Quantification and Validation of Packaging Usability Measures

Erik Hörberg & Tomasz Solak

Master Thesis

Department of Design Sciences
Lund University
ISRN: LUTMDN/TMAT-5098-SE

EAT 2007

Preface

This master thesis has been written as the final part of our education in Industrial Engineering and Management at the Faculty of Engineering at Lund University. Our study was carried out in collaboration with, under the supervision of and with assistance from the Package Company and the Division of Ergonomics and Aerosol Technology at the Department of Design Sciences.

The creation of this report has been a long, but undeniably interesting, instructive and fun journey that lasted between September 2006 and February 2007. Through good times and bad, we have endured each other's company, and after many early mornings and late nights we have finally come up with results that we hope will be satisfactory and appreciated by our clients at the Package Company.

First and foremost, we would like to present special thanks to [redacted] at the Package Company, as well as to Joakim Eriksson, who acted as our supervisor at the university, for their extensive kindness and support throughout the process. We would also like to express gratitude to other personnel at the university, who kindly helped us with our perplexities. Furthermore, we would like to show our appreciation to all test participants, both young and elderly, whose contribution helped us lay the foundation for this thesis, and our families and friends for standing by us at all times.

Last but not least, we would like to thank the staff at ICA Tuna in Lund, especially those who work at the dairy and beverages department, for their assistance. Moreover, we wish to apologize for our very large but irregular purchases that gave the store's computerized replenishment system a completely erroneous picture of the actual demand. We hope you managed to get all those superfluous packages sold.

Lund, February 16, 2007

Erik Hörberg

Tomasz Solak

Abstract

Title: Quantification and Validation of Packaging Usability Measures

Authors: Erik Hörberg
Tomasz Solak

Supervisors: Joakim Eriksson, *Department of Design Sciences*
[redacted]

Problem analysis:
A prerequisite for a product to be successful is that it fulfils or exceeds the customers' expectations. A fundamental criterion for high customer satisfaction is good usability. This can be measured by performing usability tests. The Package Company has, in close cooperation with a usability consultancy, developed a method to perform usability tests with the objective to quantify package usability. However, the method had several shortcomings and despite multiple alterations, it still did not bring justice to the usability of the products being tested. The main reason for this was that the formulas used to calculate the different usability measures omitted many important aspects of the package's performance.

Purpose:
The purpose of this thesis is to examine the existing test method thoroughly and come up with suggestions for improvements in order to make the test method generate fairer results in package usability tests. Further, the purpose of the thesis is to determine the minimum amount of test participants yielding stable and reliable test results and if the test method is applicable on all kinds of user groups.

Method:
The analysis of the test method is to a large extent based on quantitative data from usability tests conducted by the authors. The required amount of test participants needed in order to ensure statistically significant results is determined with the use of theoretical statistics.

Conclusions:
The analysis of the present test method, and the tests conducted as a part of this thesis, indicates that particularly the measure of the package accuracy needs a thorough revision. An important aspect is to widen the scope of the measurement by including additional information about the performance of the tested package.

Recommendations:
The accuracy measure should be expanded in order to consider all the different kinds of spillage that may occur, as well as possible difficulties when opening and closing the package. The other components of the overall usability measure should remain in their present state. Further, the graphical presentation of the package properties should be modified in order to enhance clearness and visibility.

The statistical analysis of the usability tests concludes that using 25 test participants will provide reliable results. This is however based on the homogeneous group of participants that formed the foundation of this particular study. In order to determine whether or not this applies to all kinds of consumer segments and packages, it is advisable to conduct additional studies in the future.

Key words:
Usability, test method, packaging, competitive benchmarking.

Sammanfattning

Titel: Kvantifiering och validering av användbarhetsmått för förpackningar

Författare: Erik Hörberg
Tomasz Solak

Handledare: Joakim Eriksson, *Institutionen för designvetenskaper*
[redacted]

Problembeskrivning:
För att en produkt ska bli lyckad bör den uppfylla eller överträffa kundens förväntningar. Ett grundläggande krav för att få nöjda kunder är att produkten är användbar och ett vanligt sätt att undersöka om så är fallet är att genomföra användbarhetstester. Tillsammans med en extern konsultbyrå utvecklade the Package Company en metod för att testa användbarheten hos förpackningar, men det visade sig snart att metoden hade många tillkortakommanden. Trots ett flertal justeringar förmådde inte testmetoden att ge en rättvis bild av produkternas beskaffenhet, då den utelämnade många viktiga delar.

Syfte:
Syftet med detta examensarbete är att undersöka den existerande testmetoden och att föreslå hur den ska förändras för att på ett bättre sätt uppfylla sin funktion och generera rättvisa resultat. Vidare är syftet med examensarbetet att fastslå hur många testdeltagare som behövs för att erhålla pålitliga resultat och undersöka om testmetoden är lämplig att använda på alla slags användargrupper.

Metod:
Analysen av testmetoden baseras i stor utsträckning på den kvantitativa data som erhållits genom att utföra användbarhetstester. Den erfordrade mängden testdeltagare har fastställts genom att utnyttja statistiska räknemetoder.

Slutsatser:
Analysen av testerna och själva testmetoden tyder på att det framför allt är måttet på förpackningens precision som måste genomgå en grundlig omarbetning. En av de viktigaste punkterna är att utöka mätningarnas omfång och inkludera fler och mer noggrant definierade typer av brister hos förpackningen.

Rekommendationer:
Precisionsmåttet bör utökas och beakta såväl samtliga typer av spill som eventuellt kan förekomma som problem vid öppnande och stängande av förpackningen. De resterande måtten som utgör användbarhetsmåttet bör ej modifieras, då de redan ger en rättvis bild. Vidare bör den grafiska presentationen av förpackningens egenskaper förändras så att betraktaren får en bättre överblick.

Den statistiska analysen av testerna visar att 25 testdeltagare är tillräckligt för att erhålla stabila och pålitliga resultat. Detta är emellertid enbart baserat på den homogena grupp deltagare som ligger till grund för den här studien. Således bör fler tester genomföras i framtiden, i syfte att undersöka testmetodens tillämplighet för andra användargrupper eller förpackningar.

Nyckelord:
Användbarhet, testmetod, förpackningar, konkurrentjämförelser.

Glossary

Accuracy A rephrase of the usability ISO definition of *effectiveness*, which means how well the product serves the user’s needs.

Cognitive load A narrowed reformulation of the *learnability* expression and, accordingly, a measurement of how well the physical features of the product concur with the experienced usability.

Dashboard In the context of this study, a *dashboard* is a summary of the different measures calculated with the gathered test data as the basis, but it also includes basic information about the product and benchmarking comparisons with its competitors. In other words, it is all the essential data and statistics gathered in one place (see Appendices G and H), analogous to the *dashboard* of a car or an aircraft.

Efficiency Measures how efficient the user can carry out different tasks.

Learnability A measure of the user’s ability to handle a product after a certain amount of training.

Least competent user An end user representing the least skilled person who may use the product.

Misthreading Problems putting the cap on the threads of the package opening.

Non-disclosure agreement A contract between two parties with the intention to make sure that confidential material remains non-public. Often abbreviated NDA or CDA (confidential disclosure agreement).

Package Company A fictitious company name.

Persona A fictional character who represents one of the main types of end users. Typically, a *persona* includes a name, demographic information and key attributes.

Probing Asking questions that encourage the participants to describe their thinking in order to get a more profound understanding of their behavior.

Satisfaction The usability *satisfaction* is the user’s subjective perception of the product.

Scenario An extended representation of tasks that a real end user would have performed using the product. It usually includes a context in order to make it more realistic and motivate the test participant.

Tape consent form A written permission from the test participant to record the test session. Applies to both audio and video recordings.

Test monitor The person in charge of the test and the one who interacts the most with the participants. Generally, the *test monitor* is present in the same room as the test participant.

Transfer of learning effect A mostly negative (although sometimes actually desirable) side effect caused by doing several similar tasks after each other, giving the participant more knowledge of how to use the product. This fact results in poorer results for tasks performed at the beginning of a test session.

Table of Contents

Table of Figures

1 Introduction

The aim of this chapter is to explain the background to the existence of this master thesis and to provide the reader with an understanding of the problem statement. Further, the purpose and focus of the study are described, as is its target group. Finally, the outline of the thesis will be stated with the intention to clarify the structure of it and to give reading guidelines.

1.1 Background

When developing a new product, it is of uttermost importance to make sure it is easy to handle for the end users, as the developer most certainly will face receding sales figures otherwise. A natural way to ensure that the members of the target groups can use the newly developed product with ease is to conduct usability tests. Obviously, the same thing holds for the packaging industry and the Package Company is no exception. This led to the fact that the Package Company developed a usability test method in cooperation with a London-based usability consultancy. During implementation of the test method, some weaknesses were discovered. The method was subsequently revised and improved by the Package Company, thus deficiencies were still found.

1.2 Problem Analysis

Despite elaborations of the test method it did not completely satisfy the needs of the Package Company. The original test method was developed on the basis of software usability, thus parts of the method were not as profound as needed while other parts were too detailed. The adjustments made by the Package Company improved mainly the detailed areas but the method still had areas where information was missing. The revised method was hard to grasp and generated complex formulas. Still, it did not bring justice to the real usability of the products being tested, as several important matters were not taken into consideration.

For instance, regarding accuracy, the original method measured the amount of spillage in terms of the weight of the liquid. Another issue is the fact that spillage is not the only negative aspect of a package, as it can be difficult to open and close or the cap can be impossible to reclose once the seal is broken. The method had furthermore never been tested in a greater scale and a proper quantity of test participants had to be laid down. The external consultants had, indeed, provided a recommended amount of participants, although it was not detailed enough, ranging between 60 and 300 participants.[1] Usability tests are both time-consuming and expensive to perform, making the Package Company eager to improve the test method further so that fewer participants can be used and still keep a fair view of the products' usability. Besides, the influence of different segments on the test result had not been established.

[1] [redacted]

1.3 Purpose

The main purpose of this thesis is to aid the Package Company in the development of a fully functional version of this test method by critically reviewing the existing one and expanding it creatively before deploying and validating the results.

Secondly, the purpose of this study is to determine the amount of test participants needed to get a stable and reliable result. This also includes the question if the result starts to stabilize at the same number of participants from different segments. Finally the thesis should determine if *test monitors* influence the results or if the method is insensitive to different *test monitors*.

1.4 Focus and Delimitations

The major focus for this thesis is to determine a suitable way to describe the usability *effectiveness* (often called *accuracy* throughout this report) of a package, as this has proven to be very challenging. Our main task is to investigate the impact on the *effectiveness* when changing the components of the formula used to calculate this value. The aim is to conclude how to account for product spillage (and other error types) considering simplicity of execution and consistency of the results. Aspects that are examined, amongst others, are the weight and the area of the spilled liquid.

The results of a large-scale usability test are evaluated and statistically revised in order to determine how the usability test should be performed to generate statistically significant results and to determine the impact of the test monitor. Other areas of the package usability test method are also evaluated, but this is not the priority.

1.5 Target Group

The target group for this report is mainly the managers and employees at the Package Company and its development department. More precisely, the study is aimed at the personnel involved in usability testing and evaluation of newly developed packages.

Moreover, the report may be of interest to other students and postgraduates within the fields of usability, ergonomics and product development. Last but not least, the thesis is of course of interest to staff at the Department of Design Sciences, both those involved in the grading of this thesis and the examination of the authors, as well as others.

1.6 Outline of the Thesis

The outline of the thesis is supposed to reflect the way the study was performed to the greatest possible extent, although this is not the case throughout the whole report. The reason for this is the fact that the study is more of an explorative and iterative nature, implying that we had to make assumptions about the outcome, investigate it experimentally and then revise our views many times over. This is evident when some methods are mentioned before being properly introduced. Irrespective of that, the thesis is essentially divided into eight different parts, which are:

- Introduction
- Research methodology
- Theoretical exposition

- Empirical description
- Analysis and results
- Conclusions and recommendations
- Generalizations and future studies
- Appendices

The first part provides a background to the study, including a description of the task itself, hence being of interest to all reader categories. The second and third parts are mainly aimed at those who have no, or limited, experience of scientific research methodology or usability testing. As a natural consequence, those two parts can be omitted by readers who find themselves familiar with these areas and are only interested in the findings of our research. Unlike some other studies, the theoretical framework in this one is mostly a tool to widen the understanding and it only serves as the basis for the conclusions made in subsequent parts to a minor extent.

The fourth and fifth parts are of particular interest for those who wish to broaden their understanding of the problem analysis that underlies the conclusions and recommendations. The fourth part presents the initial state of the problem issue, what data has been gathered and how this was performed as well as a theoretical description of the data processing. The fifth part, on its hand, summarizes the shortcomings of the initial method and presents a whole range of new generic approaches on how to improve it. In addition, it deals with the statistical reliability of the gathered data.

The sixth and seventh part contains our conclusions and recommendations for the future and they are of interest for all target groups of the study, just like the appendices at the end of this report. In some theses the appendices may be of lesser importance, however, in our report they are not. This applies particularly to Appendices I and K, illustrating the results of the chart exploration technique and the usability measures for different user groups respectively. These appendices are of such an importance that only their vast size forces us to place their content outside the continuous text. Practically, the substance of these two appendices is the essence of this thesis.

Another thing that differentiates this thesis from many others is the lack of a company presentation. This is simply because of the company’s wish to remain anonymous and a chapter of that kind would have to be classified in its entirety, thus making it redundant from the beginning.

2 Methodology

This chapter will describe the way in which this thesis was conducted, regarding several aspects like methodological approaches, various research methods and how the input data was collected. Furthermore, different analysis models will be discussed and thoughts on how to make sure that the credibility of the thesis is at an acceptable level will be presented.

