilltopreview.

Vieritz, H., Yazdi, F., Schilberg, D., Göhner, P., & Jeschke, S. (2011, November 25). *User-centered Design of Accessible Web and Automation Systems.* In A. Holzinger & K. Simonic (Eds.), *USAB'11 Proceedings of the 7th conference on Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society: information Quality in e-Health.* Paper presented at the 7th conference on Workgroup Human-Computer Interaction and Usability Engineering, Austria, (367-378). Heidelberg: Springer-Verlag Berlin. Doi: 10.1007/978-3-642-25364-5_26

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

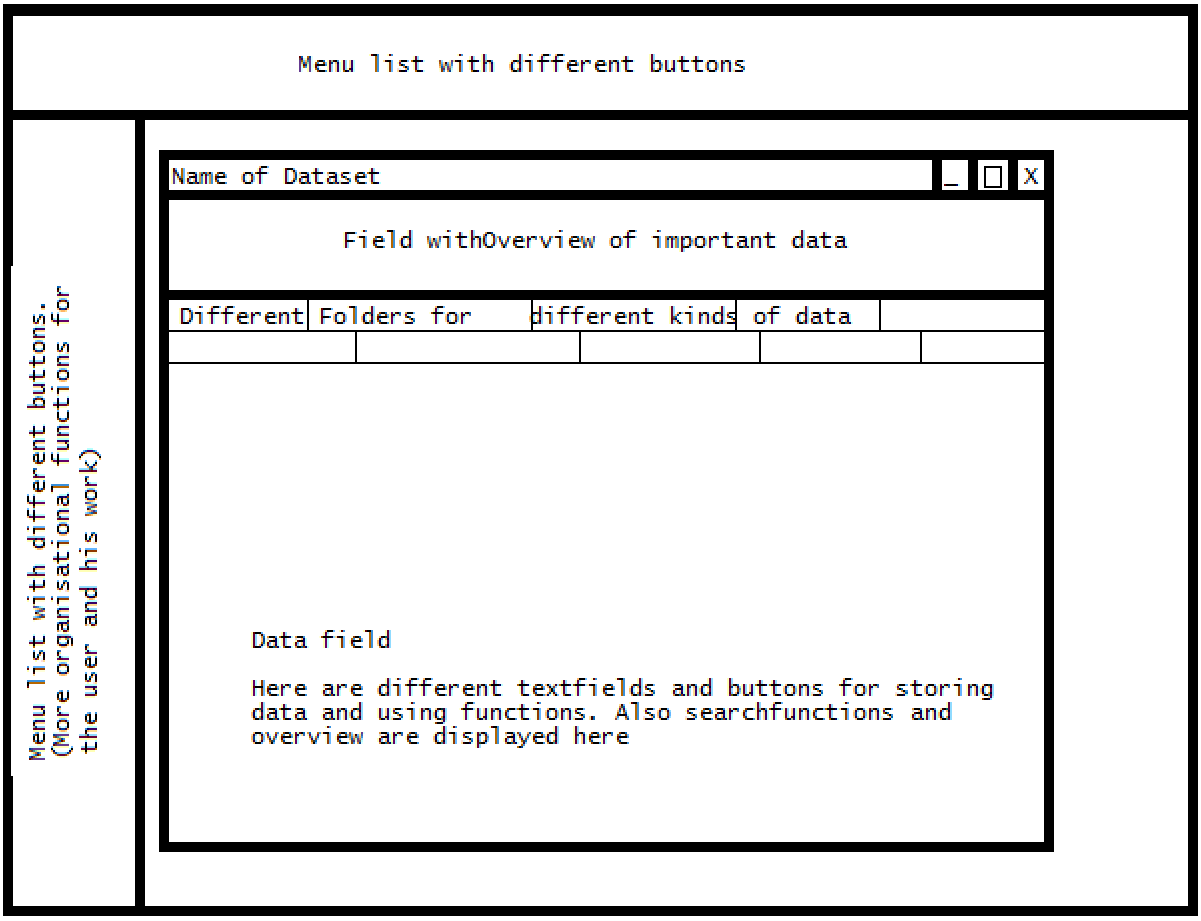
Vos, W. M. (2011). Quantitative and efficient usability testing in high risk system development: Under diversity of user groups (Unpublished Master thesis). University of Twente, Enschede, The Netherlands.

Woolrych, A., Cockton, G., & Hindmarch, M. (2004). Falsification testing for usability inspection method assessment. *In proceedings of the HCI04 Conference on People and Computers XVIII.*

Appendix

Appendix A

Illustration of the user interface. (Only an illustration for privacy and copyright reasons)


Menu list with different buttons
Menu list with different buttons.
(More organisational functions for
the user and his work)
Name of Dataset
_
□
X
Field withOverview of important data
Different
Folders for
different kinds
of data
Data field

Here are different textfields and buttons for storing data and using functions. Also searchfunctions and overview are displayed here

Illustration A. The user interface of the application used in the test including descriptions of the fields and menus.

Appendix B

Questionnaire for Demographics

Fragen zum Teilnehmer

Allgemeine Daten:	
Datum:	Teilnehmernummer:
Geschlecht: Männlich __ Weiblich __	
Alter:	
Höchste Berufsausbildung:	
Bitte beantworten Sie die folgenden Fragen in Jahren. Falls Sie weniger als ein Jahr antworten müssten, schreiben Sie bitte Monate dazu. Kreuzen Sie bei jeder Frage bitte jeweils nur eine der Antwortmöglichkeiten an.	
Wie lange arbeiten Sie schon für NAME FIRMA?	
Wie lange arbeiten Sie schon in Ihrer derzeitigen Position/Aufgabe?	
Wie lange arbeiten Sie schon mit der Anwendung?	
Wie oft verwenden Sie die Anwendung im Durchschnitt? Täglich __ 4 Tage pro Woche __ 3 Tage pro Woche __ 2 Tage pro Woche __ 1 Tag pro Woche __ Seltener als ein Tag pro Woche __	
Die Anwendung kann auf verschiedene Arten bedient werden. Welche Art der Bedienung nutzen Sie? Ich steuere nur mit der Maus __ Ich steuere nur über Tastaturbefehle __ Ich steuere mit der Maus und mit der Tastaturbefehlen __ Ich steuere auf andere Art, nämlich:	
Denken Sie über sich selbst, dass Sie ein erfahrener Nutzer der Anwendung sind? Ja __ Nein __ Ich bin mir nicht sicher __	

Appendix C

Postquestionnaire

Fragen über die Erfahrung mit dem Think aloud protocol **Teilnehmer:**

Es geht bei diesen Fragen nur um Ihre persönliche Erfahrung. Es gibt keine richtigen oder falschen Antworten.
Markieren Sie bitte den Wert, der auf der Liste am besten zu Ihrer persönlichen Erfahrung passt.
*(Schriftlich markieren Sie bitte mit einem Kreis um die Zahl, digital markieren Sie bitte die Zahl indem Sie sie **fett** drucken oder in der ersten Spalte eintragen)*

	Wie empfanden Sie die Aufgabe, während Ihrer normalen Aufgabe laut zu denken?
	Einfach 1 –– 2 –– 3 –– 4 –– 5 Schwierig
	Angenehm 1 –– 2 –– 3 –– 4 –– 5 Unangenehm
	Nicht anstrengend 1 –– 2 –– 3 –– 4 –– 5 Anstrengend
	Natürlich 1 –– 2 –– 3 –– 4 –– 5 Unnatürlich
	Nicht Zeitraubend 1 –– 2 –– 3 –– 4 –– 5 Zeitraubend

	Wie haben Sie das Arbeiten mit dem Think Aloud Protokol im Vergleich zu Ihrem normalen Arbeiten ohne Think Aloud Protocol empfunden?
	Schneller 1 –– 2 –– 3 –– 4 –– 5 Langsamer
	Weniger verwirrend 1 –– 2 –– 3 –– 4 –– 5 Verwirrender
	Ich war besser konzentriert 1 –– 2 –– 3 –– 4 –– 5 Ich war schlechter konzentriert
	Nicht mehr anstrengend 1 –– 2 –– 3 –– 4 –– 5 Anstrengender
	Erfolgreicher 1 –– 2 –– 3 –– 4 –– 5 Weniger erfolgreich
	Es machte mehr Spaß 1 –– 2 –– 3 –– 4 –– 5 Es machte weniger Spaß
	Mir sind Fehler besser aufgefallen 1 –– 2 –– 3 –– 4 –– 5 Mir sind Fehler schlechter aufgefallen
	Entspannter 1 –– 2 –– 3 –– 4 –– 5 Stressiger

Appendix D

Interview-questions (German)

Interviewleitfaden

Es folgen noch einige Fragen, die ich Ihnen gern stellen möchte.

1) Was denken Sie über die grafische Darstellung der Anwendung?
(Ist sie angenehm, übersichtlich, attraktiv, wirkt sie qualitativ? Macht es Spaß mit ihr zu arbeiten?Was kann besser?)

2) Was denken Sie über die Zugänglichkeit der Anwendung? Wie einfach es, ist sich in der Anwendung zurechtzufinden oder neue oder unbekannte Funktionen zu erlernen?
(Warum?)

3) Was denken Sie über die Strukturierung der grafischen Oberfläche?
(Ist sie intuitiv? Sinnvoll angeordnet? Themen sind sinnvoll strukturiert?)

4) Was denken Sie über die Beschriftungen der grafischen Oberfläche? Also Menütitel, Feldtitel und Überschriften?
(Sind sie gut gewählt? Irreführend? Gibt es Verwechslungsgefahr?)

5) Was denken Sie über die in der Anwendung angezeigten Fehler- und Warnhinweise?
(Sind sie ausreichend? Verständlich? Gut sichtbar?)

6) Kennen Sie 'Stolperfallen' in der Anwendung? / Muss man auf etwas besonders achten?
(Wenn ja: Wie schwer sind die Folgen? Wie oft sind Sie darüber gestolpert, bis Sie sie nicht mehr bemerkten?)

7) Machen Sie manchmal fehlerhafte Angaben oder bedienen Sie die Anwendung falsch?
(Wenn ja: gibt es eine bestimmte Stelle / Aufgabe, an der dies oft geschieht? z.B.: Falsches Feld genutzt)

8) Können Sie alle nötigen Aufgaben mit der Anwendung erfüllen?
(Wenn nein, warum nicht? Was nicht?)

9)FIP

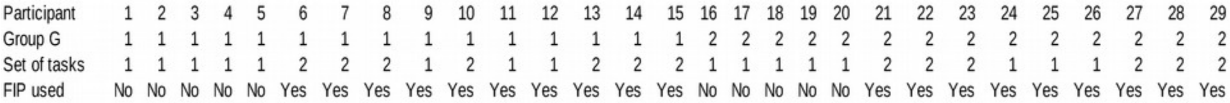
(Hilfestellungen:
Stellen Sie sich vor, dass: *Umschreibung des Ereignisses*
- Denken Sie, dass diese Beschreibung eines Vorfalles eher eine persönliche Meinung ist, oder denken Sie, dass sie auf die Anwendung zutrifft?

- Wie oft tritt so ein Fall ein?
- Wie schwerwiegend sind die Konsequenzen?
- Wie lange dauert es, bis man nicht mehr darauf hereinfällt / eine Lösung für das Problem

hat?)

Appendix E

List of participants and tasks



Participant	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Group G	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Set of tasks	1	1	1	1	1	2	2	2	1	2	1	1	2	2	2	1	1	1	1	1	2	2	2	1	1	1	2	2	2
FIP used	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table E. List of participants, their group, set of task and whether FIP was used or not.

Appendix F

Questions used in FIP (German)

Die nachfolgenden Fragen sind eher hypothetisch zu verstehen. Will sagen, nur weil ich etwas sage, heißt das noch nicht, dass es auf die Anwendung auch zutrifft. Ich würde Sie bitten, mir zu sagen, ob Sie der Ansicht sind, dass die Aussage zutrifft oder nicht.

1 Stimmt es, dass das System Veränderungen im Datensatz erst nach einem erneuten Öffnen des Datensatzes in der Ausgabe anzeigt?

2 Würden Sie sagen, dass dem System Unschlüssigkeiten zu spät oder gar nicht auffallen?

3 Würden Sie sagen, dass die Anordnung einiger Dropdownmenüs unlogisch ist?

4 Haben Sie schon mal versucht Daten in ein Feld einzutragen, das auf einem anderen Reiter lag als dem, auf dem Sie gerade sind?