2.1 Scientific Approach

There are many different approaches applicable for research. Three such approaches are presented by Arbnor & Bjerke, namely the analytical approach, the systems approach and the actors approach. In addition to these three approaches there is another classification model, covering explanatory and understanding knowledge, which are also known as explanatics and hermeneutics.[2] How these two parallel systems are related to each other is visible in Figure 2.1 below.

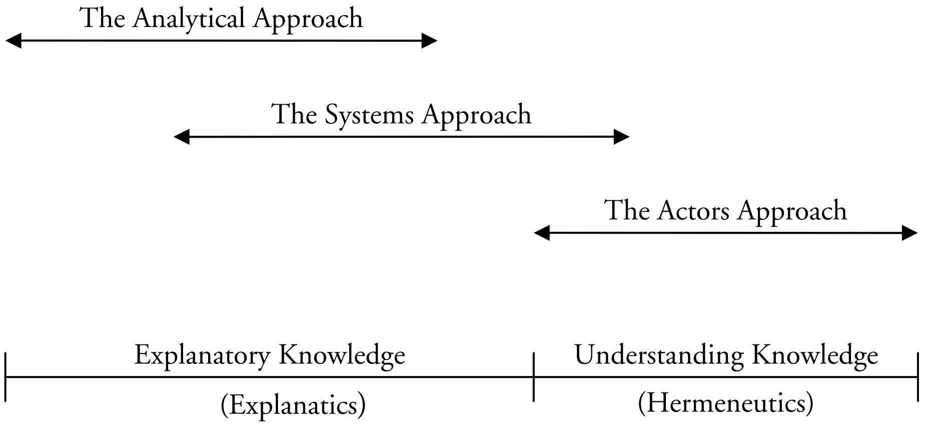


Figure 2.1 *The Relation between Explanatics and Hermeneutics*[3]

2.1.1 The Analytical Approach

The analytical approach has its roots in analytical philosophy, and as a consequence it is very common in Western thinking. The essence of this approach is that the whole is the exact sum of its parts, neither less nor more. This fact implies that a problem can be divided into a number of smaller problems and when these subproblems are solved, so is the original problem.

Another important fact regarding the analytical approach is that the knowledge created using this approach is considered to be independent of the observer. Thus, the knowledge is free to move and accessible to all having the required competence.[4]

In order to illustrate the analytical approach, consider a band consisting of the most skilled guitarist, bassist, drummer and vocalist. This would be the best band there is according to the

[2] Arbnor, I & Bjerke, B (1997), pp. 47-49

[3] *Ibid.*, p. 46

[4] *Ibid.*, p. 50

analytical approach, as no other combination of musicians can generate a higher sum of skills. However, what is overlooked is the fact that the personal chemistry between the band members may not the best or the fact the musicians may play different genres, causing the music to sound terribly bad.[5]

2.1.2 The Systems Approach

In contrast to the analytical approach, the whole differs from the sum of its parts when it comes to the systems approach. This makes not only the parts themselves essential, but likewise are the relations between the parts. In this manner, there will be some synergy effects as these relations may improve or worsen the outcome.

If the band members from the earlier example (see Section 2.1.1) were chosen on a systems approach basis, the final result would probably be much better. Not only would the personal chemistry be taken into account, but they would all play the same genre. Moreover, they would also adjust their instruments to the acoustics in the room. This means that the optimal band constellation and music instrument adjustments would depend on a vast number of factors.[6] None of these factors could be altered without affecting the outcome.[7]

The knowledge developed through the systems approach is depending on systems, and individuals are described as systems characteristics. This leads to the fact that the systems approach explains parts through the characteristics of the whole.[8]

The results of studies performed using the systems approach cannot be seen as absolute. If the study would be conducted once more, the result may very well be different of that from the first study. As a consequence, the results and conclusions from earlier studies cannot be directly transferred to another study.[9]

2.1.3 The Actors Approach

On the contrary to the systems approach, the actors approach describes the whole in terms of the characteristics of its parts. This approach is somewhat different from the above-mentioned ones as it is more interested in understanding social constructions instead of finding explanations.[10]

The fact that the reality is a social construction makes it dependent of its observers and the people that constitute the reality. This leads to that the knowledge that has been developed through the actors approach is based on how the actors act in a reality they have created themselves.[11]

Referring to the band once again, using an actors approach when choosing the band members, the main focus would be on the band as one entity and the relation between the band members, as well as the relation between them and the tour manager and the record company staff.

[5] Arbnor, I & Bjerke, B (1997), p. 50
[6] *Ibid.*
[7] *Ibid.*, p. 65
[8] *Ibid.*, p. 52
[9] *Ibid.*, pp. 67-68
[10] *Ibid.*, p. 52
[11] *Ibid.*, pp. 71-75

2.1.4 Explanatics

Researchers who assume that models and methods successfully used in natural sciences can be applied to other sciences, without any significant differences in the quality of the studies, are called explanaticists.[12] All knowledge has to be empirically verifiable if it is to be called science. Another thing of great importance is objectivity; researchers cannot be affected by unscientific values.[13] As seen in Figure 2.1, explanatics cover the same area as the analytical approach and most of the systems approach.

2.1.5 Hermeneutics

Researchers who oppose the explanaticists and claim that there is a distinction between the models used in natural sciences and other sciences, and that the models are not interchangeable, are called hermeneuticists.[14] Hermeneutics is about the understanding of the meaning of texts, symbols and actions. While trying to understand them, there is a continuous change back and forth between an overall perspective and a partial perspective. It is also important to be aware of the circumstances and how they affect a study.[15] In hermeneutics there is a distinction between explaining nature and understanding culture.[16] Hermeneutics are similar to the actors approach as they regard understanding knowledge.

2.1.6 The Scientific Approach in this Thesis

As this thesis is mainly about collecting vast amounts of data and drawing conclusion based on this data, it may seem like the analytical approach would be a suitable one. However, this approach considers the information to be independent of the observer, thus making it inappropriate for this study. The systems approach would be much better, as the results cannot be seen as absolute. If the test participants were replaced by new ones, there is little doubt the outcome would not have been the same.

The actors approach is a bit different, as it is mainly about understanding how the actor behaves and why this behavior occurs. Even though this seems not to be fully compatible with a study that includes large amounts of statistical data, one must bear in mind that to a large extent, the statistical data is more of a way to express how the actors experience the product. Following this, the results are highly dependent on the actors' interpretation of expressions like *good*, *problem*, *easy to use*, *difficult*, *practical*, *annoying* and the like.

As a result, this thesis will mainly use the actors approach, although some elements of the systems approach will be taken into consideration as well. This implies that the study will be conducted mostly using hermeneutics.

2.2 Research Methods

There are several research methods, all with different strengths and weaknesses. When making a scientific study it is very important to make clear which methods can be used in general and which one is the optimal to use in that specific case.

[12] Arbnor, I & Bjerke, B (1997), p. 45
[13] Wallén, G (1996), pp. 26-27
[14] Arbnor, I & Bjerke, B (1997), p. 45
[15] Wallén, G (1996), pp. 33-34
[16] Arbnor, I & Bjerke, B (1997), p. 45

2.2.1 Inductive, Deductive and Abductive Methods

There are two main methods when it comes to tackle scientific studies; the inductive and the deductive method.[17] The inductive method implies that the researcher gathers empirical data in order to construct theories.[18] This is done by summarizing regularities in observations of the reality. The observations have to be completely unbiased[19], which is extremely difficult to accomplish, and consequently, this fact becomes one of the weakest points of this method. Another point of criticism against the inductive method is that the theories it leads to contain nothing but what is found in the empirical data.[20]

The deductive method gives theories a more central position than induction, as they are the starting point and not the end. Using already existing theories, the researcher tries to deduce new hypotheses and subsequently test those using empirical studies.[21] In order to conduct proper hypothesis testing, the researcher has to have extensive knowledge about the subject concerned.[22] As a theory never will be complete, further studies will either strengthen or weaken the theory.[23] A weakness with the deductive method is that the researcher will be more apt to search for and, therefore, find information and data that proves the theory right.[24] Figure 2.2 illustrates the difference between the inductive and the deductive method and their relation to reality and theory.

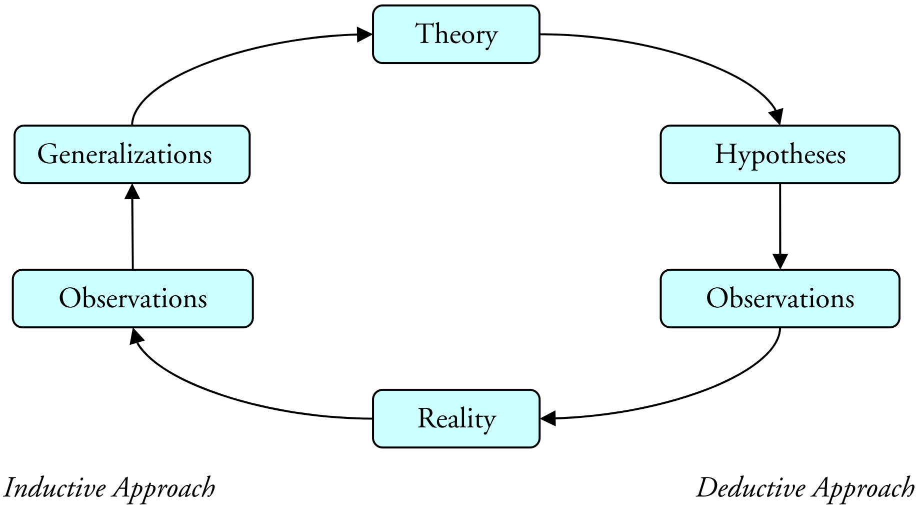


Figure 2.2 *The Relation between the Inductive and the Deductive Approach*[25]

In addition to these two main methods, there is also a third approach to use, namely the abductive method, which is a way to conclude the cause of an observation. It is somewhat different from the deductive method and bears more resemblance to the inductive method, as the researcher examines the factors causing a certain effect without actually manipulating

[17] Holme, I M & Solvang, B K (1997), p. 51
[18] Arbnor, I & Bjerke, B (1997), p. 92
[19] Wallén, G (1996), p. 47
[20] *Ibid.*, p. 89
[21] Holme, I M & Solvang, B K (1997), p. 51
[22] Wallén, G (1996), p. 48
[23] Holme, I M & Solvang, B K (1997), p. 51
[24] Jacobsen, D I (2002), p. 35
[25] Wiedersheim-Paul, F & Eriksson, L T (1991), p. 150

these factors. In other words, the researcher uses both existing theories and the gathered empirical data by turns. A negative aspect of the abductive method is that it requires thorough experience within the subject of the study.[26]

2.2.2 Qualitative and Quantitative Research Methods

There are two main methodological research methods regarding data collection; qualitative and quantitative. The most significant difference between the two is the usage of numbers and statistics. The qualitative method is not as formal as the quantitative and the purpose of such a study is to get a more profound knowledge and understanding of the problem being studied.[27] Moreover, the purpose of the researcher is to understand how people feel about themselves, their environment and their existence rather than getting an absolute measure of the same things.[28] The term qualitative refers to the fact that the collected data consists of identifiable characteristics, although they cannot be numerically graded. The data is often collected through interviews or field studies and of a narrative and expressive nature.[29]

Quantitative methods, on the other hand, are more formalized and structured. The data is collected in form of numbers instead of words, as was the case with qualitative studies. In order to draw the correct conclusions, the study must be very carefully prepared and tested in advance (pilot tests) to make sure everything will run smoothly, because no changes can be made when the study is in progress.[30] Using this method makes it easier to get a general view of the subject as this method makes it possible to statistically process the gathered data.

2.2.3 Research Methods in this Thesis

The task of this thesis was to review the existing method and to come up with possible improvements before conducting user tests to collect the data that would form the basis of our conclusions. This implies that the deductive method would be the best one to use, but the tests were spread out during a rather long period of time and several new ideas and thoughts were developed between the beginning and the end of the user testing period. As a result of this, it is fairer to say that elements of the abductive method were used in addition to the deductive one.

Regarding the methodological research methods, we have mostly used the quantitative one, as a lot of data have been gathered and processed statistically, which resulted in many graphs, comparison values and tables. Some of the values actually indicate answers that are more qualitative than quantitative, but as the task was to enhance a way of presenting the properties of a product and how good it is compared to its competitors, we have tried to quantify all qualitative data as much as possible.

2.3 Data Collection

Collecting necessary data can be made in various ways, such as observations, experiments, interviews and literature studies. Irrespective of the choice of data collecting method, it is essential that the preparations and planning of the project are decided before the data collection commences.[31]

[26] Wallén, G (1996), p. 48
[27] Holme, I M & Solvang, B K (1997), pp. 13-14
[28] Lundahl, U & Skärvad, P-H (1999), p. 101
[29] Wallén, G (1996), p. 63
[30] Holme, I M & Solvang, B K (1997), p. 81
[31] *Ibid.*, pp. 172-173

2.3.1 Primary and Secondary Information

It is important to consider whether or not the information should be collected from scratch or if there is a possibility to use data that has already been collected by someone else. When choosing which approach is the most suitable, it is of uppermost importance to decide whether the already available data can be trusted or not.[32] One must never take the reliability of a source for granted and to ensure relevance in the collected data there is a main rule stating to always search for the primary source.[33] When using secondary sources it is of great importance to establish whether or not the source can be considered objective and this could be done by seeking what benefits the author could have in distorting or altering the facts.[34]

2.3.1.1 Observations and Experiments

Observations and experiments are excellent ways of detecting the cause and effect of usability problems. This can be done by observing how a test subject is managing different artificial situations. The experiments can be conducted in numerous ways and the methods that will be used in this thesis are further described in Section 4.2. One of the main problems when conducting usability tests is that large numbers of test subjects may be needed to ensure statistical significance. Usability tests are therefore often a very time-consuming way of collecting data.[35]

2.3.1.2 Questionnaires and Interviews

Interviews assure primary information with a direct contact between the person asking the questions and the respondent. This can be carried out either in a phone interview or face to face. The difference between interviews and questionnaires is that the questionnaires are filled out by the person answering the questions whilst the interviewer notes the answers himself. An important aspect when conducting a questionnaire is to try to avoid making it too long or complicated, otherwise the participant might lose interest in filling out the form.[36] This way of collecting information often results in a large quantity of data, which makes it very important to design the questions in such a way that it facilitates the compilation.