5 Stimmt die Aussage: Es fehlen Hinweise auf weitere mögliche Daten in einigen Untermenüs. Diese werden übersehen.

6 Würden Sie sagen, dass Hinweise wenn man etwas eingeben muss fehlen.

7 Würden Sie sagen, dass die Funktion zum öffnen von externen Suchseiten unbrauchbar ist weil sie zu schnell schließt und Fehler macht?.

8 Finden Sie manchmal das Symbol der Anwendung nicht in der Taskleiste?

9 Würden Sie sagen, dass es eine unklare Beschriftung bei Erweitertem Feld gibt?

10 Würden Sie sagen, dass unnötige Textbausteine im Datensatz verbleiben?

11 Würden Sie sagen, dass die Übersicht der Daten besser werden müsste?

12 würden Sie sagen, dass Sie zu viele Möglichkeiten in der Anwendung haben?

13 Würden Sie sagen, dass es hilfreich sein würde wenn Sie die Oberfläche anpassen könnten? Also einzelne Felder verschieben oder ausblenden?

14 Und würde Ihnen das helfen wenn Sie einstellen könnten, welche Fenster automatisch morgens geöffnet werden?

15 Würden Sie sagen, dass die gesamte Darstellung der Oberfläche zu klein ist?

16 Würden Sie sagen, dass in den verschiedenen Untermenüs mit Adressen ein Adressfinder sinnvoll wäre?

17 Nehmen Sie die angezeigten Hinweise bei Pflichtfeldern wahr?

18 Würden Sie sagen, dass auf fehlende Haken bei Einträgen nicht hingewiesen wird?

19 Würden Sie sagen, dass es Ihnen schon einmal passiert, dass Sie, wenn Sie einen Eintrag neu anlegen möchten, aus Versehen auf ein Suchergebnis klicken und dadurch das Suchergebnis unbemerkt übernehmen?

20 Würden Sie sagen, dass die Anordnung in Drop-Down Menüs manchmal unlogisch ist?

21 Würden Sie sagen, dass die Mauswege von einem Fenster in der rechten unteren Ecke bis zum Suchen oder Neu Button lang sind?

22 Würden Sie sagen, dass es sinnvoll wäre, wenn man in der Ausgabe markieren könnte?

23 Benutzen Sie das integrierte Suchsystem (Strg+I)? (Resultat: fehlerhafte Umschreibung)

24 Würden Sie sagen, dass ein Hinweis auf weitere mögliche Daten in den Sub-menüs hilfreich wäre?

25 Würden Sie sagen, dass das Logo in der Startleiste schlecht zu finden ist?

26 Würden Sie sagen, dass die A-Übersicht ausreichend ist? Oder fehlt dort etwas?

27 Würden Sie sagen, dass Ihnen bei der Arbeit eine angezeigte Checkliste für sinnvolle Fragen helfen würde? Also wenn sie zum Beispiel angezeigt würde, wenn Sie mit der Maus über einem Reiter verharren?

28 Würden Sie sagen, dass die Oberfläche zur Verteilung der Aufgaben übersichtlich und leicht verständlich ist?

29 Würden Sie sagen, dass der Cursor im Suchfeld ungünstig positioniert ist und Sie ihn lieber auf dem Dropdownmenü hätten?

30 Suchen Sie manchmal Felder?

31 Wie würden Sie es finden, wenn das System bereits zu den passenden Ergebnissen springt, während Sie tippen?

32 Würden Sie sagen, dass es stört, dass die Ausgabe anders strukturiert ist als in dem Datensatz?

33 Würden Tags, also Kürzel bei den Aufträgen helfen um schneller zu erkennen, um was es bei dem Auftrag geht?

34 Würden Sie sagen, dass ein kleines, integriertes Suchfeld in der Anwendung sinnvoll wäre?

35 Würden Sie sagen, dass eine Auswahl bei Nachträgen für 'unbenannt' existieren sollte?

36 Würden Sie sagen, dass die Oberfläche zur Verteilung der Aufgaben übersichtlich und leicht verständlich ist?

37 Finden Sie, dass die Arbeitsaufträge unübersichtlich und unlogisch in der Abarbeitung sind?

38 Würden Sie sagen, dass es ein Problem ist, dass man zum Beispiel Adresseinträge erst vollständig ausfüllen muss um einen anderen Eintrag editieren zu können?

39 Würden Sie sagen, dass es ein Problem ist, dass Textbausteine in der Ausgabe nicht zu einem zusammenhängenden Text zusammen gefügt werden?

40 Würden Sie sagen, dass es besser wäre, wenn bei einer spezifischen Auswahl, zum Beispiel Gewerbebetrieb, auch automatisch die anderen Felder auf die passenden Einträge springen würden? Also zum Beispiel Herkunft: Gewerbeamt?

41 Würden Sie sagen, dass ein IBAN Feld fehlt?

42 Würden Sie sagen, dass es ein Problem ist, dass das Namensfeld nur 20 Zeichen fassen kann?

43 Würden Sie sagen, dass ein Hinweis fehlt, dass das Änderungsdatum geändert werden müsste?

44 Würden Sie sagen, dass es ein Problem ist, dass zum Beispiel bei einem Adresswechsel die Gültigkeit bereits einen Tag vor dem Beginn der neuen Adresse abläuft?

45 Würden Sie sagen, dass es Ihnen helfen würde, wenn Sie als zu versenden markierte Post noch eine gewisse Zeit aufhalten könnten?

46 Würde Ihnen eine Auswahlmöglichkeit helfen, die automatisch alle von Ihnen veränderten Felder als Herkunft auf das von Ihnen zuvor gewählte setzt?

47 Würden Sie sagen, dass es sinnvoll wäre, wenn bei versendeten Konzepten ein Hinweis stände, wann das letzte Konzept verschickt wurde?

48 Haben Sie schon einmal beobachtet, dass die Anwendung Textbausteine an einer falschen Stelle in die Ausgabe gesetzt hat?

49 Würde es Ihnen helfen, wenn Sie einen Hinweis bekommen würden, wenn Sie vergessen haben, das Änderungsdatum zu ändern?

50 Würden Sie sagen, dass eine Funktion fehlt, die einfach nur eine neue Immobilie anlegt?

Also ein Button dafür?

Appendix G

Selected list of problems and transcripts

Problem: **Font is chosen problematically**
Overall the text is too small and the font size is not always good to identify. Several users have problems to read the text they need for their tasks, making their job exhausting.

Transcript Only the font is too small for me. I would like to have a bigger font. You can work with enlargement but with this small font … you tend to have an error-frequency that you cannot perceive the numbers correctly or something like that. (*Nur die Schrift ist für mich zu klein. Ich hätte gerne ne größere Schrift. Man kann zwar mit Vergrößerung arbeiten aber es ist dann mit dieser kleinen Schrift ist .... man neigt zu einer Fehlerhäufigkeit dass man die Zahlen nicht richtig erkennt oder sowas. -Participant: G1O2*)

Sometimes I think the font could be a little bigger. The font is a little small. (*Manchmal denke ich die Schrift könnte ein bisschen größer sein. Die Schrift ist ein bisschen klein.-Participant: G1O1*)

Problem: **Structure: outer menus bad perceived**
Buttons and functions located in the upper menu in the main-window are frequently not perceived as belonging to the sub-window. Users search for options in the sub-window and cannot find them because they are located at the upper menu. The choice where the buttons are located seems to be frequently random.

Transcript <u>Interviewer:</u> Are you sometimes searching for fields?
<u>Participant:</u> Yes sure! For example now what we had with the truck. There are the one or other things you have to search for. And that was also not self explaining at that moment. That you find those positions and those fields there under 'Branch' and 'Extended'. (Note: The button 'Extended' is located in the outer menu only visible in the tab 'Branch') Somebody who does not know this stuff would have his problems to find this. That is, like I said, not self explaining.
(*Interviewer: Suchen Sie manchmal Felder?*
Participant: Ja sicherlich. Jetzt Beispielsweise was wir mit den LKW hatten. Da sind doch so die ein oder anderen Sachen, die man suchen muss. Und das ist dann in dem Moment auch nicht selbsterklärend gewesen. Dass man dann unter der Branche und Erweitert diese Position und diese Felder da findet. Also da hätte Jemand der sich nicht auskennt seine Probleme das zu finden. Das ist wie gesagt, nicht selbst erklärend. -Participant: G1HM7)

And then there are such one, two fields again….where I have to think about… (Participant searches nearly 5 minutes for the button mentioned, finding it at 'Branch' in the outer menu as 'Extended' in the end.) Precise here.
(*Und dann gibt es da noch so ein zwei Felder.... wo ich auch dran denken muss.... (Teilnehmerin sucht fast 5 Minuten den betreffenden Button, findet ihn am Ende unter Branche im äusseren Menü als Erweitert.) Genau hier.*
-Participant: G2BK2)

Problem: **Structure: Order in dropdown-menus illogically**
The structure of the options in some dropdown-menus is illogical. For example, the male version of a description is located at the top ten of the menu while the female version is almost at the end of the long list. This leads to errors when not known and irritates users.

Transcript And than I insert the owner (female) here, … that is always written further down here, I do not know why either. I would find it more reasonable when, even if the most owners are male, the owner (female) would be put directly beneath the owner and not far way down. Because it is nearly no difference whether it is a male owner or a female owner. That would be a suggestion for improvement for me. (Note: In German there are two different words for male and female owners. - Inhaber (male) and Inhaberin (female))
(*Und dann gebe ich hier die Inhaberin,... die hier immer weiter unten steht, warum weiß ich auch nicht. Ich würde es jetzt zum Beispiel sinnvoller finden wenn man, auch wenn die meisten Inhaber Männer sind, die Inhaberin gleich unter den Inhaber machen würde und nicht ganz weit unten. Weil es ja kaum ein Unterschied ist ob ein Inhaber oder eine Inhaberin. Das wäre für mich ein Verbesserungsvorschlag.-Participant G1O1*)

Well it is a little bit illogical yet because for example we have the owner (female) written at the end and the owner (male) at the top… you could put them below each other, but that does not bother me now because at some point I know where what is located. Because you work with it every day.
(Also es ist schon ein bisschen unlogisch, weil die Inhaberin zum Beispiel ganz unten steht und der Inhaber ganz oben.... das könnte man untereinander setzen, aber es stört mich jetzt auch nicht, weil ich irgendwann auch weiß wo was steht. Dadurch, dass man da jeden Tag mit arbeitet. -Participant G1O6)

Problem: **Structure: Left menu not as menu recognized**
This false positive resulted from the test-setup. Users had trouble to find the correct functions on the left sided menu. This was because they had optimized this menu to their needs in the actual application several years ago. In the test-environment it was the default menu they were not used to. Users stated this problem will never occur in the actual application

Transcript including observation and false positive confirmation

The participant searches for a field in the menu located on the left side.
Interviewer: I have just seen you were searching for a field on the sidebar for a while. Does this happen to you frequently, that you have to search a menu item there?
Participant: No. Because I have customized this, when I log in into the application, the user interface directly appears. And then there are the different sub-fields sorted as I have sorted them. That is only here in the test-environment right now.
(*Interviewer: ich habe jetzt gerade gesehen, dass Sie da schon auf der Seitenleiste ein bisschen am Suchen waren wo das Feld ist. Passiert Ihnen das öfter, dass Sie einen Menüpunkt da suchen müssen?*
Participant: Nein. Weil ich das für mich eingestellt habe wenn ich mich in der Anwendung anmelde dann erscheint da ja gleich die Benutzeroberfläche. Und da sind dann ja wiederum die ganzen anderen Unterfelder sortiert, so wie ich mir das sortiert habe. Das ist jetzt nur hier in der Testumgebung. -Participant G1HM2)

Problem:

The system only shows changes in the data after reopening the datafile in the output
One user mentioned the output would show new data only when the data file is closed and then opened again. This was confirmed false positive by other users because it is only necessary to save the data-file since the output is created from the database and not the displayed window. When closing a data-file the user is always asked whether he wants to save. Resulting, the user who mentioned the problem got the misconception he would need to close the file.