2.3.1.3 Literature Studies

Literature studies are more straightforward than interviews but do not always provide the exactly right type of information. Outdated information is the main problem when using scientific literature.[37] Seeking information in literature is nevertheless very important in the early stages of an investigation to obtain a broader understanding of the problem. Many types of scientific literature are often secondary which is associated with a risk that the information will not be fully objective or partially altered.[38]

2.3.2 Data Collection in this Thesis

Our intention is to search for information in primary sources as far as possible. When not possible, as may be the case for much of the theory parts where we have to rely on scientific literature that has been summarized in various books, we trust these authors not to have a hidden agenda in changing the contents of the main source.

[32] Wiedersheim-Paul, F & Eriksson, L T (1991), p. 76

[33] Ejvegård, R (1996), p. 15

[34] *Ibid.*, p. 18

[35] Bell, J (2000), p. 21

[36] Holme, I M & Solvang, B K (1997), p. 173

[37] *Ibid.*, p. 138

[38] *Ibid.*, pp. 131-132

This thesis is an evaluation of an existing usability test method. Our intention is, however, to generate all the needed data by ourselves to establish credibility in our report. As a consequence, the data and test values collected as a part of the original development of the test method will not be taken into consideration.

Our analysis and results will mainly be based on the observations done in a usability laboratory where we monitor test subjects whilst performing different tasks. This will undeniably provide us with primary source data to work with.

2.4 Analysis

The analysis is the process where the empirical data is compared with the theory and problem statement to obtain an understanding of areas that may need an improvement and will hopefully render in the finding of the best possible solution.[39] The analysis can be divided into a quantitative and qualitative approach.

2.4.1 Quantitative and Qualitative Analysis

The quantitative analysis is based on a large amount of data that has been collected from the observations of the test subjects. The data can be summarized in frequency tables and calculations of standard deviation and mean values can be conducted to clarify whether the empirical values support the hypothesis or not. These calculations help out in understanding if the results are statistically significant or not.[40]

The analysis of the quantitative data can be divided into three steps;

- *Description* – This step prompts us to ask ourselves the following questions; what kind of information can we detect in the collected data? What does it tell us?

- *Categorization* – It is important to categorize and reduce the immense amount of collected information to make it more manageable and to get a better overview.

- *Combination* – When the categorizing has been done it is time to interpret the data to search for generalizations or causes of trouble. This should lead to the finding of hidden relationships between different kinds of information.

One of the big advantages of the qualitative analysis, on its hand, is that the difference between planning, implementation and analysis are rather insignificant. It is very well possible to carry out a number of interviews or observations, analyze them and then return to the problem statement phase and change the work methodology. After this the tests can be run all over again, as shown in Figure 2.3. It is therefore possible to assure that the project method always will be adapted according to new knowledge gained during the project.[41]

[39] Lundahl, U & Skärvad, P-H (1999), p. 106
[40] *Ibid.*, p. 98
[41] Jacobsen, D I (2002), p. 216

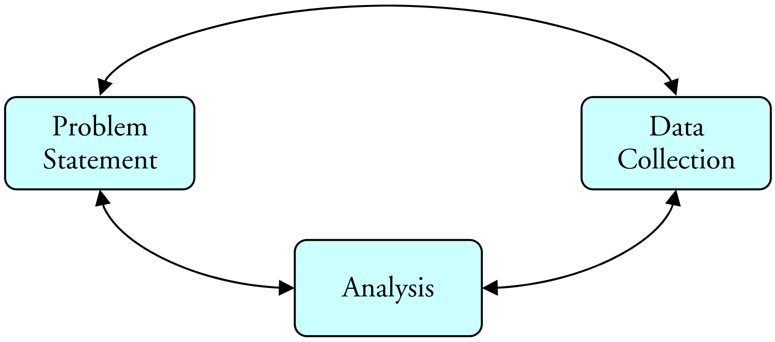


Figure 2.3 *The Problem Solving Process Using the Qualitative Analysis Approach*[42]

2.4.2 Analysis in this Thesis

As mentioned earlier, a quantitative research method will be used resulting in a rather large amount of graphs, tables and statistical values. As a consequence of this, the analysis will mainly be quantitative as well.

2.5 Credibility

This section will clarify the meaning of the terms validity, reliability and objectivity and their importance when writing a scientific report.

2.5.1 Reliability

Reliability is the question of whether a researcher would get the same results if the study would be repeated once more.[43] The concept of reliability measures the dependability and usability of a measuring instrument and the unit of measurement. The measuring instrument is often constructed by the scientist himself/herself, for example a questionnaire, which makes it feasible not to be entirely objective.[44] When using a questionnaire, the reliability can be tested in four ways;

- *Double testing* – All test subjects have to go through the test twice to make sure that the test results are reliable.

- *Divided testing* – The test answers are separated into two parts and compared to each other. A small divergence between the two parts indicates high reliability.

- *Parallel testing* – Two different questionnaires are used to measure the same thing. The reliability can be considered good if the tests reveal the same results.

- *Control questions* – Some of the questions in the questionnaire have the same essence, but are asked in different ways. The test has a high level of reliability if the answers to these questions are identical.[45]

[42] Lundahl, U & Skärvad, P-H (1999), p. 107
[43] Nielsen, J (1993), p. 165
[44] Ejvegård, R (1996), pp. 67-68
[45] *Ibid.*, pp. 68-69

In this thesis we will investigate the number of test participants that is required to get sufficiently reliable results. The aim is to find the peak where an increase of the test participants will not have a significant impact of the test result.

2.5.2 Validity

The term validity is used to ensure that what has been measured actually coincides with the goals of the measurement. One way of assuring a high level of validity is to have strict guidelines for how to measure. This is important for minimizing unnecessary fluctuations in the results. A low level of reliability always leads to low validity, whereas a high level of reliability is a necessity, but not a warrant, for obtaining a good validity.[46] The validity in this thesis is kept at a high level by video taping all tests and having a second look at them afterwards. All measurements are also done by both the authors and they were both present during each and every test.

2.5.3 Objectivity

It is always important to uphold a high level of objectivity when writing a scientific report. The meaning of this term is to try to manage and compile information without a preconceived notion. One way of solving the problem is to always name the reference where the information was gathered and manifest which parts are the authors' own opinions, assessments, interpretation and predictions. All literature should always be assessed for bias or propaganda.[47] It is our opinion that we have made our best to maintain a high level of objectivity throughout the study, although conducting tests on almost 50 participants will inevitably lead to some kind of prejudged opinions towards the end.

2.6 Thesis Procedure

After having an initial meeting with representatives from the Package Company where we were given our task, the first thing we did was to create a thesis statement which stated the purpose, aims and target groups of the thesis. This was followed by the creation of a preliminary timetable that was continuously updated throughout the work. We then began to perform literature studies regarding research methodology and usability engineering, as well as starting to write parts of this report.

The next phase of the thesis writing was to go through the existing test method and to come up with ideas on how to improve it. After doing that, we began the process of acquiring test participants and creating guidelines for the tests before conducting a pilot test as well as regular ones. After collecting all the required data from the test sessions (see more detailed information about the test data in Section 4.2.2), we started to work on the *dashboards* (see Glossary). Simultaneously, we analyzed the test results, discussed possible changes in the test method and the *dashboard* files, and continued the process of acquiring participants for subsequent tests.

The next step was to perform a thorough statistical analysis of the test results in order to see how many test participants are required to get stable results and to investigate the possible impact of the test monitor or the characteristics of the participants. After completing this part, the same test was conducted on elderly people in order to see if the test can be applied to this

[46] Ejvegård, R (1996), pp. 69-72
[47] *Ibid.*, pp. 17-18

consumer segment as well, and if the results start to stabilize in a similar manner to that of the students.

Finally, we made *dashboards* and statistical analysis for the elderly people test before concluding the optimal way to measure the products' performance and completing the writing of this thesis. The whole thesis process is illustrated in Figure 2.4.

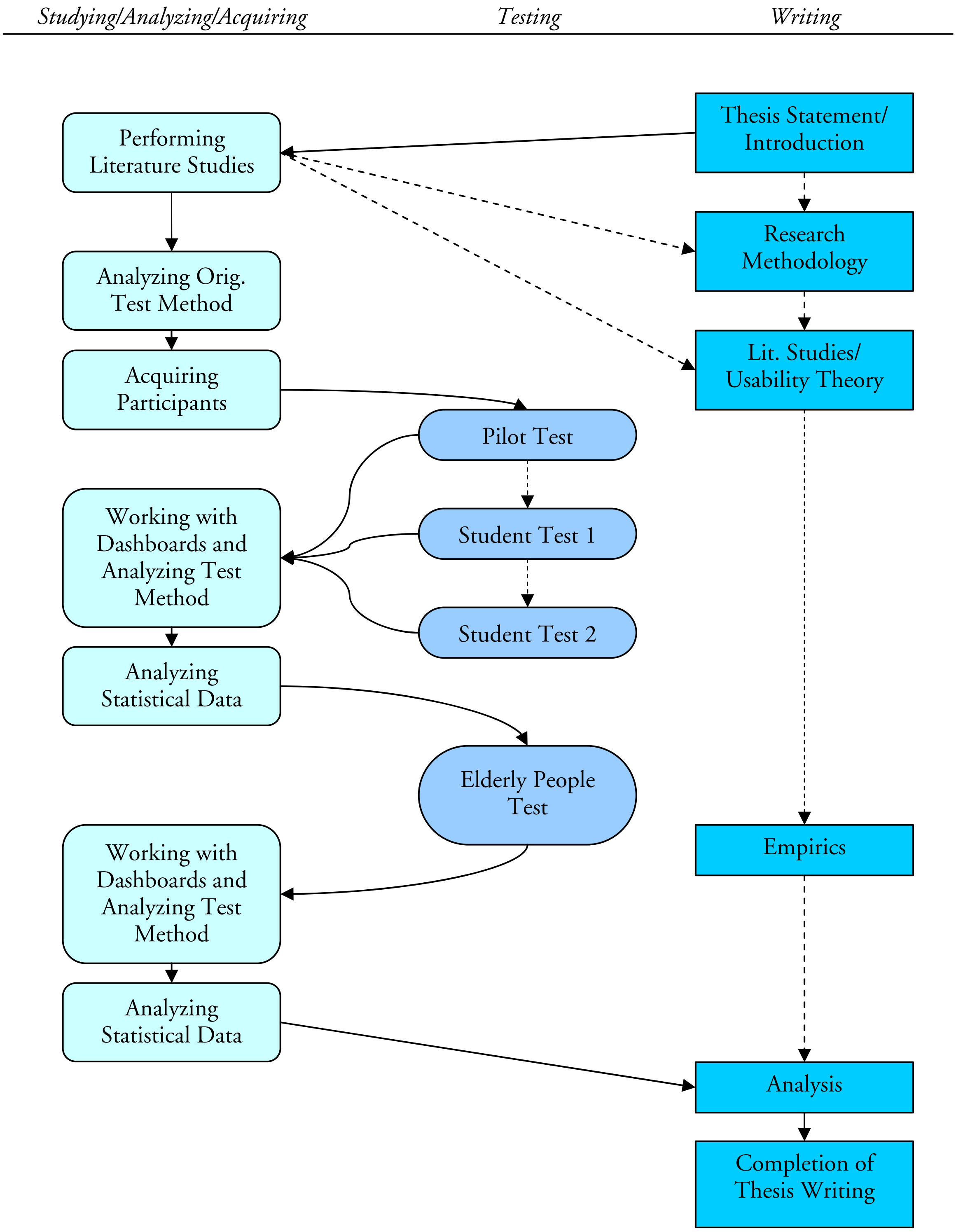

Studying/Analyzing/Acquiring
Testing
Writing
Performing Literature Studies
Thesis Statement/ Introduction
Research Methodology
Analyzing Orig. Test Method
Lit. Studies/ Usability Theory
Acquiring Participants
Pilot Test
Working with Dashboards and Analyzing Test Method
Student Test 1
Student Test 2
Analyzing Statistical Data
Elderly People Test
Empirics
Working with Dashboards and Analyzing Test Method
Analyzing Statistical Data
Analysis
Completion of Thesis Writing

Figure 2.4 *The Thesis Process*

3 Literature Studies

This chapter will provide a frame of reference to the concept of usability. A usability engineering walkthrough will be presented as well as the routine of usability testing. Further, the intention of this chapter is to provide the reader with a thorough understanding of the whole product development process, a prerequisite of comprehending why usability testing is undertaken.

3.1 Usability Engineering Methodology

The ISO 9241-11 definition of usability reads as follows.

> *"[Usability is the] extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."*[48]

Usability engineering is a way of defining goals and specifications for a product with the aim to establish an acceptable usability level through testing. According to Carlshamre, usability engineering can be divided into user profiling, task analysis, setting usability goals, making design decisions, generation of prototypes and evaluation, as seen in Figure 3.1. These steps will be explained further in the following subchapters.[49]

The last three steps in the model depicted in Figure 3.1, to be precise design decisions, generation of prototypes and evaluation, are known as rapid prototyping. Rapid prototyping is used when the user is unsure what the final product should be or look like. This cycle goes on until the product developers have found what they believe to be the optimal final product.[50]

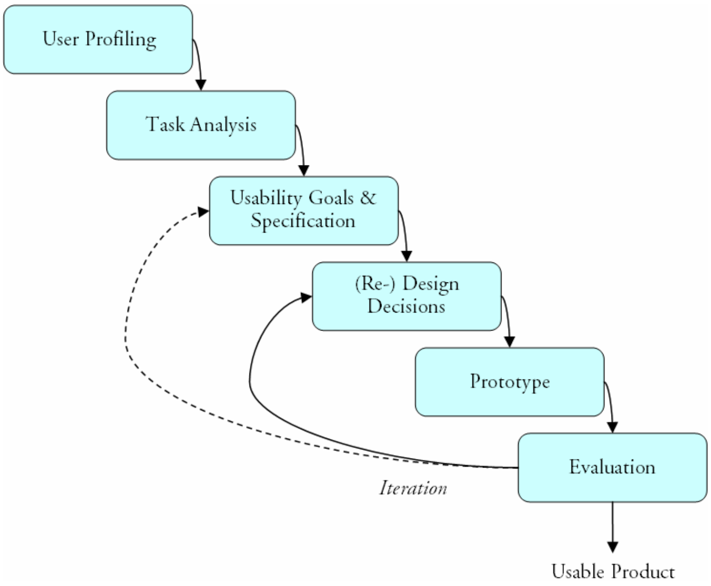


Figure 3.1 *The Usability Engineering Sequence of Work*[51]

[48] Carlshamre, P (2001), p. 33
[49] *Ibid.*, pp. 29-31
[50] Faulkner, X (2000), pp. 106-107
[51] Carlshamre, P (2001), p. 31

3.1.1 User Profiling

An important aspect of usability engineering is getting to know the user. The final usability of a product cannot ultimately be determined in a laboratory. Instead, it is up to the final consumer whether or not a product can be considered to have a high level of usability.[52] The process of user profiling is based on gathering information about the different users before dividing them into specific profiles, or *personas*. This makes it possible to ensure whether the product has a desirable level of usability for a specific group of users or not.[53] A checklist for gathering user information can contain the following information (it can obviously be adapted to contain information of a higher importance for the kind of product that is developed as well).

User information:

- *Age range*
- *Educational background*
- *Skills*
- *User classification*

Use of the product:

- *Discretionary or mandatory user*

Job details:

- *Brief job description*
- *Main tasks*
- *Responsibilities*
- *Control of work load*[54]

3.1.2 Task Analysis

Task analysis is used in order to understand what characteristics a product is supposed to have. This can be achieved by looking at the goals and then work out the processes involved in reaching that goal. Every task can be divided into subtasks, until a level that is required for solving the problem is reached. A task can be simplified into a model with input, output and a process which transforms the input to the output, see Figure 3.2.[55]

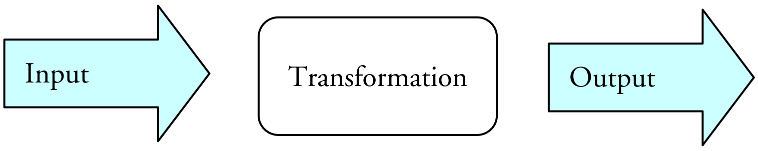


Figure 3.2 *Simplified Model of a Task*[56]

In order to break down the tasks into subtasks, there are some basic questions that need to be answered, according to Faulkner. When defining the inputs of a task, these are:

[52] Carlshamre, P (2001), pp. 30-31
[53] http://www.usability.uk.com/glossary/user-profiling.htm (2006-09-11)
[54] Faulkner, X (2000), p. 47
[55] *Ibid.*, pp. 63-64
[56] *Ibid.*, p. 64

- What information is needed to perform the task?
- What are the characteristics of the information sources?
- What is the availability of information?
- What possible errors might occur?
- Who or what initiates the task?

To define the outputs of the task, the following questions must be considered:

- What are the performance criteria?
- What happens to the output?
- How does the task performer get feedback about task performance?

The transformation between the input and the output will be highlighted by the following questions:

- What is the nature of the decision making?
- What strategies exist for decision making?
- What skills are needed?
- What interruptions are likely to occur and when?

The task composition of each job will also be considered:

- How often is the task done and when?
- Does the task depend on any other task?
- What is normal/abnormal workload?
- What control does the task performer have over the workload?[57]

All this information helps the usability engineer to develop a design that matches the user's mental model of the task. This makes the product more user friendly.[58]

3.1.3 Setting Usability Goals

According to the ISO definition of usability in the beginning of this chapter, usability goals should be reached with *effectiveness*, *efficiency* and *satisfaction*. These three words can also be transformed into the REAL model, which stands for *Relevance, Efficiency, Attitude and Learnability*.[59] The product can later on in the design process be compared to these goals to examine whether the product manages to fulfil the usability goals or not. The following subchapters will briefly explain the components of the REAL model.

3.1.3.1 Relevance

The *relevance* describes how well the product serves the user's needs. This is based on the extent to which the user utilizes all of the product's capacity and how big percentage of the tasks that are completed.

[57] Faulkner, X (2000), pp. 65-66

[58] http://www.usability.uk.com/glossary/task-analysis.htm (2006-09-11)

[59] Carlshamre, P (2001), pp. 33-35

3.1.3.2 Efficiency

The *efficiency* measures how efficient the user can carry out different tasks. Examples of things to measure are how many times an error occurs and the time spent for each and every task.

3.1.3.3 Attitude

The *attitude* is a subjective notion from the users of their feelings towards the product. A way of measuring these subjective feelings might be to register the number of times the test subject expresses frustration or satisfaction.

3.1.3.4 Learnability

The *learnability* of a product determines whether or not the product is easy to learn, and if the user remembers how to use it. Things to measure can be the number of repetitions of failed attempts and how frequent the user needs to use help or documentation.[60]

3.1.4 Design Decisions

Design must be seen as an iterative process and cannot be determined at once. For usability reasons, the design should primarily focus on the usability goals of the product as described in the preceding chapter. The design process can be illustrated by the following five core activities, specified in Figure 3.3 and briefly described below.[61]

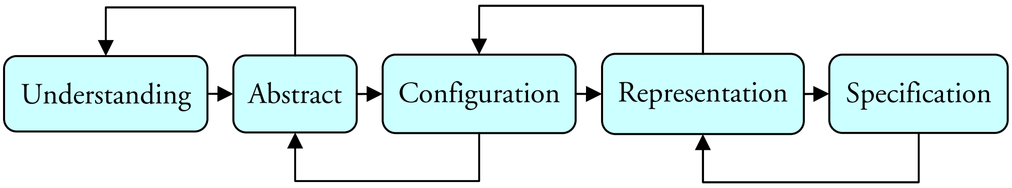


Figure 3.3 *The Design Process*[62]

- *Understanding* – Analyze the problem and figure out which subproblems need to be solved.

- *Abstract* – Sum up the main parts and determine what is most important and what is of lesser importance.

- *Configuration* – Determine how the different parts are linked together and how they can be configured for utmost user friendliness.

- *Representation* – Present the design with the aid of sketches and prototypes.

- *Specification* – Determine the final design.[63]

[60] Carlshamre, P (2001), p. 35
[61] Gulliksen, J & Göransson, B (2002), p. 238
[62] *Ibid.*, p. 239
[63] *Ibid.*, pp. 239-240

3.1.5 Prototyping

There are several ways of illustrating a design. It can be made as a computer simulation, a physical prototype or something in between. The prototypes can be used for the customer to examine the product or for the design team to evaluate the usability. Creating a prototype provides an opportunity to discover weaknesses, examine new solutions, determine demands and test the functionality of a product.[64]

Prototypes can be classified along two dimensions. The first dimension is to what extent the prototype is physical or analytical. The physical prototype is a tangible artifact in contrast to the analytical prototype, which often is a mathematical or computer-made model. The second dimension of prototype classification runs along an axis between being comprehensive and being focused. The comprehensive prototype implements most of the attributes of a product, as opposed to the focused prototype, which implements one, or a few, of the attributes. The relations between the above-mentioned dimensions and prototypes can be seen in Figure 3.4.[65]

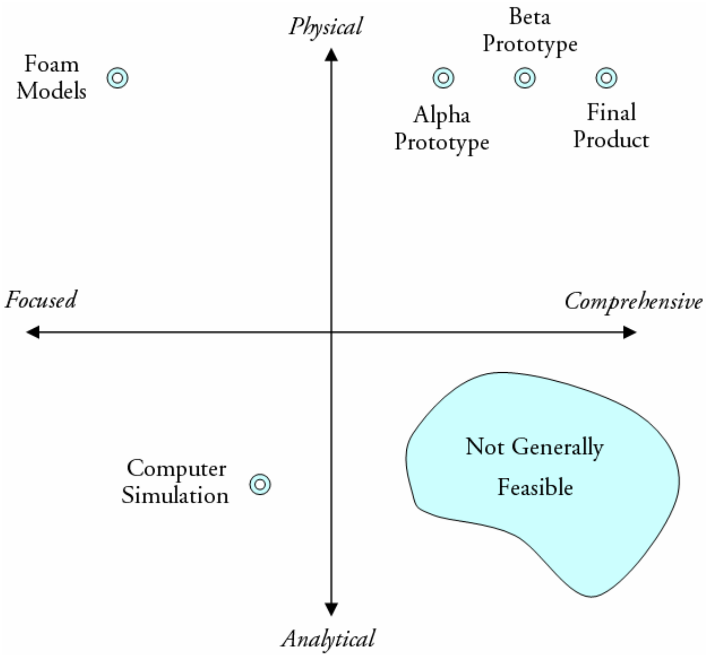


Figure 3.4 *Classification of Prototypes*[66]

3.1.6 Evaluation

There are several evaluation techniques to be used when trying to determine the usability of a product. Examples of such techniques are expert evaluations, benchmarking, workshops, questionnaires, field studies, usability lab evaluations, *scenarios*, *personas* et cetera. According to Gulliksen & Göransson, there are some basic guidelines which will explain how and when these methods should be used;

- Combining different evaluation methods is important in order to make sure that a majority of the usability problems are discovered. If only one method is used, there is a huge possibility that some of the problems go undetected.

[64] Gulliksen, J & Göransson, B (2002), p. 242
[65] Ulrich, K T & Eppinger, S D (2003), p. 247
[66] *Ibid.*, p. 249

- It is often better to use simple methods rather than complex ones in the early stages of the design process and, instead, use the time and money saved for finding alternative solutions.

- All evaluations should be documented.

- The final documentation of the findings should not be overdone. Keeping it simple will make sure that no evaluation time will be lost for unnecessary activities.

- Trying to maintain a positive attitude in the documentation of the findings will ensure that the developers actually respect the results. It is not unusual that developers see the usability evaluation as negative criticism against their design.

- The usability evaluation can sometimes be asked for by the developers only to get a receipt that the product functions satisfactory. In order to guarantee that the results will be used for the intended purpose of ensuring usability, it is important to document and report the actual findings.

- Informing the users that have been involved in the evaluation process about the findings is recommended so that they do not feel that their contribution has been in vain.[67]

3.2 Usability Testing

This subchapter regards different methods and important aspects when evaluating the usability of a product. Information about how to set up plans and goals, acquire test participants and choosing *test monitors* will be presented. Moreover, several different test types will be described as well as a detailed description of the usability testing procedure will be given.

3.2.1 Test Goals and Plans

Writing a proper test plan is an initial and crucial part of the test as it describes how to conduct the test, when and where to do it and how to choose the test participants. Due to time pressure, it is possible that tests are made without having a detailed test plan. This may, however, turn out to be a great mistake as a comprehensive test plan serves as a blueprint for the test, telling exactly what should be done and what should not. It can also be seen as the main communication tool within the development team, including the main developer, the test monitor and the rest of the staff. Following the fact that everybody in the development team has to review the test plan continuously, it is a good way to collect feedback and determine what resources are necessary to be able to accomplish the test. Further, it helps the test monitor to work in a systematic manner.[68]

Even though the test plan is most likely to be revised to some extent during the test process, it is important to write it down before commencing with the tests. Among the many issues that could be addressed in the test plan there are several very important ones, such as what the purpose of the test is and problem statements, what characteristics the test users are supposed to have, how many test users are needed and what test tasks have to be done. Other issues to

[67] Gulliksen, J & Göransson, B (2002), pp. 261-264
[68] Rubin, J (1994), pp. 81-83

be addressed are definitions of when the test participants have finished a task (it is not always obvious), how long the test is estimated to take, who will be the test monitor, what kind of assistance will be available to the test participant and in what kind of environment the test will take place. The test plan should, in addition to the above-mentioned issues, also state what data is to be collected by the test team and define the successful completion criteria.[69] The following subchapters will describe the different sections of a test plan.

3.2.1.1 Purpose

The purpose of the test does not have to be described at a too specific level. The more detailed description of the problem should be left for the problem statement section and instead, a shorter description from an overall point of view can be presented in this section.

3.2.1.2 Problem Statement

As this section describes the problems and issues that need to be solved, it may very well be considered the most important one in the entire test plan. The better the problem definition is regarding precision, accuracy, clearness and measurability, the better the whole test plan will be. Without a proper test plan, there is a risk of conducting time-consuming and expensive tests and still not get any answers or explanations to the key concerns of the developers.

Incomplete and vague problem statements such as "*Does the product work?*" do not make any sense as they do not say what to measure or how to do it. Besides, a problem statement like the one just mentioned is more likely to bias the research and the results favorably. A good problem statement will definitely either state whether something has been accomplished or not or what the underlying reasons to a specific issue are.