Transcript first observation:

There we have something special on the system again. It does not display changes here until I pop out and in again. So do not wonder about this, it is changed there. (referencing to the data-file)
(*Da haben wir wieder etwas Besonderes an dem System. Mir zeigt er hier erst dann Änderungen an, wenn ich raus und wieder rein gehe. Also nicht wundern, der ist dort geändert. -Participant G1O2*)

Disconfirmation:

No, that is not correct. It is sufficient if you save once and than open the output.
(*Nein, das ist ja nicht richtig. Es reicht, wenn man einmal speichert und dann die Ausgabe aufruft. -Participant G1O6*)

Appendix H

Distribution graphics

Appendix H-1: Group G1 including false positives

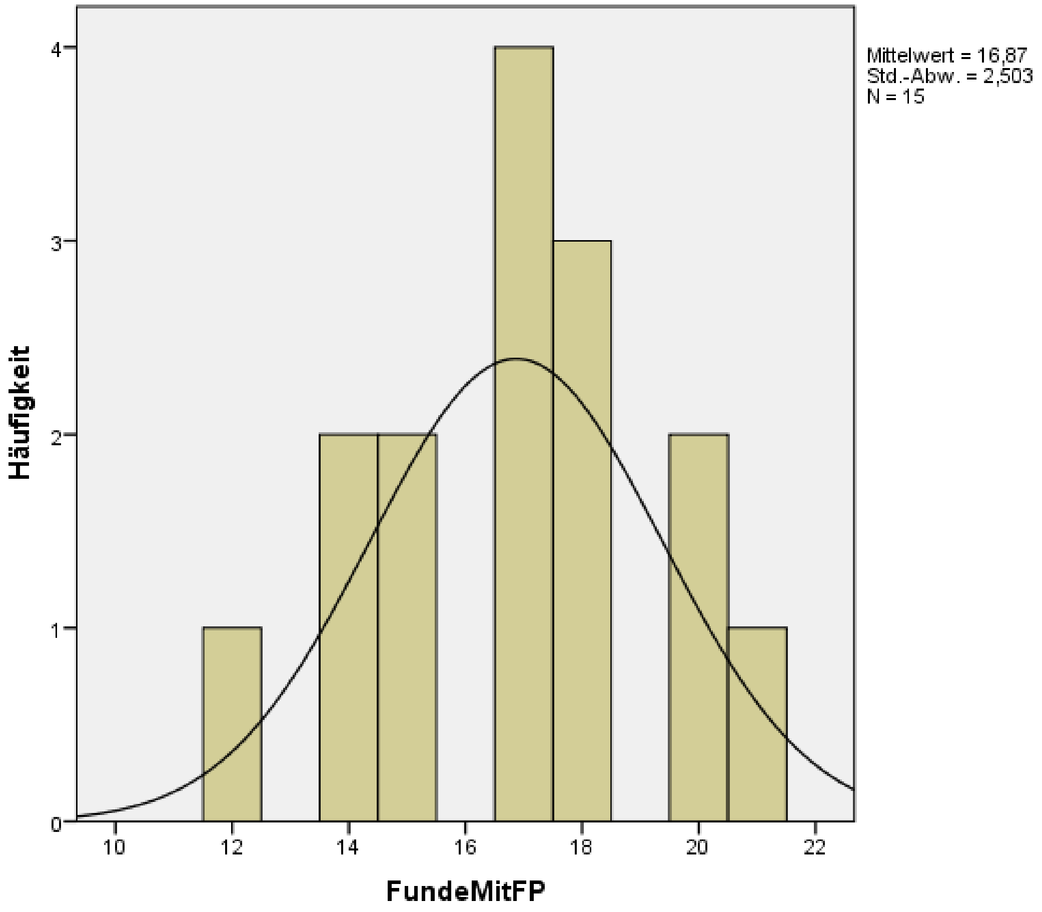


Figure H-1. Distribution graphic for group G1 including false positives.

Table H-1

Results of the Kolmogorov-Smirnov-test for normal distribution. Test for group G1 including false positives.

Kolmogorov-Smirnov-Anpassungstest

		FundeMitFP
N		15
Parameter der Normalverteilung[a,b]	Mittelwert	16,87
	Standardabweichung	2,503
Extremste Differenzen	Absolut	,188
	Positiv	,125
	Negativ	-,188
Statistik für Test		,188
Asymptotische Signifikanz (2-seitig)		,162[c]

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

c. Signifikanzkorrektur nach Lilliefors.

Appendix H-2: Group G1 excluding false positives

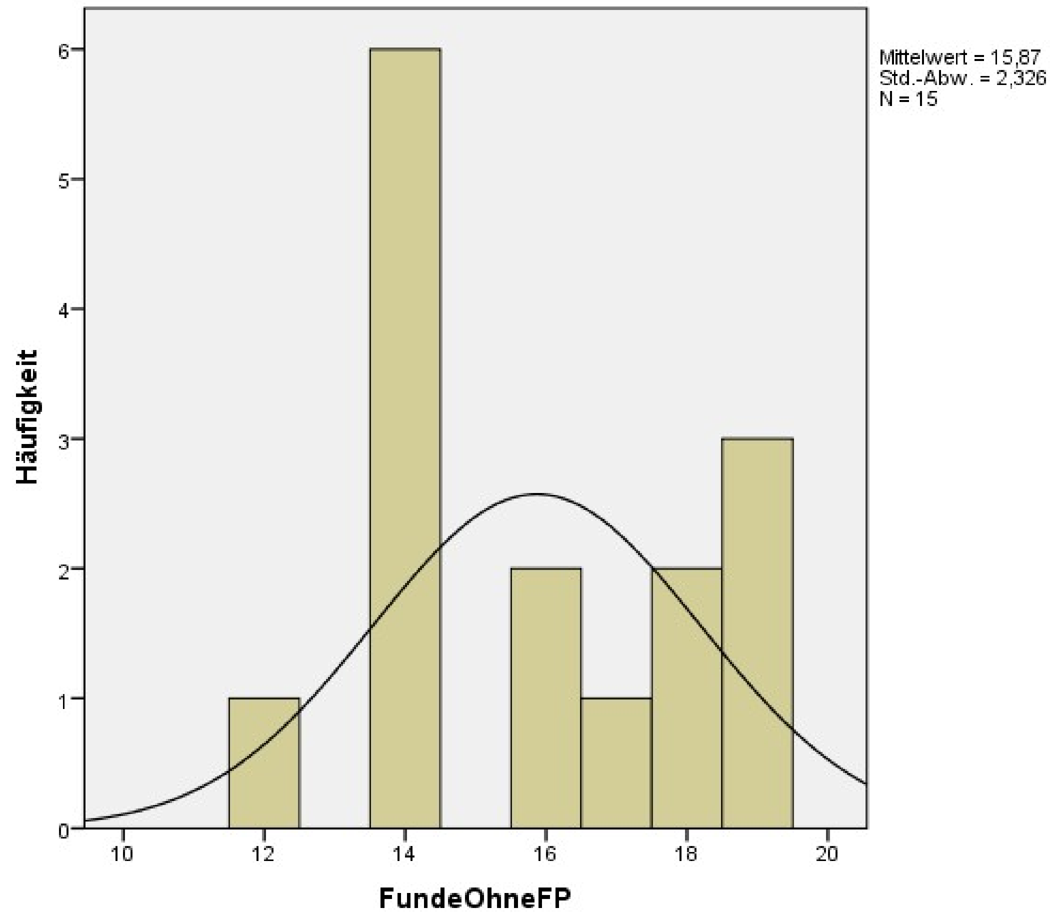
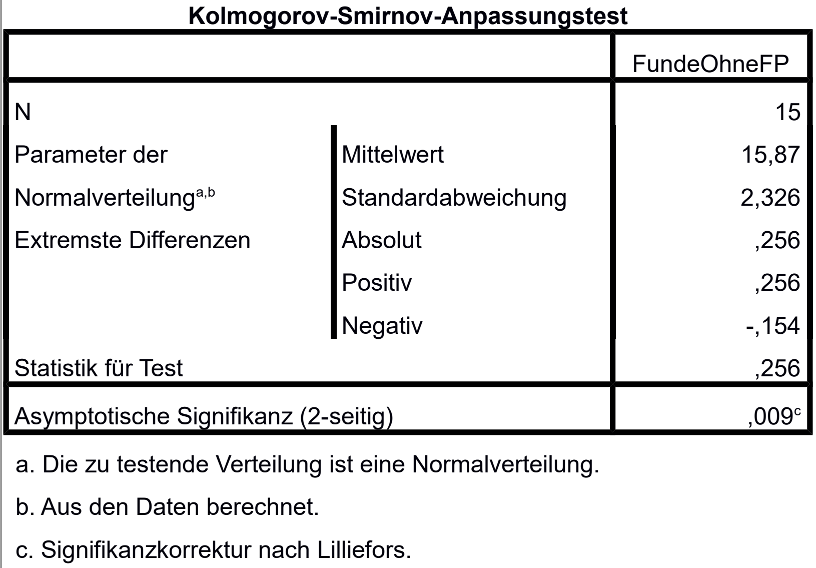


Figure H-2. Distribution graphic for group G1 excluding false positives.

Table H-2

Results of the Kolmogorov-Smirnov-test for normal distribution. Test for group G1 excluding false positives.



Kolmogorov-Smirnov-Anpassungstest

		FundeOhneFP
N		15
Parameter der Normalverteilung[a,b]	Mittelwert	15,87
	Standardabweichung	2,326
Extremste Differenzen	Absolut	,256
	Positiv	,256
	Negativ	-,154
Statistik für Test		,256
Asymptotische Signifikanz (2-seitig)		,009[c]

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

c. Signifikanzkorrektur nach Lilliefors.

Appendix H-3: Group G2 including false positives

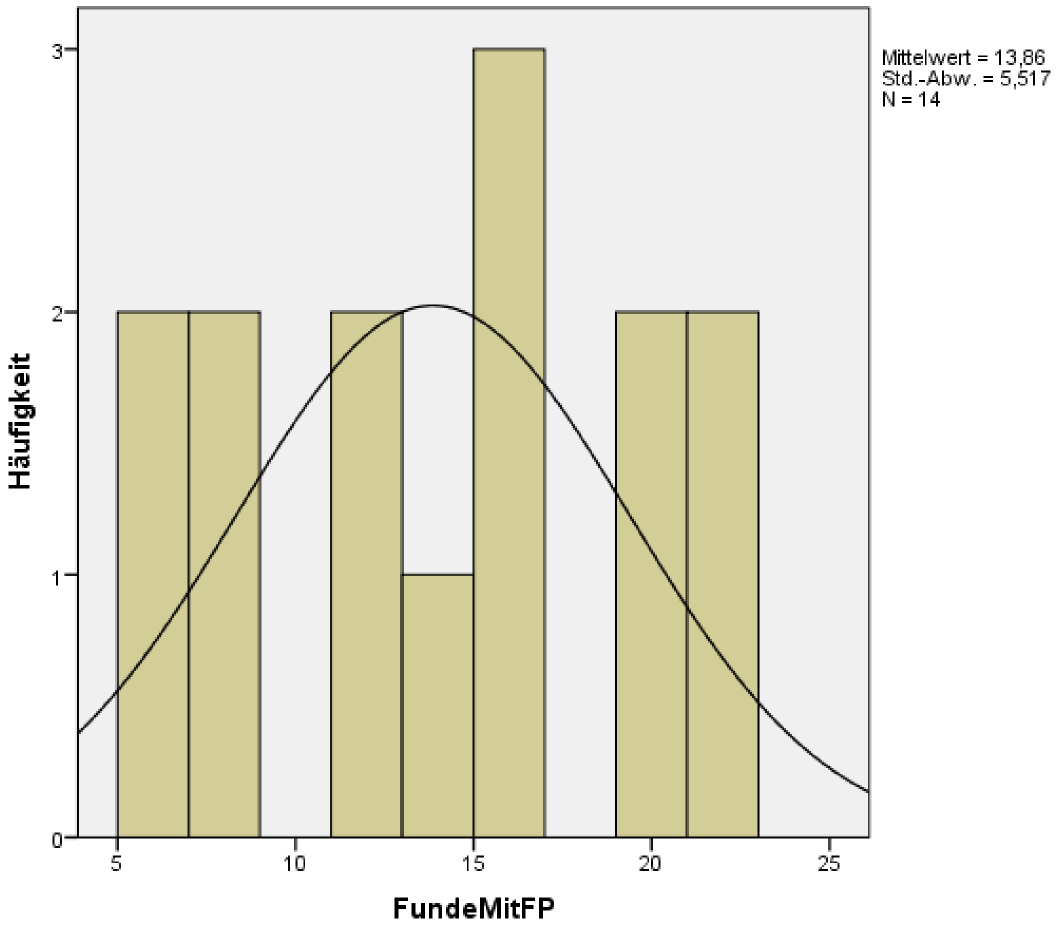
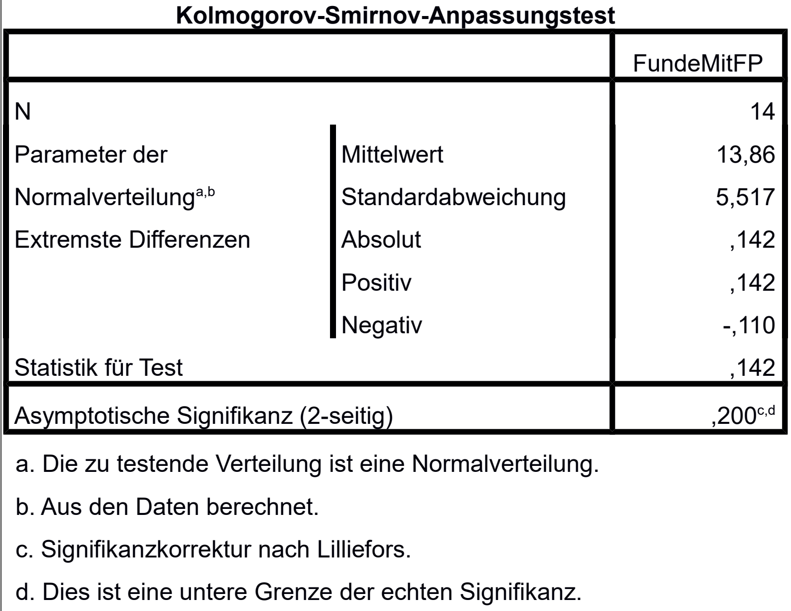