3.2.1.3 User Profile

Specifying the characteristics of the product's target group is a necessity in order to be able to test it on an adequate group of people, which in turn is required to be able to produce a useful product. More information about user profiling was earlier given in Section 3.1.1.

3.2.1.4 Method

The method section is supposed to describe how the test session itself is going to be performed. It has to cover the entire session, from when the test participants arrive to the point they leave and be rather detailed. A detailed description is necessary in getting other to understand what the test is all about and to come up with ideas and comments. Additionally, it will give emphasis to the need of communication with possibly forgotten resources within the development team and it is more or less a requirement for using several test monitors.

3.2.1.5 Task List

This section of the test plan contains descriptions of the tasks the participants will perform during the test. This includes a simple explanation of the task itself, a list of the materials and machines required and what state they are supposed to be in, a description of what exactly is meant by successful completion of the tasks and, if possible, the average time and maximum time limit to accomplish each task.

3.2.1.6 Test Environment and Equipment

This section aims at describing the environment where the test will take place and the equipment that will be used by the participants. It is important to simulate the conditions

[69] Nielsen, J (1993), pp. 170-171

where the product will be used later on in order to get the participants to behave like the end users and to get a better prediction of the outcome.

3.2.1.7 Test Monitor Role

This section of the test plan describes what the test monitor will do and under what circumstances intervention like *probing* or role playing is likely to occur. This is of importance if there will be persons present that are not used to the testing process. Information regarding the amount of help and on what occasions this help will be given to the participant is also a part of this section.

3.2.1.8 Data to be Collected

This section summarizes what data is to be collected during the test, as well as after it is finished. There should be a connection between the problem statement and the collected data, which can be divided into two groups, performance and preference data. The former measures participant behavior, mainly through numerical data like the number of errors and the time to accomplish tasks. Preference data, on the other hand, measures the opinion of the participant, including questionnaire answers and expressed opinions. Generally performance data would be considered as equivalent to quantitative data and, consequently, the same relationship would exist between preference and qualitative data. However, both performance and preference data may be used both quantitatively and qualitatively, depending on the objectives of the test.[70]

3.2.2 Acquiring Test Participants

The single most important rule when it comes to acquiring test participants is to find such with a behavior similar to that of the end users. If the test, for some reason, is conducted using a small amount of participants, these should mainly represent average users. A wider range of end user characteristics should only be represented if there is a possibility to use a larger amount of participants. It is recommended to avoid using sales people and internal staff members as test participants because of the risk of getting misleading results as these people most often do not wish to criticize the work of their own company and fellow colleagues.

3.2.2.1 User Categorization

An important user categorization is novice and experts users. Tests on products with target users spread all over the experience and knowledge ranges have to be thoroughly done as the difference in characteristics between novice and expert users can be enormous.[71] Among the novice users it can be very insightful to include one or a couple of *least competent users*. This term means end users representing the absolute lowest degree of knowledge. This might seem pointless at a glance but will provide a deeper understanding to the development team. If the *least competent user* can manage a task, practically everybody will and if the *least competent user* does not manage it, there is still a lot to learn regarding the end users' conceptual model of the product and what the major problems of the product are.[72]

3.2.2.2 Sufficient Number of Test Participants

The number of test participants is dependent on several factors. Some of these are the required reliability of the test results, the available resources for the conduct of the test, the availability of participants, the length of the preparations and the equivalent of test session itself.

[70] Rubin, J (1994), pp. 83-106
[71] Nielsen, J (1993), pp. 175-177
[72] Rubin, J (1994), p. 129

If the purpose of the test is to find as many usability problems as possible it should be enough to use just a few participants, but if statistically valid results are required, which is the case in this report, it is most likely that a considerable increase in the number of participants will be necessary.[73]

3.2.3 Test Designs

Two popular test designs are the independent groups design and the within-subjects design. When using the first one, every part of the product is tested by separate user groups. This reduces the so called *transfer of learning effect* that is caused by doing certain tasks after having done other very similar tasks and, as a result, having more knowledge on how to act.

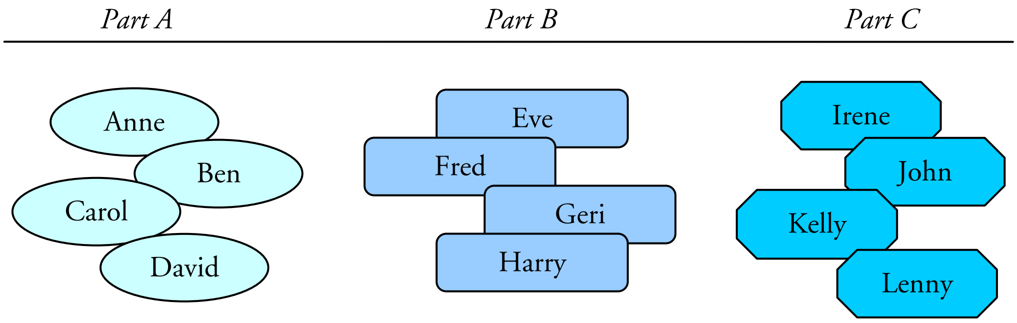


Figure 3.5 *Independent Groups Design*[74]

As seen in Figure 3.5 above, there will be no *transfer of learning effect*, as the different parts are tested by different participants. However, in this particular example, the independent groups design requires not less than 12 participants and sometimes there is a need to keep the number of participants to a minimum. Under such circumstances, the within-subjects design may be favorable to use, as it implies that one group of participants will test all the parts in turns. Applying this to the above-mentioned example, just four test participants are needed. This will of course lead to *transfer of learning effects*, though they may be omitted by using a technique called counterbalancing. In such a case the test participants perform the tasks in different sequences. This is visualized in Figure 3.6 where the parts are tested in three turns.

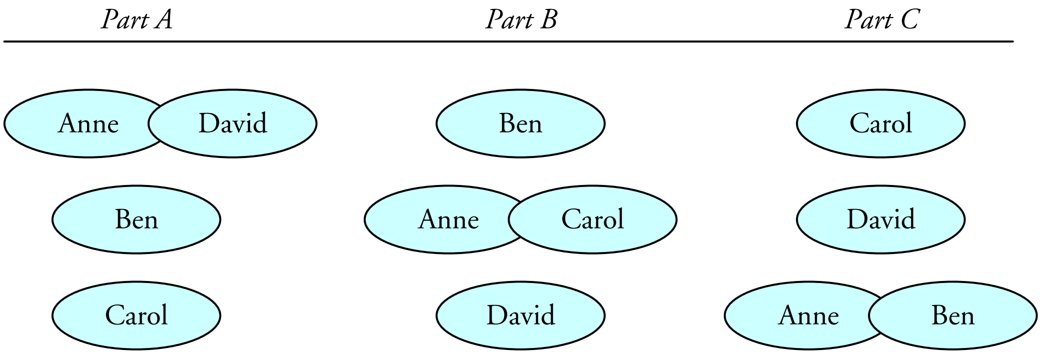


Figure 3.6 *Within-Subjects Design Using Counterbalancing*[75]

[73] Rubin, J (1994), p. 128
[74] *Ibid.*, p. 89
[75] *Ibid.*, p. 91

A weakness of the within-subjects design is that the sequence of performance may be unnatural to the user and *transfer of learning effects* may actually be desired in some cases.

Besides these two test designs, it is also common to use 2x2 matrices when comparing two different products on two different user groups simultaneously or if there are two sets of user group categorizations. Two examples of such categorizations are level of experience and gender. Both the independent groups design and the counterbalancing technique are applicable on 2x2 matrices.

3.2.4 Choosing Test Monitors

Regardless what method is used to conduct the test, somebody has to be the leader who runs the whole test, the *test monitor*. This is the person of the development team that interacts with the test participant by giving instructions and assistance, handing out questionnaires and performing the debriefing session. There are some personal characteristics that the *test monitor* should possess, including;

- Good knowledge about usability engineering in general
- Good knowledge about the test method used in particular
- The ability to learn fast
- Making test participants feel comfortable
- Have a good memory
- Be a good listener
- Have tolerance for ambiguity
- Be unafraid of deviating from the test plan
- Have a lot of patience
- Show empathy
- Be a good communicator
- Be a good organizer

There are however some pitfalls for the *test monitor*, such as;

- Behaving leading and unintentionally giving clues to the participant
- Being too busy collecting data to observe what is happening
- Acting too knowledgeable, causing the participants to ask for help all the time instead of thinking themselves
- Being too rigid with the test plan
- Not relating too well to the participants
- Jumping to conclusions too fast and overlook what takes place in later tests due to preconceived notions[76]

3.2.5 Test Types

There are different types of tests and which one is to be used is dependent of the present stage in the product development life cycle. Figure 3.7 shows when the different test types are applicable.

[76] Rubin, J (1994), pp. 67-73

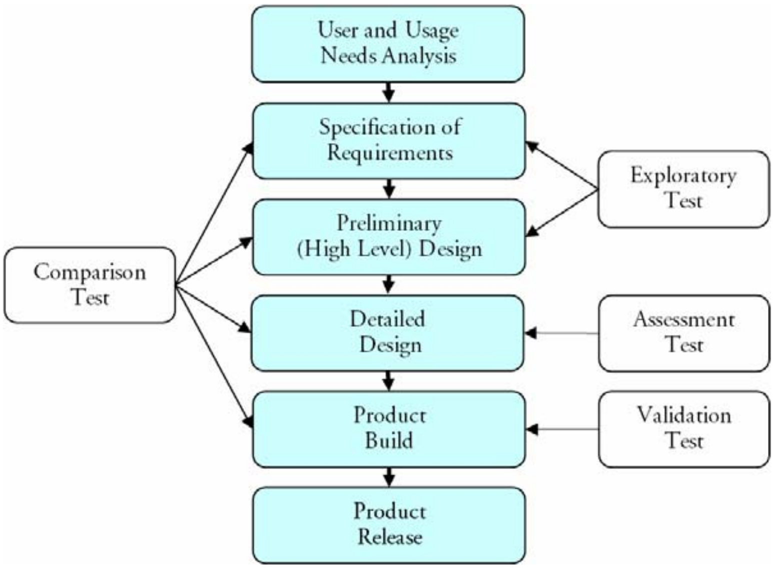


Figure 3.7 *Product Development Life Cycle*[77]

It is worth to emphasize that these test types are not to be regarded as substitutes, but complementary to one another and should, thus, all be used during the development life cycle.

3.2.5.1 Exploratory Test

The exploratory test is used at an early stage in the development process, while constituting the specification of requirements or early in the design phase. The purpose of the exploratory test is to evaluate how the users' conceptual model of the product looks like. This kind of testing is important due to the fact that if the project begins wrong, it will most likely continue to be wrong throughout the life cycle. Exploratory tests usually have a lot of interaction between the *test monitor* and the participant and they are often carried out using prototypes or mockups. The task that the participant has to do is often to just use the product randomly and meanwhile make comments about what is happening. The question that exploratory testing seeks to answer is not regarding how well the test participant is performing, but rather why the participant is behaving like he or she does.

3.2.5.2 Assessment Test

Assessment tests are normally conducted during the design phase and are perhaps the most common usability test of the four types described. The objective is to localize imperfections in the design and the tests are normally conducted on partially or completely functional prototypes. There is less interaction between the *test monitor* and the participant compared to the exploratory test, but on the other hand there is more emphasis on the participant's behavior. The participant will rather perform specific tasks than just walking through and randomly explore the product. The data collection mainly regards quantitative measures.

3.2.5.3 Validation Test

Validation tests are normally used late in the development life cycle, close to the release of the product. The intention of such tests is to make sure that the product's usability meets the requirements. This kind of validation is usually done by comparing a product partially or completely ready for release with some kind of standard or benchmark. There is practically no

[77] Rubin, J (1994), p. 32

interaction between the *test monitor* and the participant, or at least it is kept to a minimum. A validation test needs more rigorous and consistent experiments than assessment tests.

3.2.5.4 Comparison Test

Unlike the three above-mentioned tests, this type is not connected to a development life cycle phase in the same way. On the contrary, it can be conducted practically whenever during the development process. Comparison tests can be used in combination with any of the other test types and the objective is to compare different design alternatives in order to find out which one is the optimal choice.[78]

3.2.6 Usability Laboratories

Usability tests are usually conducted in laboratories specially dedicated for the purpose. Figure 3.8 shows how such a laboratory may look like, although some additional components can be used as well. The depicted laboratory design is to be regarded as a very moderate one.

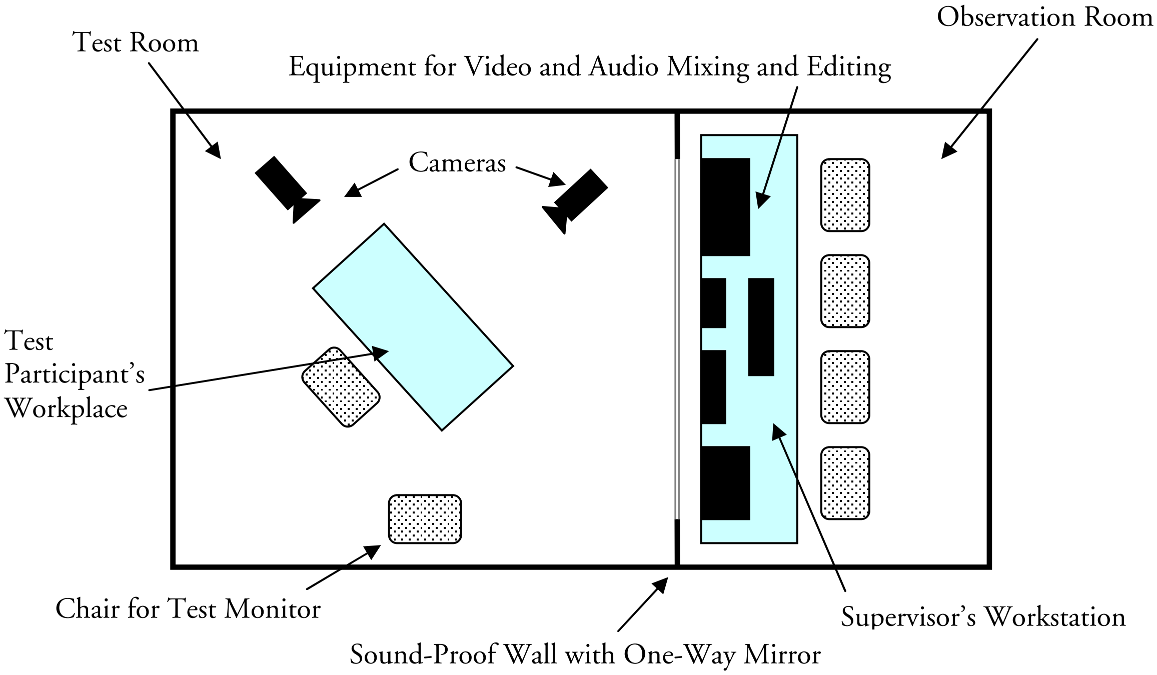

Test Room
Equipment for Video and Audio Mixing and Editing
Observation Room
Cameras
Test Participant's Workplace
Chair for Test Monitor
Sound-Proof Wall with One-Way Mirror
Supervisor's Workstation

Figure 3.8 *Design of a Typical Usability Laboratory*[79]

An essential part of this kind of laboratory is the one-way mirror that separates the test room from the observation room. As it is sound-proof, the test participant will neither see nor hear the development team and usability specialists in the observation room. The main reason for using a mirror like this is that participants tend to ignore unseen observers and, thus, act in a more natural way. Other important parts of the equipment are the video cameras and the microphones. These are necessary to record the test properly and are remote-controlled from the observation room. It is recommended to use several cameras to be able to focus on several things simultaneously, like the test participant's face, the product and an overall view of the whole test.

[78] Rubin, J (1994), pp. 30-45

[79] Nielsen, J (1993), p. 201

In some cases it may however be unnecessary to record the test session as it takes approximately 3 to 10 times the duration of the test itself to analyze the recorded material. If the purpose of the test is to discover the major problems, recording may be left out, but the benefits of video recording are in general larger than those of avoiding it. If the developers are hard to persuade in some specific case, showing a recording of a test participant struggling with the problem may very well do the trick.[80]

3.2.7 Test Tasks

The main rule regarding test tasks is to choose them to be as representative as possible to the real future tasks. Further, the chosen tasks should cover as many important parts of the product as possible and be neither too trivial nor too difficult and time-consuming. It is also important that the test tasks accurately specify the results that the test participant is expected to achieve.

A test session is not the time for fooling around. The part ordering the test wants results, which implies that the participant has to be focused and take the test seriously. This can be facilitated by handing over the tasks in writing. In addition, they should be written in a formal, realistic and serious way, like a *scenario* for example.[81]

3.2.8 Test Stages

A usability test consists of more than the test itself. The different stages of the test process will be explained in this section.

3.2.8.1 Preparation

As a part of the preparations for a usability test, the test monitor should develop a vast amount of materials, such as background questionnaires, *non-disclosure agreements* and *tape consent forms*, pre-test and post-test questionnaires, orientation scripts, task scenarios and debriefing guidelines.[82] Additionally, the *test monitor* is responsible for preparing the test room and making sure it is ready to use, including all things in it, like computers, video cameras, microphones et cetera.[83]

Other parts included in the preparations are tests of the test itself and revisions of the product. The former includes the test team taking the test themselves in order to find design problems in it, as well as conducting a pilot test. This kind of test is exactly like a real test, though it is conducted with just one or a few users and the intention is to get the bugs out of the test procedure and improve potentially problematic parts of it. The pilot test may also help the test team to identify areas of the tested product so problematic that they have to be fixed even before conducting the real test.[84]

3.2.8.2 Introduction

The introduction part of the test includes the *test monitor* welcoming the participant and explaining what is about to happen. Optionally, the *test monitor* may describe the product setup, but the following statements are usually said during a test introduction:

[80] Nielsen, J (1993), pp. 200-204
[81] *Ibid.*, pp. 185-186
[82] Rubin, J (1994), p. 142
[83] Nielsen, J (1993), p. 187
[84] Rubin, J (1994), p. 225

- The aim of the test is to evaluate the product and not the participant.

- The *test monitor* has no personal interest in the result, in order to avoid that the participant will be afraid of hurting the *test monitor's* feelings.

- The test is confidential. This may be said even if it is not true, as to prevent the participant to discuss the test with future test participants.

- Taking the test is voluntary and the participant may choose to quit at any time for any reason.

- Information about video and audio recordings if such take place.

Following the introduction the *test monitor* gives written instructions to the participant, including the tasks to be performed. If the participant does not have any questions, the test itself may commence.[85]

3.2.8.3 The Test Itself

During the test, the *test monitor* should keep a low profile and try not to interact with the participant more than necessary. It is important that the participant does not notice any preferences from the *monitor*, who therefore should not openly show any reactions regarding the actions of the participant. Further, the *test monitor* must be careful about the voice and body language, as it is easy to unintentionally bias the test participant. Another important issue is to treat every test participant as a unique individual. With too short breaks between the tests, it is most likely that the *monitor* will be affected by the results of the earlier participants.

Test monitors tend to help out participants when they encounter difficulties. A better thing to do in such a case is to encourage the participant to express their problems and feelings about them. Besides, to see the participant struggle with a problem is a good way to get an understanding of what has to be improved.

Although direct observations of the test participant, combined with audio and video recordings and different kinds of questionnaires, are a good way to gather information on how the participant is thinking, it may not be enough. Under such circumstances, the thinking aloud method can be useful. This is a straightforward method to capture the thoughts of the participants by asking them to vocally express what they are thinking of constantly throughout the test session. In case the *test monitor* still does not get the answers needed, it might be useful to use *probing*, though this method should be adopted infrequently.[86]

3.2.8.4 Debriefing

After finishing the test, the participants are asked to fill out some questionnaires regarding it, as the comments may become biased if discussions take place before this step. During the debriefing session, the participant gets the possibility to explain things that the *test monitor* was not able to see or present suggestions for future revisions of the product. The debriefing

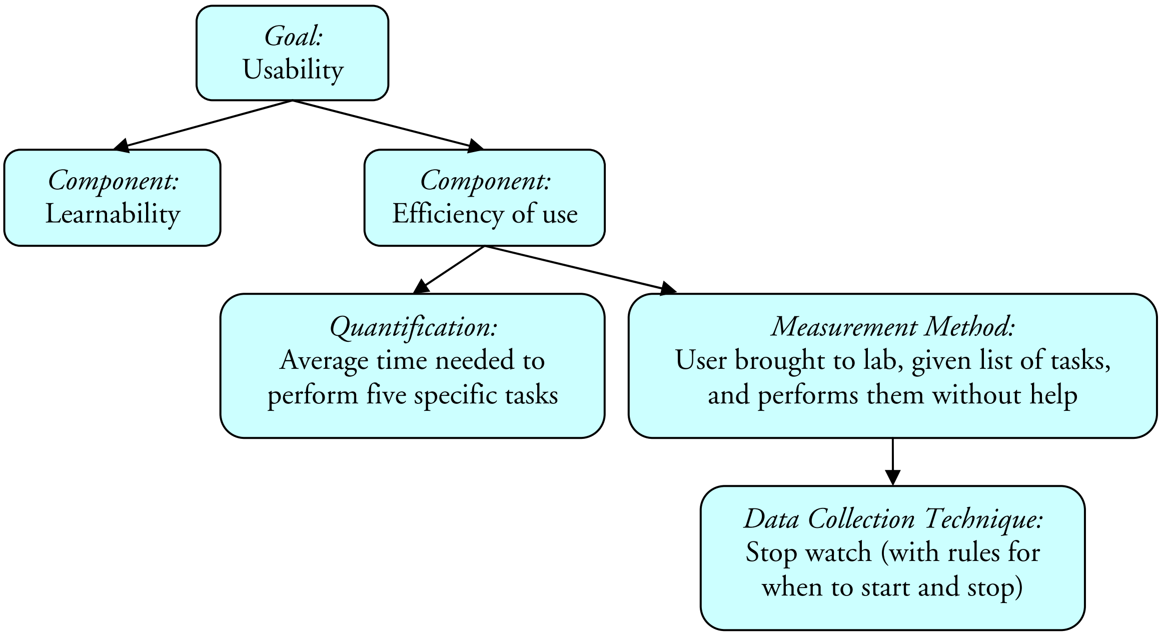
[85] Nielsen, J (1993), pp. 188-190

[86] Rubin, J (1994), pp. 215-219

is also a good occasion to clarify what the intention of the test participant was in cases of controversial or strange behavior.[87]

3.2.9 Performance Measurements

Performance measurement studies are important tools to see if a product meets the usability goals and to compare the tested product with its competitors. The most common way to measure the users' performance is to let a number of test users perform certain tasks and then gather data like how long time the tasks took to accomplish and how many errors were made.



Goal: Usability
Component: Learnability
Component: Efficiency of use
Quantification: Average time needed to perform five specific tasks
Measurement Method: User brought to lab, given list of tasks, and performs them without help
Data Collection Technique: Stop watch (with rules for when to start and stop)

Figure 3.9 *Model of Usability Measurement*[88]

In the case illustrated by Figure 3.9, the goal is to have a product with good usability. This may seem a bit abstract and vague, but goals are often rather abstract. As a consequence, they are often broken down into subcomponents on several levels. In the illustrated example, the goal is broken down into two components, namely *learnability* and *efficiency*. The component *efficiency of use* needs, however, to be quantified. An example of how to quantify it may be the average time it takes to perform, for instance, five specific tasks. When the definition of what is to be quantified is made, there is still a need to define a method to measure the performance. Possible methods in the described case could be either to bring test participants to a usability laboratory and perform a test there or to observe them in their natural environment. Finally, there is a need to define in what way the measurement is going to be conducted. A possible option in the current case is to measure the average time manually with a stop watch.

[87] Nielsen, J (1993), p. 191
[88] *Ibid.*, p. 192

Some examples of typical quantifiable usability measures:

- The time needed to perform a certain task completely
- The number of committed errors
- The number of positive/negative emotions expressed[89]

3.2.10 Analysis and Transforming Data into Conclusions

The first thing that needs to be done is to summarize the raw data so it becomes more manageable. When this step is completed, the data has to be analyzed and a good start is to determine which tasks were at a reasonable level of difficulty to the participants. By setting a percentage that has to be reached, it is easy to see if a task should be taken into consideration when evaluating the product. According to Rubin, 70% is a reasonable criterion and this implies that if at least 70% of the test participants completed the task successfully and within the given time limit, the task should not be seen as problematic, and thus applicable for analyzing. After sorting out the adequate tasks, user errors and difficulties should be identified as well. Following this step, a source of error analysis should be conducted in order to find out what really caused the difficulties discovered during the test. The next step would then be to prioritize the problems by criticality, which is a measure considering how severe the problem is combined with the likeliness that it will occur. The reason for this is to focus on the problems with most impact on the future product. It may also be good to analyze differences between the results for different users groups or product versions.

The last step in the usability testing process is the most important one, namely to transform the now summarized raw data into conclusions and recommendations for future actions. An important thing is to focus on the solutions with the largest impact on the end product. Other things to think of when making recommendations are ignoring considerations about whether the solution is doable or not and providing both long-term and short-term recommendations. It is likewise of great importance to present possible fields for future studies. The recommendations should be made by a group of people and it is advisable to wait a couple of days after finishing the testing before drawing conclusions and present recommendations.[90]

3.2.11 Limitations of Testing

Even high usability test scores cannot with 100 percent certainty ensure that a product has a high usability when released onto the market. Some of the reasons to this are:

- A usability test is an artificial situation which tends to make the participants more aware of what is expected of them. This may cause them to act in an unnatural way during the test.
- The test results are dependent on how the test was conducted and are not an absolute truth. Even if the results are statistically significant, it does not prove that the product works, it merely states that that the results were not due to chance.

[89] Nielsen, J (1993), pp. 192-194

[90] Rubin, J (1994), pp. 274-288

- The participants of a test are rarely fully representative of the target population. The development team has to define the correct target population and then acquire a representative selection of test participants to be available for the tests.[91]

3.3 Data Processing

The data gathered in connection with the usability tests is not only used to calculate different usability measures, but also to determine how many test participants are required to get stable results. The following subchapters will briefly explain the underlying mathematical and statistical calculations that will be used in this report.

3.3.1 Chart Exploration

The chart exploration method is perhaps not the most advanced one there is, but it is definitely easy to perform and it gives a good view of how the test results vary depending on the number of participants used in the test. The graph below shows an example of this for the *satisfaction* measure and it is constructed by starting with the total number of participants, in this case 35, and determining the cumulative average of the *satisfaction* value. The equivalent value for 34 participants is then simply obtained by removing the input data for the 35th one. This procedure is subsequently repeated until only one participant is remaining. All the cumulative average values are then plotted in a chart like the one below.