Figure H-3. Distribution graphic for group G2 including false positives.

Table H-3

Results of the Kolmogorov-Smirnov-test for normal distribution. Test for group G2 including false positives.



Kolmogorov-Smirnov-Anpassungstest

		FundeMitFP
N		14
Parameter der Normalverteilung[a,b]	Mittelwert	13,86
	Standardabweichung	5,517
Extremste Differenzen	Absolut	,142
	Positiv	,142
	Negativ	-,110
Statistik für Test		,142
Asymptotische Signifikanz (2-seitig)		,200[c,d]

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

c. Signifikanzkorrektur nach Lilliefors.

d. Dies ist eine untere Grenze der echten Signifikanz.

Appendix H-4: Group G2 excluding false positives

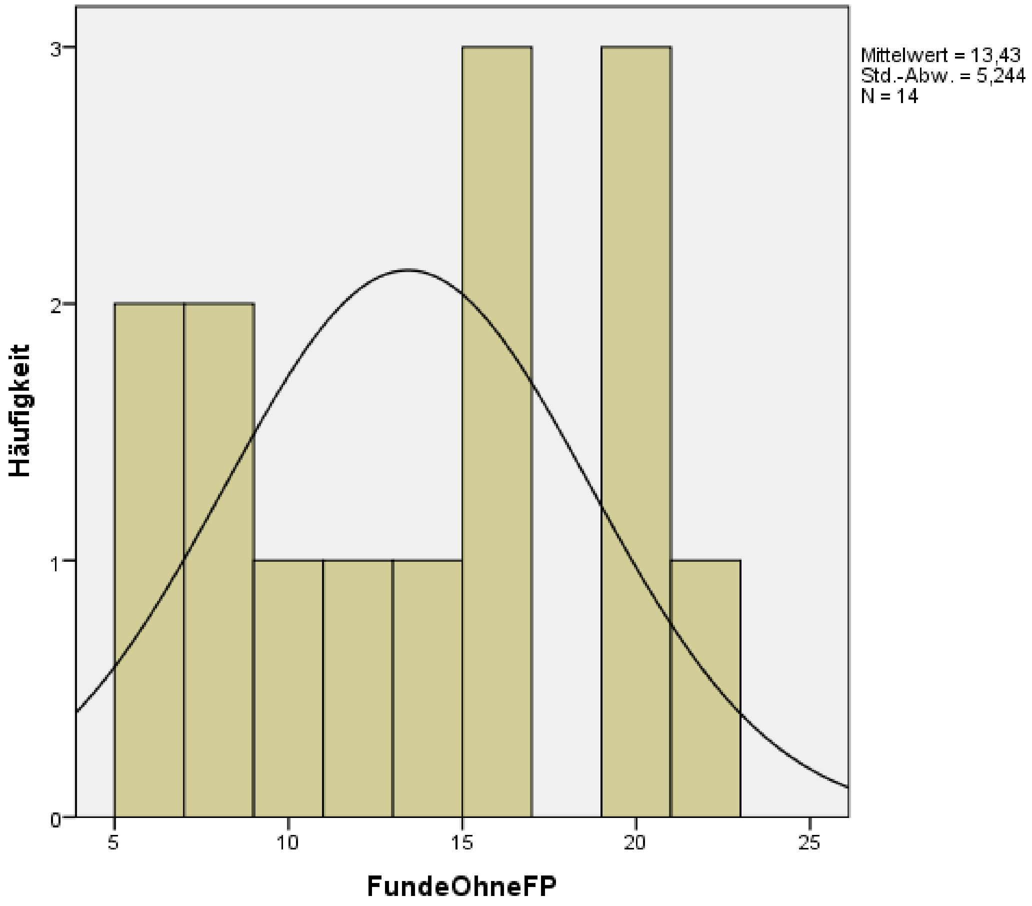
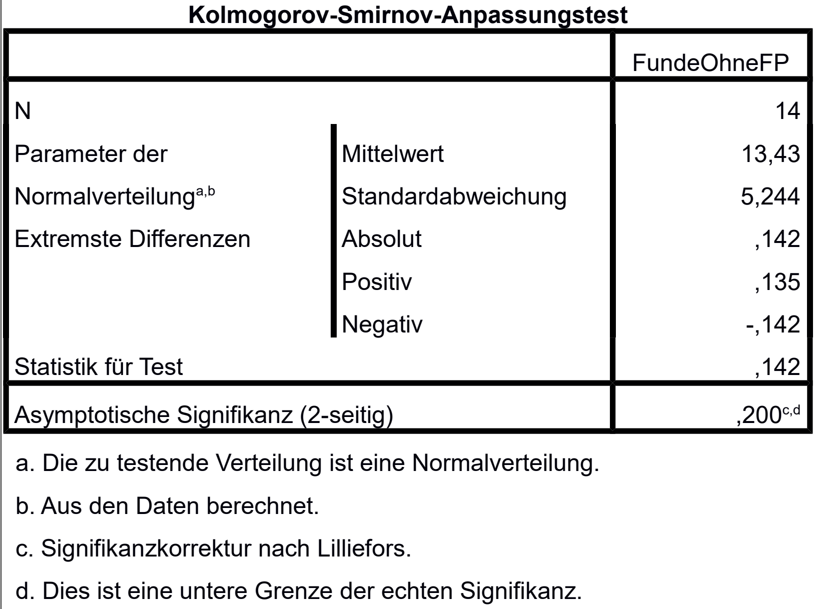


Figure H-4. Distribution graphic for group G2 excluding false positives.

Table H-4

Results of the Kolmogorov-Smirnov-test for normal distribution. Test for group G2 excluding false positives.



Kolmogorov-Smirnov-Anpassungstest

		FundeOhneFP
N		14
Parameter der Normalverteilung[a,b]	Mittelwert	13,43
	Standardabweichung	5,244
Extremste Differenzen	Absolut	,142
	Positiv	,135
	Negativ	-,142
Statistik für Test		,142
Asymptotische Signifikanz (2-seitig)		,200[c,d]

a. Die zu testende Verteilung ist eine Normalverteilung.

b. Aus den Daten berechnet.

c. Signifikanzkorrektur nach Lilliefors.

d. Dies ist eine untere Grenze der echten Signifikanz.

Appendix I

R-outputs of the calculations for the late control strategy

The data is analyzed for a specific set of data named before each output. Each output is compiled and regarded as a whole (Notebook) and not further divided or described inside the output.

Relevant variables in these outputs are:

sessions = number of participants

obs_problems = number of observed problems

s = standard deviation

D.hat = estimated completion rate

D.total = estimated number of total problems existing in the product

D.null = estimated number of undiscovered problems

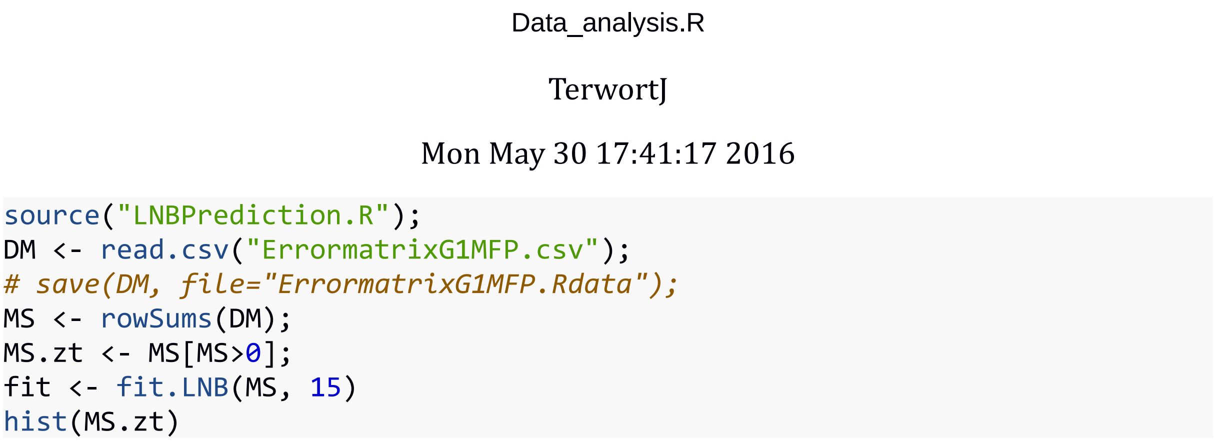
quantile(boot$D.hat… = confidence interval for the estimated completion rate

qlngeom = estimated required sample size

quantile(boot.85, … = confidence interval for estimated sample size

The illustrations contained in the output show the frequencies of new findings per participant and the calculated geometrical function that is basis for the variablescalculations.