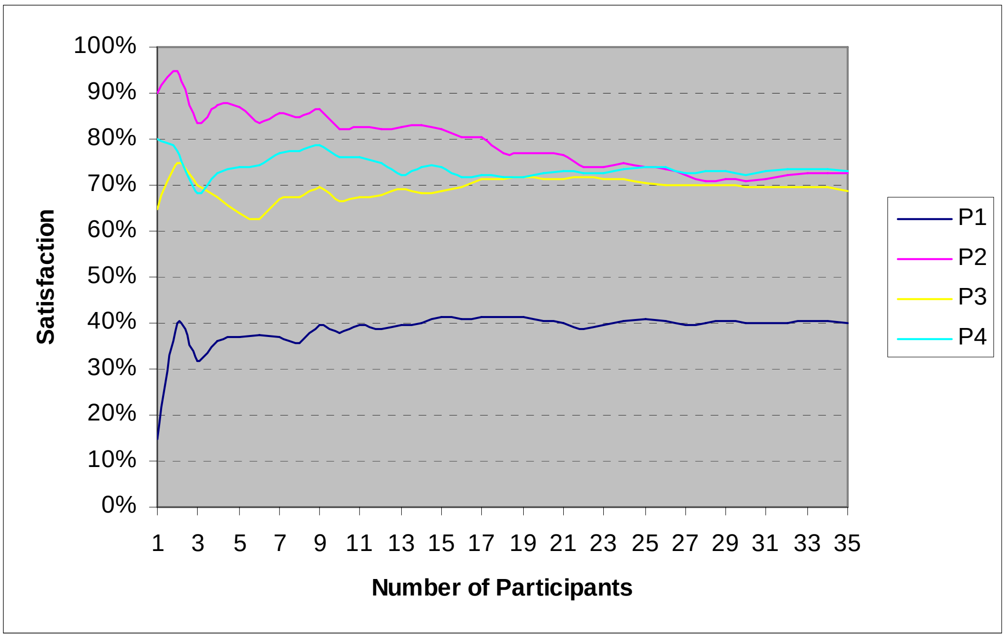


Figure 3.10 *An Example of How the Test Results Vary with the Number of Participants*

As seen in Figure 3.10, it is fairly obvious that the results stabilize rather quickly and get more or less totally stable for all packages when using around 25 participants, as none of the curves hardly oscillate any more when adding more participants from that moment on.

[91] Rubin, J (1994), p. 27

3.3.2 Statistical Calculation

As already faintly outlined, there are more sophisticated methods than chart exploration. Calculating the standard deviation of the test data and use it to determine the minimum amount of test participants required to get stable results within a certain confidence interval is a possible option.

There are two formulas for determining the standard deviation; the so called unbiased method (also known as the “n-1” method) and the biased method (also called the “n” method). The former only needs a sample from the total population, while the data for the latter represents the whole population, although it can still be used if only the number of samples is fairly large (at least 30 samples). The formula for the unbiased method looks like the following:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

where

$\sigma =$ standard deviation
$n =$ sample size (amount of participants)
$x =$ observation data[92]

It is now easy to calculate the required amount of test participants using the following formula:

$$2 \cdot \lambda_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = \delta$$

which is transformed into

$$n = \left(2 \cdot \lambda_{\alpha/2} \cdot \frac{\sigma}{\delta}\right)^2$$

where

$\lambda_{\alpha/2} =$ $\alpha/2$ quantile
$\alpha =$ confidence coefficient (= 1 – level of confidence)
$\delta =$ length of confidence interval[93]

In the underlying statistical material for Figure 3.10, the P2 package had a standard deviation of 6.37 percentage units. If we would like to know the required amount of participants based on that figure, all we have to do is set a level of confidence (for instance 95%) and a length of the confidence interval (for example 5 – this means the result may differ 2.5 percentage units). The $\alpha/2$ quantile for a 95% confidence interval is 1.96 which gives us[94]:

[92] Blom, G & Holmquist, B (1998), p. 34
[93] Wiktorsson, M (2006-11-29)
[94] Blom, G & Holmquist, B (1998), p. 363

$$n = \left(2 \cdot 1.96 \cdot \frac{6.37}{5}\right)^2 = 24.9$$

As seen in Figure 3.10, this value coincides quite well with the spot where the curve representing the P2 package turns stable and the recommended amount of required participants that this entails.

The results calculated this way are more statistically correct than those of chart exploration, but may on the other hand recommend using several hundred participants. The reason for this is that all the values are treated equally when calculating the standard deviation. As a consequence, large volatility in the cumulative average when using just a few participants may affect the recommended amount of test participants unfavorably, even if the results stabilize very quickly.

3.3.3 Wilcoxon Signed Rank Sum Test

One way to see if odd results and large standard deviations are a result of a few values differing very much from the rest is to perform a Wilcoxon test. This kind of test is normally used to see whether there is symmetry around a value or not. As there is no “correct” value for the *accuracy*, *efficiency* and the like, the cumulative mean for all participants will be used as it is the best value available. There will most likely be no asymmetry around the mean value for obvious reasons, but if it is fairly close, then it can be seen as a warning sign, telling us that the large standard deviation is not caused by extreme volatility in the test results.

The idea of the Wilcoxon Signed Rank Sum Test is to test a null hypothesis of symmetry around a certain value. If we, for instance, have the following data

158 125 156 157 141 155 201 140 117 96 134 136 126 190 169

it may be interesting to know whether or not the distribution is symmetrical around 150. The following null hypothesis is therefore stated:

H_0: the distribution is symmetrical around 150.

The next step is to calculate the difference between each of the observed values and the hypothesized value; $d_i = x_i - M$, where d is the difference, x stands for the observed value and M denotes the hypothesized value 150. The following differences are generated:

8 -25 6 7 -9 5 51 -10 -33 -54 -16 -14 -24 40 19

Now the absolute values of these differences have to be ranked, with the smallest $|\, d_i \,|$ being assigned rank 1 and the largest $|\, d_i \,|$ out of the n differences will be assigned rank n.

Following this, each rank is labeled with its sign, according to the sign of d_i, before calculating W^+, which is the sum of the ranks of the positive differences and W^-, which is the sum of the ranks of the negative ones. The differences above will correspond to the following ranks, with negative ranks in italics:

4 *11* 2 3 *5* 1 14 *6* *12* *15* *8* *7* *10* 13 9

If W^+ is larger than W^-, the values larger than *M* deviate more from *M* compared to the values smaller than *M*. [95] We are interested in the values deviating the most from *M*, which means that *W* is

$$W = \max\left(W^+, W^-\right)$$

If H_0 is true and n is large, we can approximately say that $W \in N(m, \sigma)$. To test the hypothesis, these formulas are used:

$$m = \frac{n(n+1)}{4}$$

$$\sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$u = \frac{(W - m)}{\sigma}$$

The absolute *u* value is then compared to the $\alpha/2$ quantile. Using a level of confidence of 95%, the hypothesis of symmetry around *M* cannot be rejected if | *u* | is lower than $\lambda_{0.025}$.

In the example above, W^+ is 33 and W^- is 87, making *W* equal to the latter. Following the fact that *n* is 15, it is easily seen that *m* is 60, *σ* is 17.61 and, finally, *u* is 1.53. This value is lower than 1.96, which indicates that the hypothesis of symmetry around 150 cannot be rejected, although it is fairly close. If applying this to the statistical calculations, we could have neglected recommendations of immense amounts of participants if having such a large *u* value.[96]

[95] http://mlsc.lboro.ac.uk/documents/wsrt.pdf (2006-12-04)

[96] Blom, G & Holmquist, B (1998), pp. 256-257

4 Empirics

This chapter will provide a description of the present package usability test method, as it was designed before we made any changes. Furthermore, a detailed explanation on how the usability tests were conducted will be given, followed by an enumeration of what kind of data was provided or withdrawn from the usability tests and how this was done.

4.1 The Present Method

The test method is based on the ISO 9241-11 definition of usability, taking *effectiveness*, *efficiency* and *attitude* into consideration. These values are added together with the weights of 40-40-20 to receive an overall competitor benchmark for the package in question, as seen in Figure 4.1.[97] The percentages illustrate how good the package is relative to a perfect package. What is considered a perfect package is depending on how the different values are calculated.

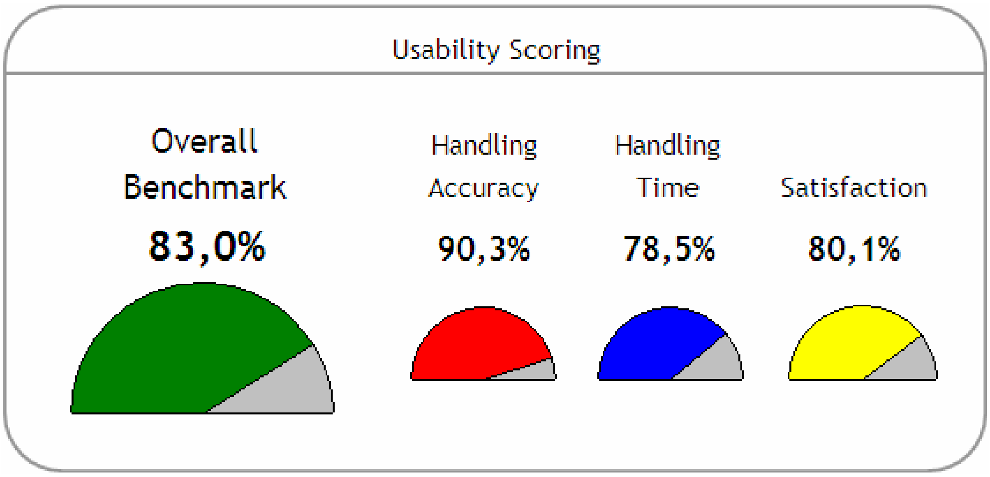


Figure 4.1 *Presentation of the Different Usability Scores*[98]

4.1.1 Accuracy

Accuracy, also known as relevance or *effectiveness*, measures how well the product serves the users' needs. In the present test method, a package's *accuracy* is determined by the pouring accuracy and the incidence of spillage.

When measuring the *accuracy* in the part of the test that includes drinking (see Section 4.2.1 for more information on the tests), the *test monitor* looks for any spillage and answers the question with a yes or no. This provides two different values for every test participant, one for drinking while sitting and one for drinking while walking. These values are then added and an all in all result is generated for every participant which shows if the test participant has spilled or not in either of the two tests. The number or spilling participants is then compared to the total number of participants to attain an *accuracy* percentage for the tested package.[99]

[97] [redacted]

[98] [redacted]

[99] [redacted]

$$Accuracy = 1 - \left(\frac{number\ of\ spilling\ participants}{number\ of\ participants} \right)$$

The *accuracy* measure for the pouring part of the test is somewhat different and calculated in a more complex way compared to the drinking test. First of all, the test participants are supposed to fill four glasses with liquid as close as possible to a marked line. Two of the glasses should be poured carefully, while the remaining two should be poured as quickly as possible, ignoring lack of accuracy and spillage.[100] The *accuracy* per participant is then calculated as:

$$1 - \left(\frac{\left|\Delta TPC\right| + \left|\Delta TPQ\right|}{4 \cdot content\ of\ a\ glass} + \frac{SC \cdot TPSPC}{2 \cdot content\ of\ a\ glass} + \frac{SQ \cdot TPSPQ}{2 \cdot content\ of\ a\ glass} \right)$$

where

$|\Delta TPC| =$ absolute Total amount of liquid by which the marked line was missed when Pouring Carefully
$|\Delta TPQ| =$ absolute Total amount of liquid by which the marked line was missed when Pouring Quickly
$SC =$ Spillage when pouring Carefully; yes = 1, no = 0
$TPSPC =$ Total number of Participants Spilling when Pouring Carefully
$SQ =$ Spillage when pouring Quickly; yes = 1, no = 0
$TPSPQ =$ Total number of Participants Spilling when Pouring Quickly

This method was discarded in consultation with the Package Company early in the work process and was replaced by a another method, much like the one called *Categorized Spillage,* which will be described in the analysis chapter later on (see Section 5.2.1.3). The method is, however, included in the report in order to investigate the pros and cons of all methods that have been important in one way or another.

4.1.2 Efficiency

The usability *efficiency* measures how efficient a task can be performed using a specific product. This is a time-based measure that reflects how effortless or complicated it is to perform a task with the product.

In the present test method, the fastest test participant's time, for the package with the shortest mean time, is inserted into a database and used as a benchmark value to determine the other participants' relative results. This is added up and the usability *efficiency* is presented as a percentage that describes how well the mean time is compared to the fastest time. When measuring the *efficiency* of packages to drink from, this will generate two different *efficiency* values, one for drinking while sitting and one for drinking while walking. These two values are weighted equally to receive an overall *efficiency* value.[101] The formula for the *efficiency* according to the present test method looks like follows:

[100] [redacted]

[101] [redacted]

$$Efficiency = \frac{\left(\frac{fastest\ handling\ time,\ sitting}{mean\ handling\ time,\ sitting}\right) + \left(\frac{fastest\ handling\ time,\ walking}{mean\ handling\ time,\ walking}\right)}{2}$$

This method is obviously also applicable on packages to pour from, but instead of drinking sitting and walking the participant is supposed to pour carefully and quickly.

4.1.3 Satisfaction

The following subchapters will describe the two ways that are used to illustrate the users' experienced *satisfaction* with the product.