Appendix I-1: Output for group G1 including false positives

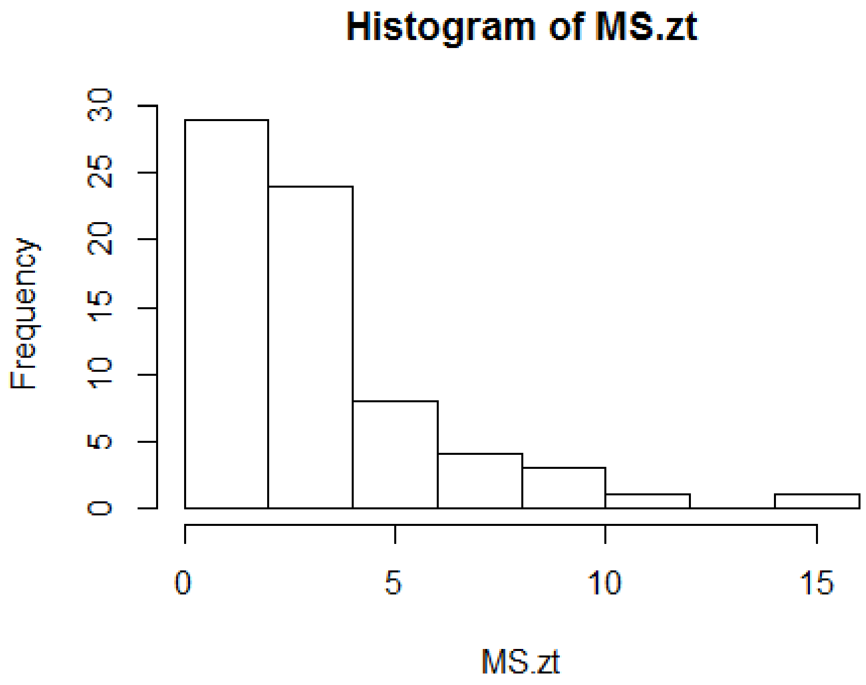


Data_analysis.R

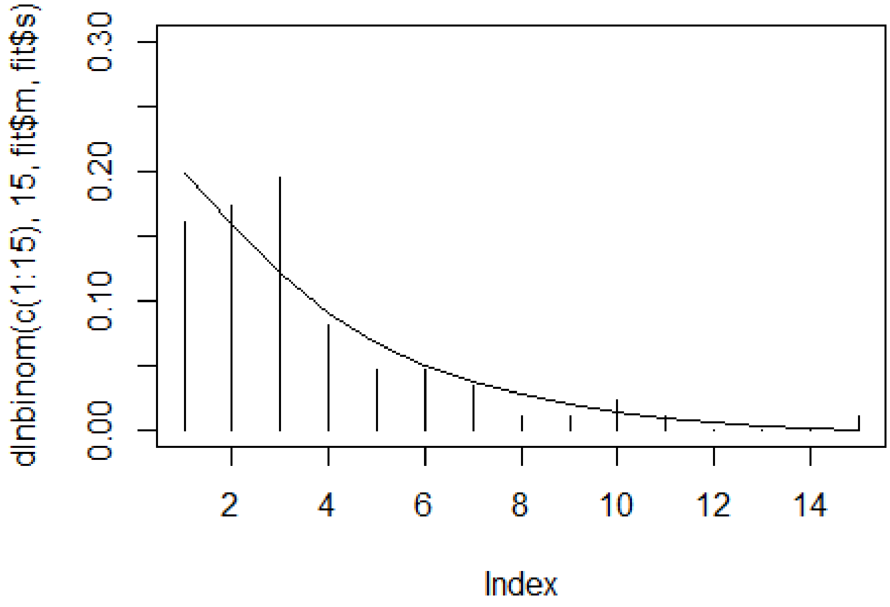
TerwortJ

Mon May 30 17:41:17 2016

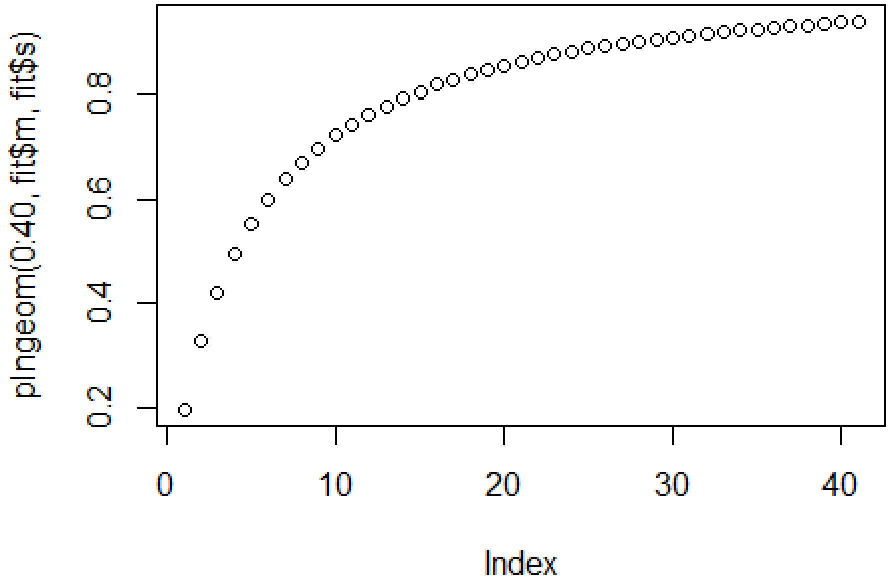
```
source("LNBPrediction.R");
DM <- read.csv("ErrormatrixG1MFP.csv");
# save(DM, file="ErrormatrixG1MFP.Rdata");
MS <- rowSums(DM);
MS.zt <- MS[MS>0];
fit <- fit.LNB(MS, 15)
hist(MS.zt)
```



```
lnb_summary(fit)
##   sessions obs_problems       mu        s     D.hat  D.total   D.null
## 1       15           70 -1.76444 1.324989 0.8064621 86.79887 16.79887
boot <- boot.LNB(MS.zt, 15, noruns=100, startval=c(-1.8,2.2));
## ==
## Warning in dbinom(x, size, exp(m)): NaNs wurden erzeugt
## =====
quantile(boot$D.hat, c(0.1, 0.9), na.rm=TRUE);
##       10%       90%
## 0.7137181 0.8790691
qlngeom(0.85, fit$m, fit$s);
## [1] 20
boot.85 <- qlngeom(0.85, boot$m, boot$s);
quantile(boot.85, c(0.1,0.9), na.rm=TRUE);
##  10%  90%
## 13.0 34.2
 plot(dlnbinom(c(1:15), 15, fit$m, fit$s),ylim=c(0,0.3), t="l");
 points(tabulate(MS.zt)/(fit$problem.n+D.null(fit)),t='h')
```



```
plot(plngeom(0:40, fit$m, fit$s))
```



```
## How many participants are neccesary for 99% completeness
  qlngeom(.99, fit$m, fit$s)
## [1] 119
```

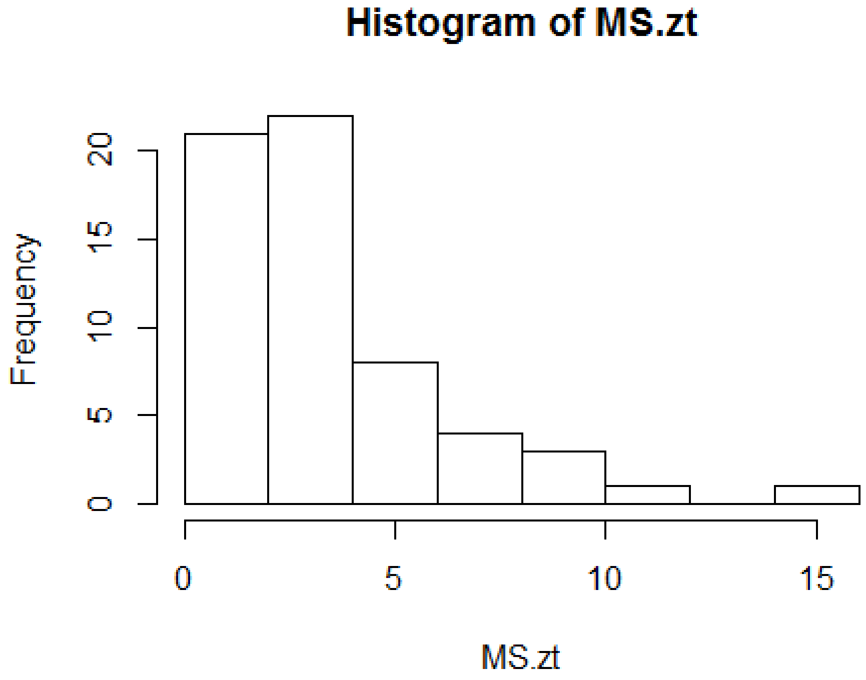
Appendix I-2: Output for group G1 excluding false positives

Data_analysis.R

TerwortJ

Mon May 30 17:46:56 2016

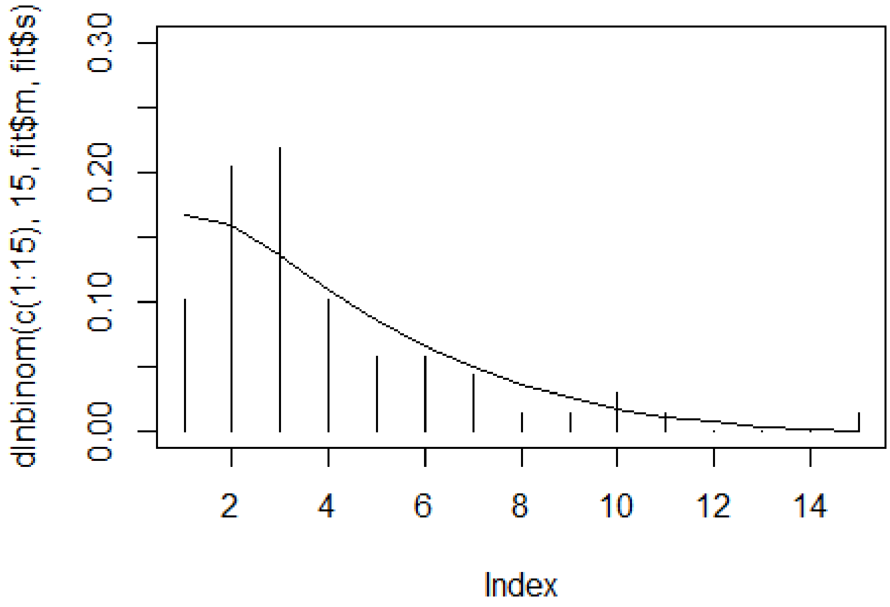
```
source("LNBPrediction.R");
DM <- read.csv("ErrormatrixG1OFP.csv");
# save(DM, file="ErrormatrixG1OFP.Rdata");
MS <- rowSums(DM);
MS.zt <- MS[MS>0];
fit <- fit.LNB(MS, 15)
hist(MS.zt)
```



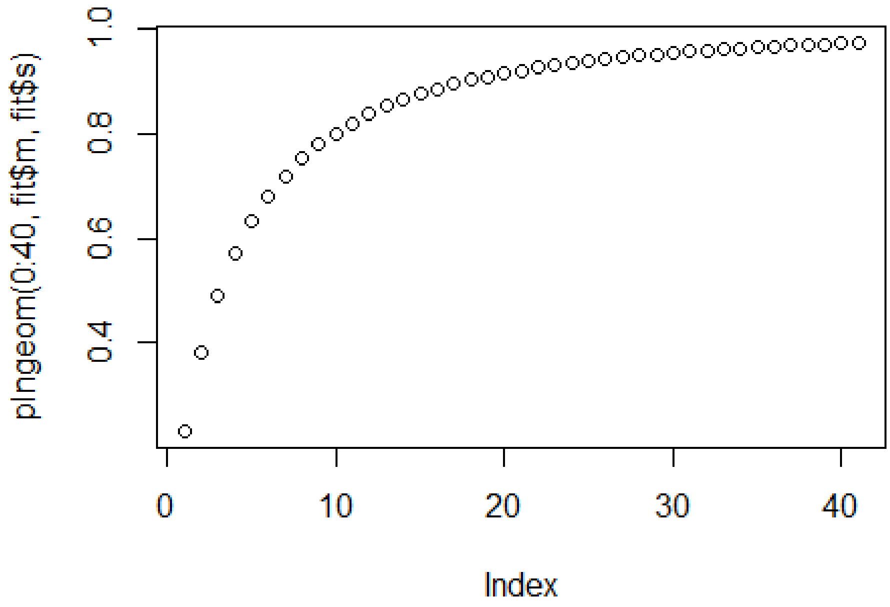
```
lnb_summary(fit)
##   sessions obs_problems        mu        s     D.hat  D.total   D.null
## 1       15           60 -1.438812 1.031953 0.8751654 68.55847 8.558467
boot <- boot.LNB(MS.zt, 15, noruns=100, startval=c(-1.8,2.2));
## ==

## ====
quantile(boot$D.hat, c(0.1, 0.9), na.rm=TRUE);
##       10%       90% 
## 0.8176893 0.9404227
qlngeom(0.85, fit$m, fit$s);
## [1] 13
boot.85 <- qlngeom(0.85, boot$m, boot$s);
```

```
quantile(boot.85, c(0.1,0.9), na.rm=TRUE);
##   10%   90%
##   9.0 18.3
 plot(dlnbinom(c(1:15), 15, fit$m, fit$s),ylim=c(0,0.3), t="l");
 points(tabulate(MS.zt)/(fit$problem.n+D.null(fit)),t='h')
```



```
 plot(plngeom(0:40, fit$m, fit$s))
```



```
## How many participants are neccesary for 99% completeness
  qlngeom(.99, fit$m, fit$s)
## [1] 68
```

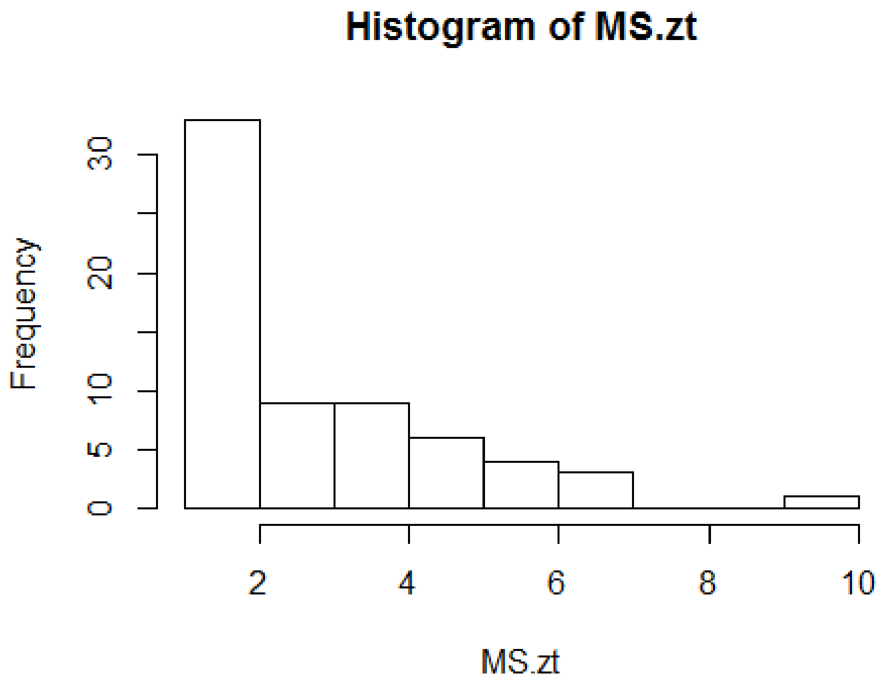
Appendix I-3: Output for group G2 including false positives

Data_analysis.R

TerwortJ

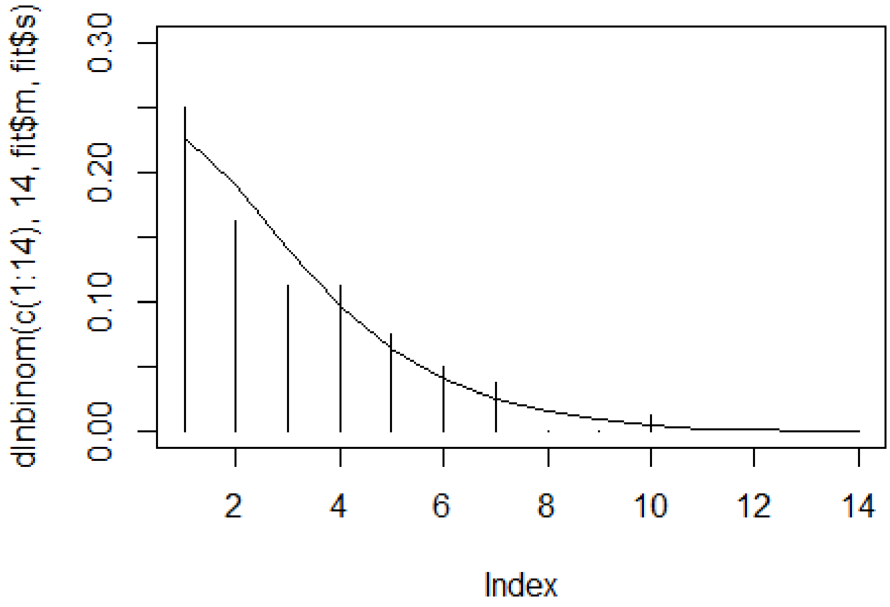
Mon May 30 17:51:15 2016

```
source("LNBPrediction.R");
DM <- read.csv("ErrormatrixG2MFP.csv");
# save(DM, file="ErrormatrixG2MFP.Rdata");
MS <- rowSums(DM);
MS.zt <- MS[MS>0];
fit <- fit.LNB(MS, 14)
hist(MS.zt)
```

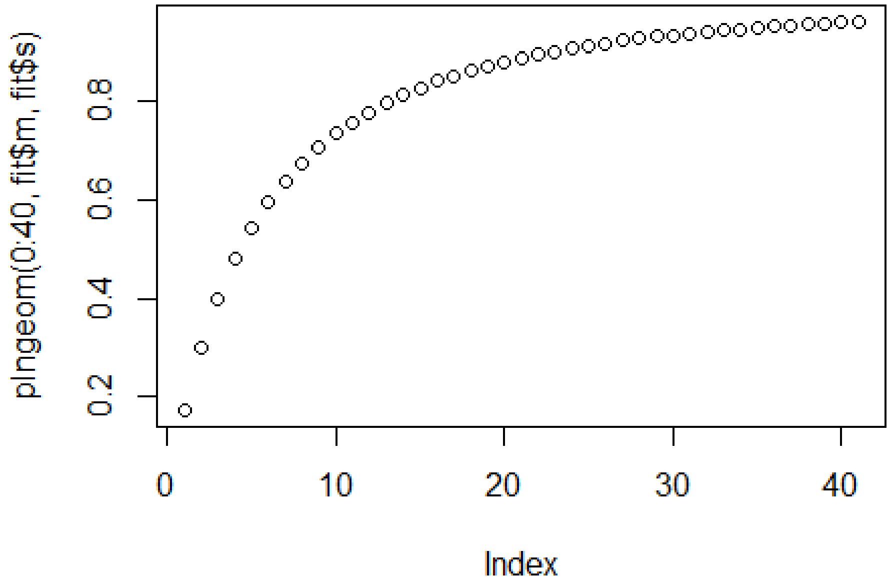


```
lnb_summary(fit)
##   sessions obs_problems        mu         s     D.hat  D.total   D.null
## 1       14           65 -1.799882 0.8026086 0.8132662 79.92463 14.92463
boot <- boot.LNB(MS.zt, 14, noruns=100, startval=c(-1.8,2.2));
## ===
## ===============
quantile(boot$D.hat, c(0.1, 0.9), na.rm=TRUE);
##       10%       90%
## 0.7134031 0.9023798
qlngeom(0.85, fit$m, fit$s);
## [1] 17
boot.85 <- qlngeom(0.85, boot$m, boot$s);
```

```
quantile(boot.85, c(0.1,0.9), na.rm=TRUE);
##  10%  90% 
## 10.9 28.1
 plot(dlnbinom(c(1:14), 14, fit$m, fit$s),ylim=c(0,0.3), t="l");
 points(tabulate(MS.zt)/(fit$problem.n+D.null(fit)),t='h')
```



```
 plot(plngeom(0:40, fit$m, fit$s))
```



```
## How many participants are neccesary for 99% completeness
  qlngeom(.99, fit$m, fit$s)
## [1] 78
```

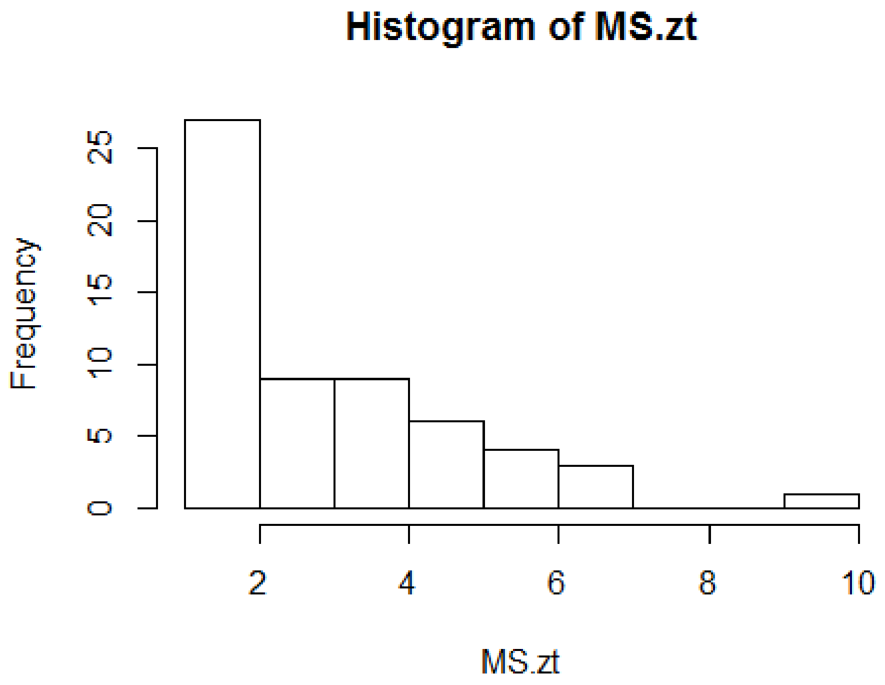
Appendix I-4: Output for group G2 excluding false positives

Data_analysis.R

TerwortJ

Mon May 30 17:54:11 2016

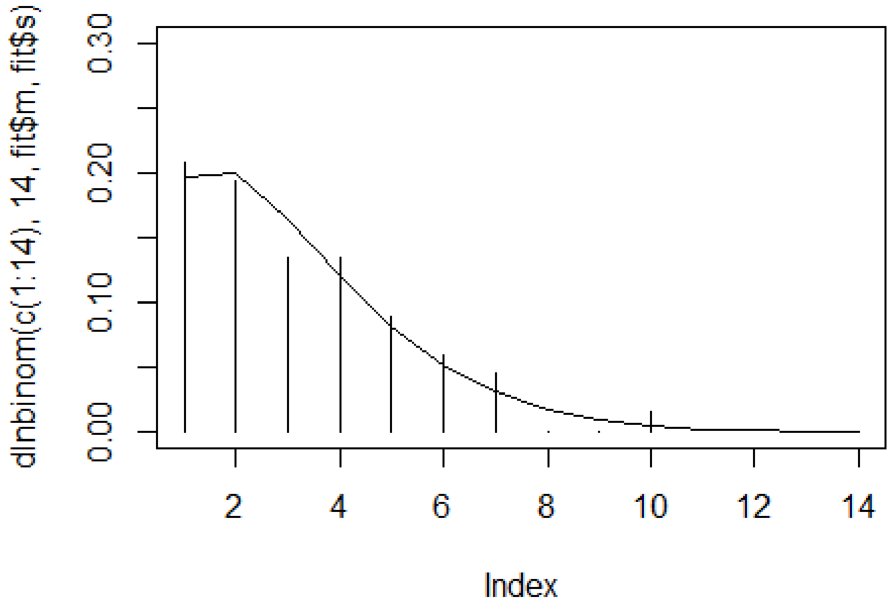
```
source("LNBPrediction.R");
DM <- read.csv("ErrormatrixG2OFP.csv");
# save(DM, file="ErrormatrixG2OFP.Rdata");
MS <- rowSums(DM);
MS.zt <- MS[MS>0];
fit <- fit.LNB(MS, 14)
hist(MS.zt)
```



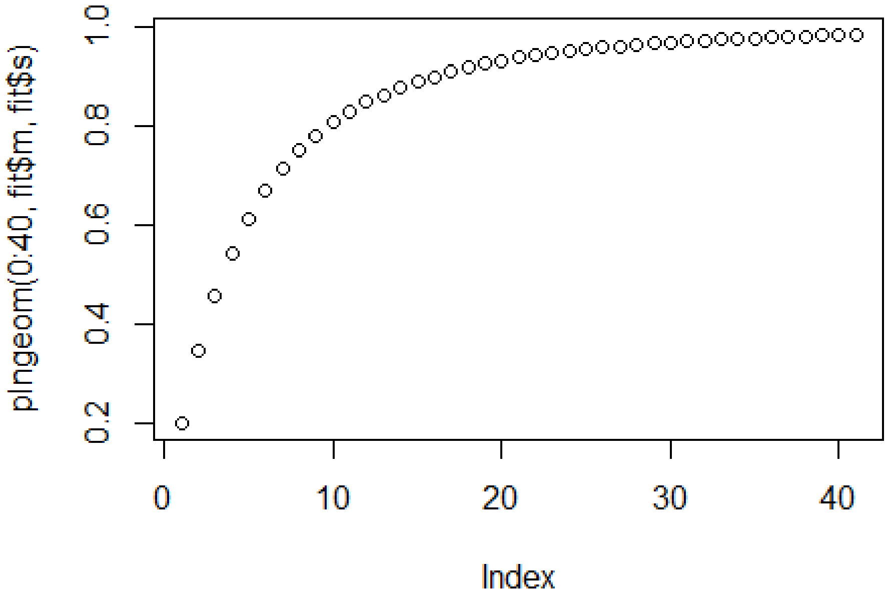
```
lnb_summary(fit)
##   sessions obs_problems        mu         s     D.hat  D.total   D.null
## 1       14           59 -1.542138 0.5585293 0.8782175 67.18154 8.181539
boot <- boot.LNB(MS.zt, 14, noruns=100, startval=c(-1.8,2.2));
## =
## Warning in dbinom(x, size, exp(m)): NaNs wurden erzeugt

## =
quantile(boot$D.hat, c(0.1, 0.9), na.rm=TRUE);
##       10%       90%
## 0.8196702 0.9640791
qlngeom(0.85, fit$m, fit$s);
```

```
## [1] 13
boot.85 <- qlngeom(0.85, boot$m, boot$s);
quantile(boot.85, c(0.1,0.9), na.rm=TRUE);
## 10% 90%
##   8  17
 plot(dlnbinom(c(1:14), 14, fit$m, fit$s),ylim=c(0,0.3), t="l");
 points(tabulate(MS.zt)/(fit$problem.n+D.null(fit)),t='h')
```



```
 plot(plngeom(0:40, fit$m, fit$s))
```



```
## How many participants are neccesary for 99% completeness
  qlngeom(.99, fit$m, fit$s)
## [1] 48
```

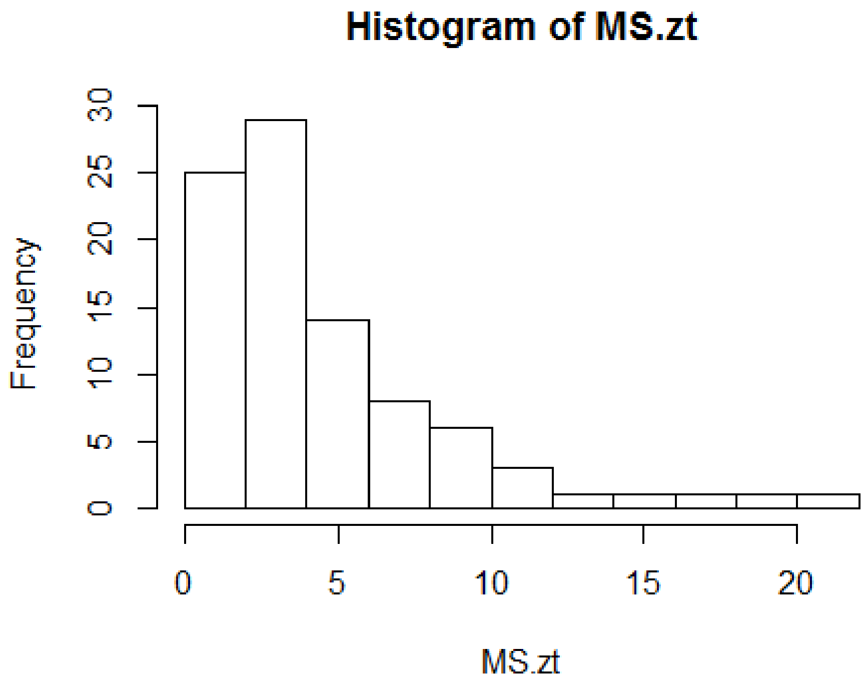
Appendix I-5: Output for group Ges including false positives

Data_analysis.R

TerwortJ

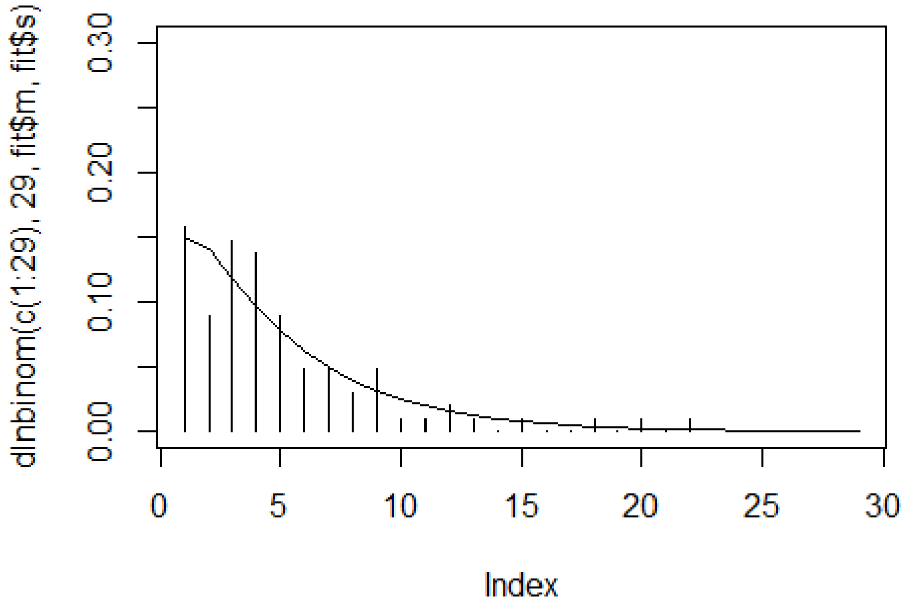
Mon May 30 17:57:41 2016

```
source("LNBPrediction.R");
DM <- read.csv("ErrormatrixGesMFP.csv");
# save(DM, file="ErrormatrixGesMFP.Rdata");
MS <- rowSums(DM);
MS.zt <- MS[MS>0];
fit <- fit.LNB(MS, 29)
hist(MS.zt)
```

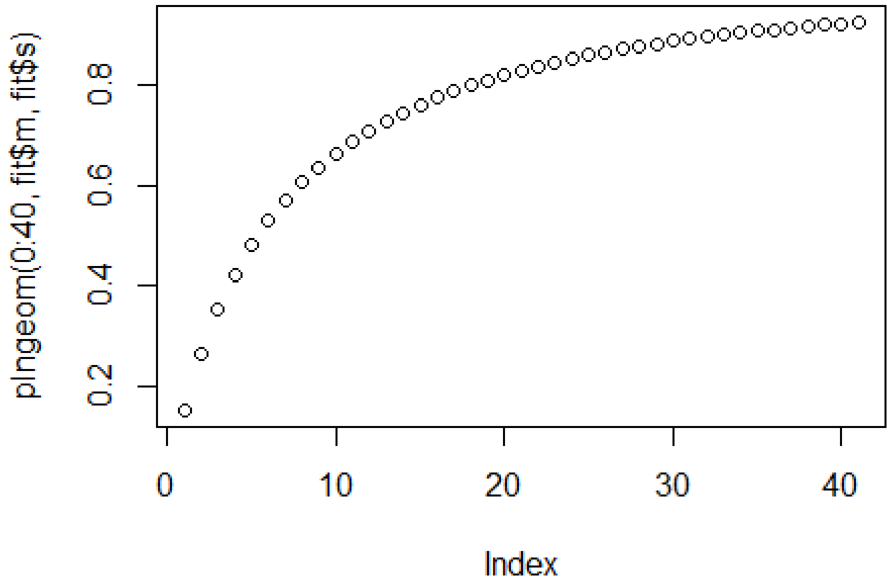


```
lnb_summary(fit)
##   sessions obs_problems        mu       s     D.hat  D.total   D.null
## 1       29           90 -2.055808 1.07999 0.8828232 101.9457 11.94566
boot <- boot.LNB(MS.zt, 29, noruns=100, startval=c(-1.8,2.2));
## =====
## Warning in dbinom(x, size, exp(m)): NaNs wurden erzeugt
## ==
quantile(boot$D.hat, c(0.1, 0.9), na.rm=TRUE);
##       10%       90%
## 0.8350201 0.9308501
qlngeom(0.85, fit$m, fit$s);
## [1] 24
```

```
boot.85 <- qlngeom(0.85, boot$m, boot$s);
quantile(boot.85, c(0.1,0.9), na.rm=TRUE);
## 10% 90%
##  18  32
 plot(dlnbinom(c(1:29), 29, fit$m, fit$s),ylim=c(0,0.3), t="l");
 points(tabulate(MS.zt)/(fit$problem.n+D.null(fit)),t='h')
```



```
 plot(plngeom(0:40, fit$m, fit$s))
```



```
## How many participants are neccesary for 99% completeness
  qlngeom(.99, fit$m, fit$s)
## [1] 129
```

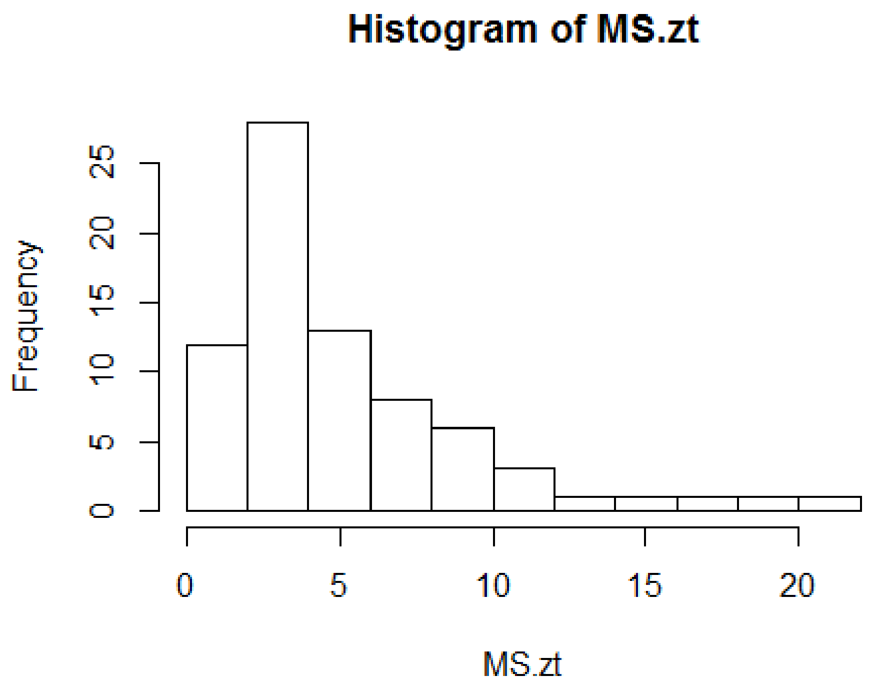
Appendix I-6: Output for group Ges excluding false positives

Data_analysis.R

TerwortJ

Mon May 30 18:00:01 2016

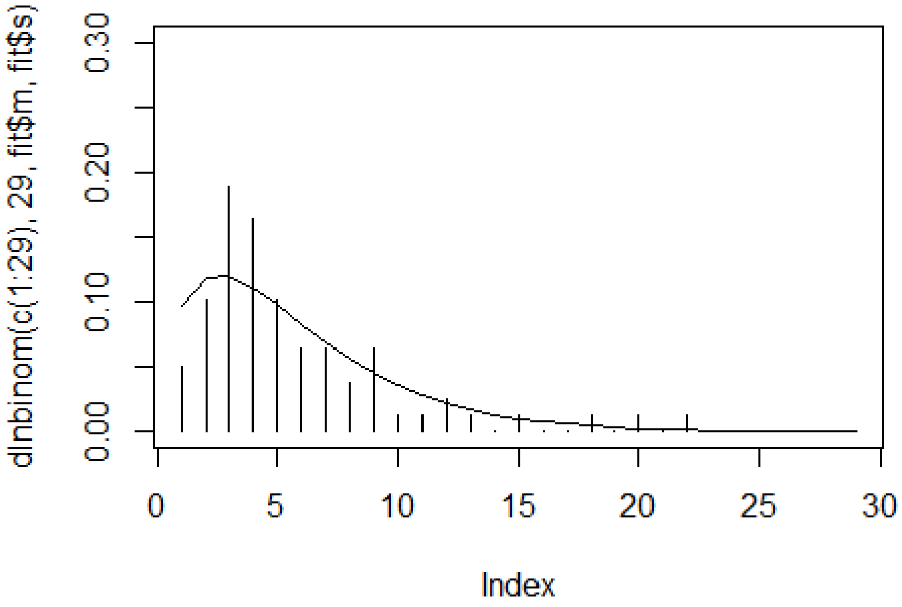
```
source("LNBPrediction.R");
DM <- read.csv("ErrormatrixGesOFP.csv");
# save(DM, file="ErrormatrixGesOFP.Rdata");
MS <- rowSums(DM);
MS.zt <- MS[MS>0];
fit <- fit.LNB(MS, 29)
hist(MS.zt)
```



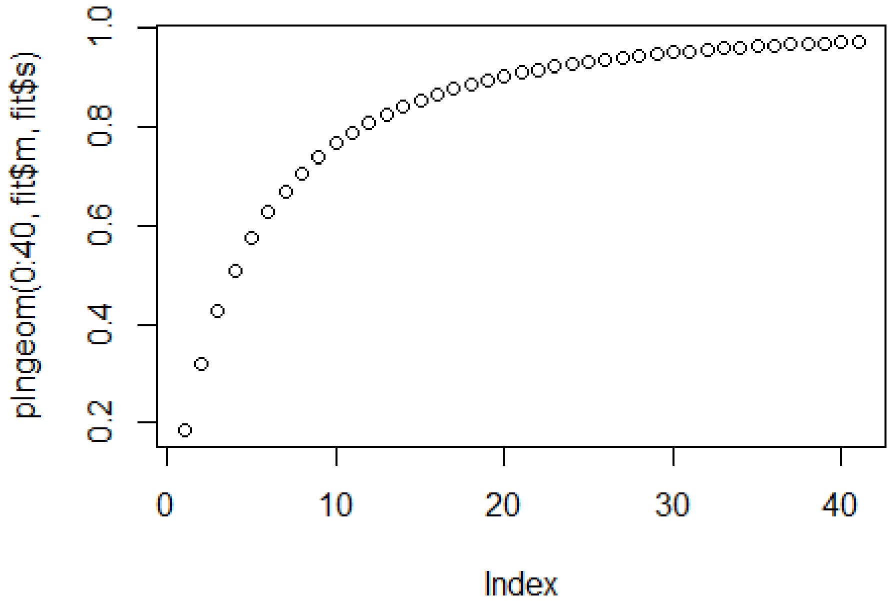
```
lnb_summary(fit)
##   sessions obs_problems         mu         s     D.hat  D.total    D.null
## 1       29           75 -1.683102 0.7080455 0.9480931 79.10615 4.106152
boot <- boot.LNB(MS.zt, 29, noruns=100, startval=c(-1.8,2.2));
## Warning in dbinom(x, size, exp(m)): NaNs wurden erzeugt

quantile(boot$D.hat, c(0.1, 0.9), na.rm=TRUE);
##       10%       90%
## 0.9272159 0.9702908
```

```
qlngeom(0.85, fit$m, fit$s);
## [1] 15
boot.85 <- qlngeom(0.85, boot$m, boot$s);
quantile(boot.85, c(0.1,0.9), na.rm=TRUE);
##  10%  90%
## 13.0 17.3
 plot(dlnbinom(c(1:29), 29, fit$m, fit$s),ylim=c(0,0.3), t="l");
 points(tabulate(MS.zt)/(fit$problem.n+D.null(fit)),t='h')
```



```
 plot(plngeom(0:40, fit$m, fit$s))
```



```
## How many participants are neccesary for 99% completeness
  qlngeom(.99, fit$m, fit$s)
## [1] 64
```

Appendix I-7: Output for additional group

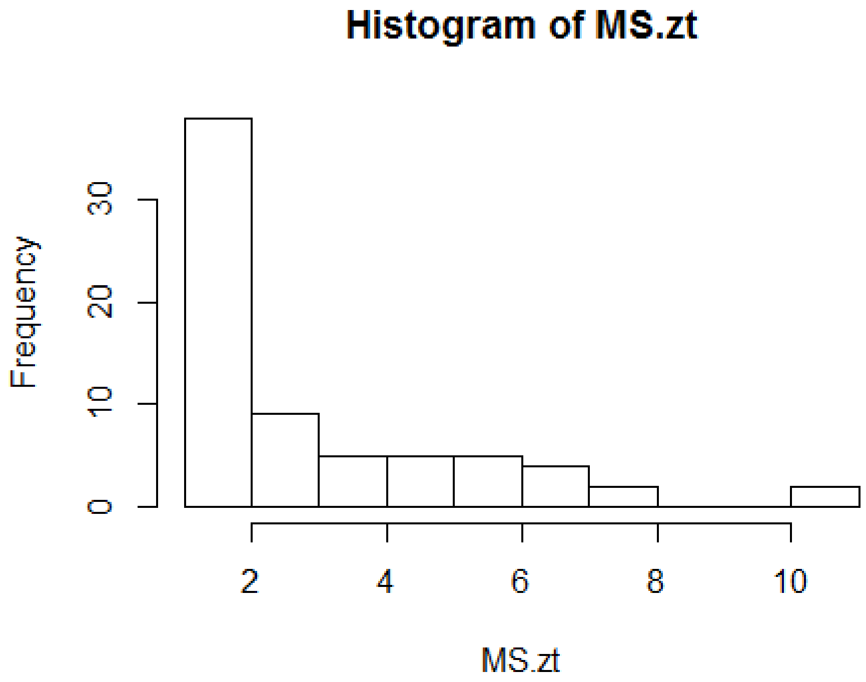
Output for the additional group only containing first half of the sample for each location excluding false positives

Data_analysis.R

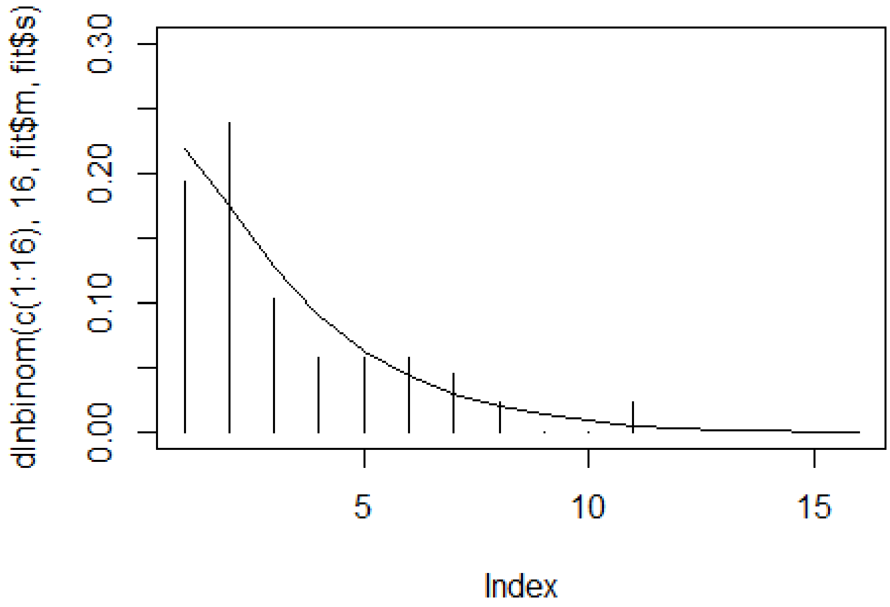
TerwortJ

Mon May 30 18:18:21 2016

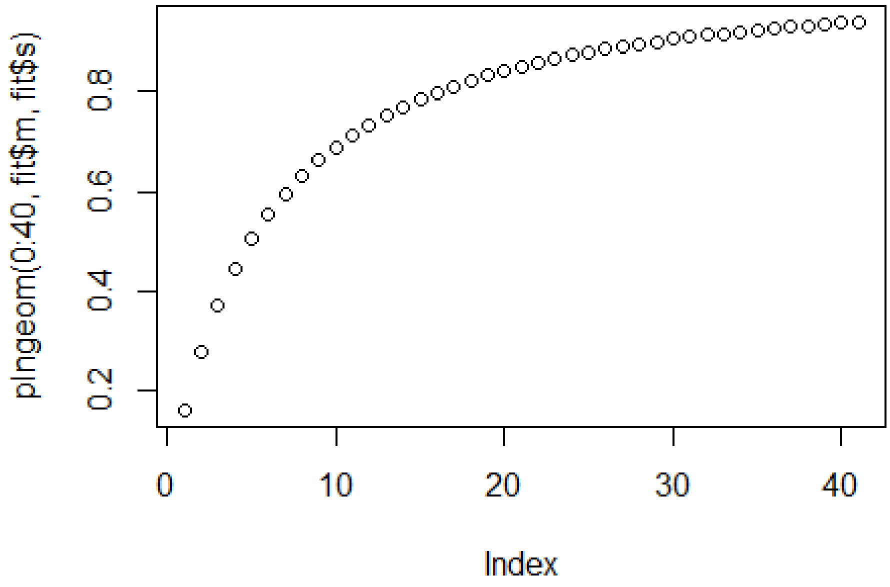
```
source("LNBPrediction.R");
DM <- read.csv("FirsthalfofsampleOFP.csv");
# save(DM, file="FirsthalfofsampleOFP.Rdata");
MS <- rowSums(DM);
MS.zt <- MS[MS>0];
fit <- fit.LNB(MS, 16)
hist(MS.zt)
```



```
lnb_summary(fit)
##   sessions obs_problems        mu        s    D.hat  D.total   D.null
## 1       16           70 -1.955577 1.027011 0.797814 87.73975 17.73975
boot <- boot.LNB(MS.zt, 16, noruns=100, startval=c(-1.8,2.2));
## ==
## Warning in dbinom(x, size, exp(m)): NaNs wurden erzeugt
## =
quantile(boot$D.hat, c(0.1, 0.9), na.rm=TRUE);
##       10%       90%
## 0.7230118 0.8782400
qlngeom(0.85, fit$m, fit$s);
## [1] 22
boot.85 <- qlngeom(0.85, boot$m, boot$s);
quantile(boot.85, c(0.1,0.9), na.rm=TRUE);
##  10%  90%
## 14.0 31.1
 plot(dlnbinom(c(1:16), 16, fit$m, fit$s),ylim=c(0,0.3), t="l");
 points(tabulate(MS.zt)/(fit$problem.n+D.null(fit)),t='h')
```



```
 plot(plngeom(0:40, fit$m, fit$s))
```



```
## How many participants are neccesary for 99% completeness
  qlngeom(.99, fit$m, fit$s)
## [1] 112
```

Appendix I-8: Output of the first four participants

Output based on only the first four participants of group G1 including false positives

```
> source("LNBPrediction.R");
> DM <- read.csv("FirstFour.csv");
> # save(DM, file="FirstFour.Rdata");
> MS <- rowSums(DM);
> MS.zt <- MS[MS>0];
> fit <- fit.LNB(MS, 4)
> hist(MS.zt)
>
> lnb_summary(fit)
  sessions obs_problems        mu        s     D.hat  D.total   D.null
1        4           37 -3.579836 6.534598 0.2696111 137.2347 100.2347
>
>
> boot <- boot.LNB(MS.zt, 4, noruns=100, startval=c(-1.8,2.2));
===========================================================================
===============================
There were 34 warnings (use warnings() to see them)
> quantile(boot$D.hat, c(0.1, 0.9), na.rm=TRUE);
       10%        90%
0.07514785 0.68815266
> qlngeom(0.85, fit$m, fit$s);
[1] 384
> boot.85 <- qlngeom(0.85, boot$m, boot$s);
 Show Traceback

 Rerun with Debug

Error in integrate(f, 0, 1, x, size, m, s, rel.tol =
.Machine$double.eps^0.5,  :
  the integral is probably divergent

> quantile(boot.85, c(0.1,0.9), na.rm=TRUE);
 10%  90%
13.0 34.2
>  plot(dlnbinom(c(1:4), 4, fit$m, fit$s),ylim=c(0,0.3), t="l");
>  points(tabulate(MS.zt)/(fit$problem.n+D.null(fit)),t='h')
>  plot(plngeom(0:40, fit$m, fit$s))
```