4.1.3.1 Handling Questionnaire

In the present test method, the *satisfaction* is measured by the participants' scoring, between one and five, for the words *hold, open, drink, close* and *easy to use,* according to a *handling questionnaire*, like the one found in Appendix A. All the participants' scores are then added up and a mean value for every category is determined. The five mean values are weighted equally and compared to the theoretically optimal score, which is five, to achieve an overall *satisfaction* value:[102]

$$Satisfaction = \frac{MV_{hold} + MV_{open} + MV_{drink} + MV_{close} + MV_{easy\ to\ use}}{5 \cdot 5}$$

where

$MV =$ Mean Value

4.1.3.2 Word Choice List

In addition to the *handling questionnaire*, the participants choose five words that they think are the most descriptive for the package from a *word choice list* containing 36 words (as seen in Appendix B). These words are put together in a chart which shows the frequency of every word, according to Figure 4.2.[103]

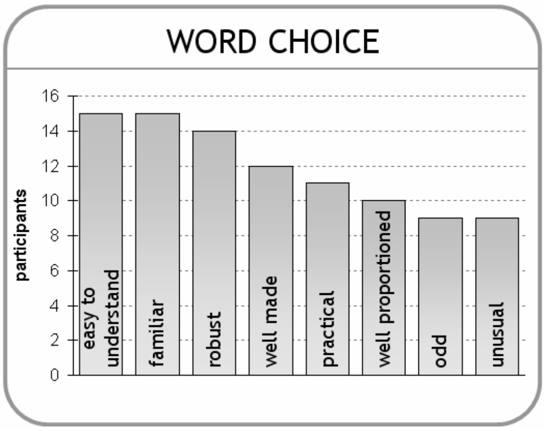


Figure 4.2 *Original Presentation of Word Choices*[104]

[102] [redacted]

[103] *Ibid.*

[104] *Ibid.*

Early in the thesis process, this list was revised by the Package Company to contain only 22 words and the corresponding chart was also revised in order to act as a benchmark tool for the different packages. This can be seen in Figure 4.3, where the *word choices* for four packages are shown.

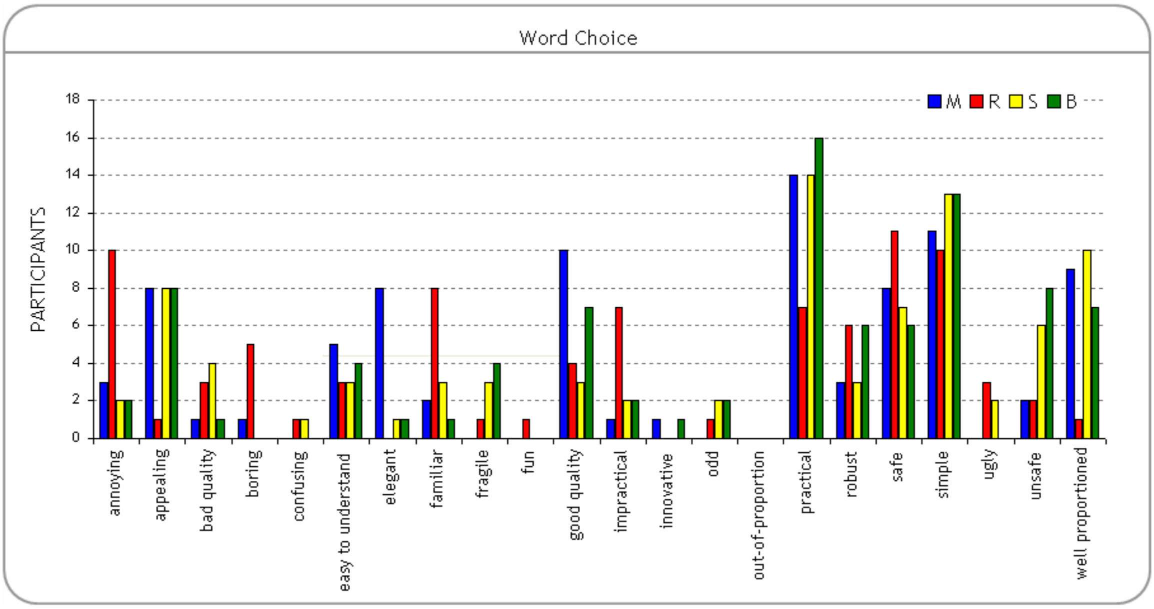


Figure 4.3 *Word Choice Benchmark*[105]

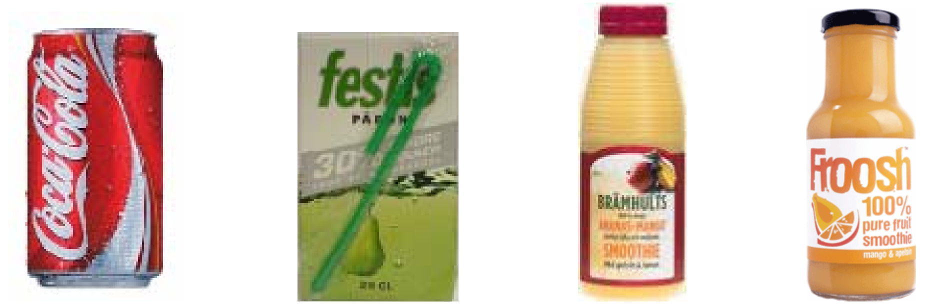
4.2 Test Conduct

In order to learn how the company carried out the usability tests, we started off by studying the present test method. We went through a written document containing the present test plan, including application criteria, equipment, preparations, test execution, data processing and post-test procedures. We also received several video recordings that were thoroughly studied in order to get a clearer view of how to conduct the tests.

4.2.1 Test Procedures

When a proper amount of knowledge and understanding of the usability test had been gathered, the preparations for our own tests began. The first step was to choose the packages that were supposed to be used during the tests. The aim was to find four smaller packages of different kinds for the drinking test and four larger packages for the pouring test. A requirement was that the packages were supposed to have different shapes and opening mechanisms and preferably be of different materials as well. After some discussions with the company regarding the appropriateness of the packages that were initially proposed, we ended up with the decision to use the ones pictured in Figures 4.4 and 4.5.

[105] [redacted]



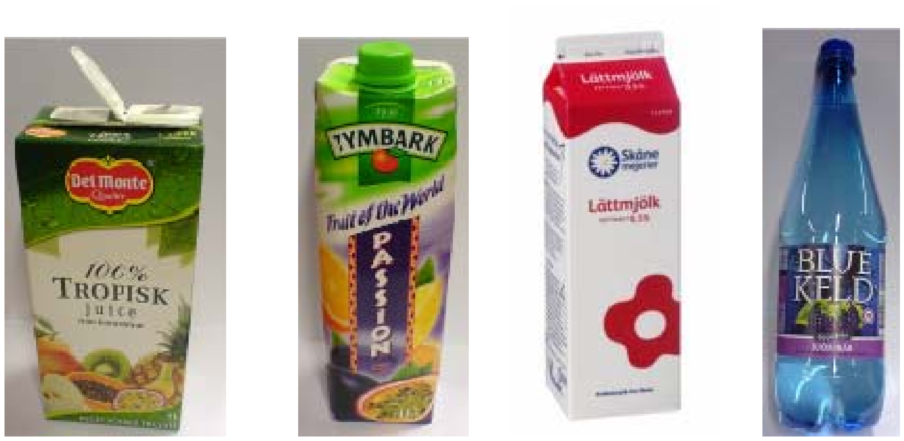
Aluminum Can 33 cl

Brik with Straw 25 cl

Plastic Bottle 38 cl

Glass Bottle 25 cl

Figure 4.4 *Packages for the Drinking Test*



Brik 100 cl

Prisma 100 cl

Gable Top 100 cl

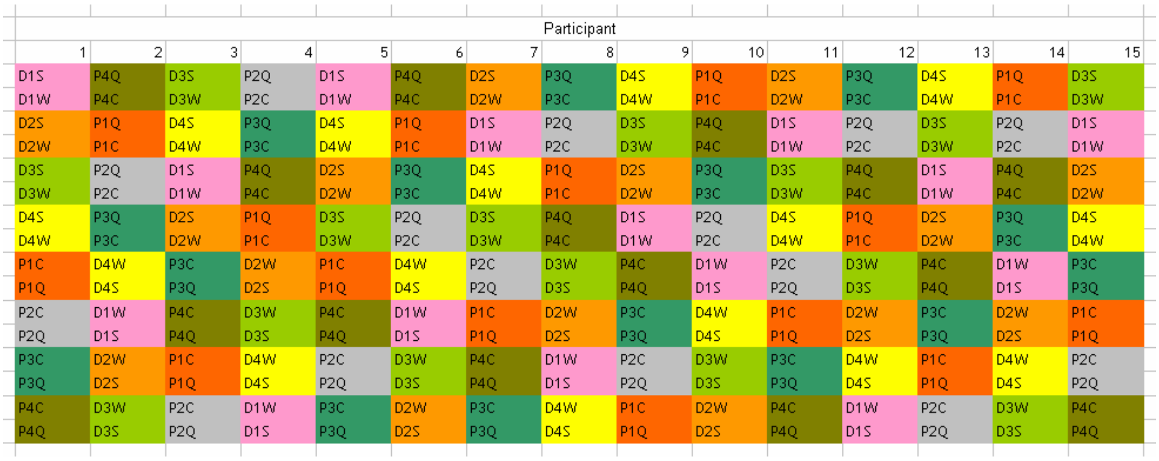
PET Bottle 100 cl

Figure 4.5 *Packages for the Pouring Test*

Subsequently, we explored the usability lab at the university, where the tests were to be held, in order to acquaint ourselves with the equipment and the surroundings. We had, de facto, already used the same usability lab earlier during the course of our education, although it was most definitely necessary to refresh our proficiency. Simultaneously, all necessary test participants were recruited. As we were looking for students it might have seemed like an easy task, although we were forced to use our entire circle of acquaintances, as well as the equivalent circles of theirs. Despite the tough situation we finally managed to acquire 35 students as participants, whereof 20 were male and the remaining 15 were female.

The following step was to set up a rotation scheme in order to prevent any *transfer of learning effects* or similar factors that might have affected the outcome. Considering four different packages per test, there were 24 possible sequences. As the largest single test consisted of 15 participants, we used 15 out of these 24 combinations so that every package would be used in every of the eight possible positions. Figure 4.6 graphically presents the test order for 15 participants, where one column indicates the test sequence for one participant, starting from the top. To be able to compare the impact of the test monitor, the exact same order was used for participant 16 to 30. The pilot test used the five columns to the far left and, later on, the elderly people test (14 participants) used all but the last column.

As seen in the figure below, the participants started alternately with the drinking test or the pouring test. The running order of the subtasks was likewise changed from one participant to another.



Participant														
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
D1S	P4Q	D3S	P2Q	D1S	P4Q	D2S	P3Q	D4S	P1Q	D2S	P3Q	D4S	P1Q	D3S
D1W	P4C	D3W	P2C	D1W	P4C	D2W	P3C	D4W	P1C	D2W	P3C	D4W	P1C	D3W
D2S	P1Q	D4S	P3Q	D4S	P1Q	D1S	P2Q	D3S	P4Q	D1S	P2Q	D3S	P2Q	D1S
D2W	P1C	D4W	P3C	D4W	P1C	D1W	P2C	D3W	P4C	D1W	P2C	D3W	P2C	D1W
D3S	P2Q	D1S	P4Q	D2S	P3Q	D4S	P1Q	D2S	P3Q	D3S	P4Q	D1S	P4Q	D2S
D3W	P2C	D1W	P4C	D2W	P3C	D4W	P1C	D2W	P3C	D3W	P4C	D1W	P4C	D2W
D4S	P3Q	D2S	P1Q	D3S	P2Q	D3S	P4Q	D1S	P2Q	D4S	P1Q	D2S	P3Q	D4S
D4W	P3C	D2W	P1C	D3W	P2C	D3W	P4C	D1W	P2C	D4W	P1C	D2W	P3C	D4W
P1C	D4W	P3C	D2W	P1C	D4W	P2C	D3W	P4C	D1W	P2C	D3W	P4C	D1W	P3C
P1Q	D4S	P3Q	D2S	P1Q	D4S	P2Q	D3S	P4Q	D1S	P2Q	D3S	P4Q	D1S	P3Q
P2C	D1W	P4C	D3W	P4C	D1W	P1C	D2W	P3C	D4W	P1C	D2W	P3C	D2W	P1C
P2Q	D1S	P4Q	D3S	P4Q	D1S	P1Q	D2S	P3Q	D4S	P1Q	D2S	P3Q	D2S	P1Q
P3C	D2W	P1C	D4W	P2C	D3W	P4C	D1W	P2C	D3W	P3C	D4W	P1C	D4W	P2C
P3Q	D2S	P1Q	D4S	P2Q	D3S	P4Q	D1S	P2Q	D3S	P3Q	D4S	P1Q	D4S	P2Q
P4C	D3W	P2C	D1W	P3C	D2W	P3C	D4W	P1C	D2W	P4C	D1W	P2C	D3W	P4C
P4Q	D3S	P2Q	D1S	P3Q	D2S	P3Q	D4S	P1Q	D2S	P4Q	D1S	P2Q	D3S	P4Q

Figure 4.6 *Rotation Scheme for the Packages of the Tests*

The following denotations apply in the scheme above:

D1 – Aluminum Can	*P1* – Tetra Brik
D2 – Tetra Brik with Straw	*P2* – Tetra Prisma
D3 – Plastic Bottle	*P3* – Gable Top
D4 – Glass Bottle	*P4* – PET Bottle
S – Sitting	*C* – Carefully
W – Walking	*Q* – Quickly

The next task was to purchase all the necessary packages, plastic cups, napkins and garbage bags. In total, almost 800 packages were purchased, weighing well above half a metric ton. Furthermore, trays were borrowed to facilitate and expedite the pouring test and all the required questionnaires, *word choice lists* and other sheets necessary for the *test monitor* to be able to note whether the test participants had any problems or not were printed. All these forms can be viewed in Appendices A, C, D and E.

With all the packages purchased and all paperwork done, we conducted the test on ourselves in order to practice the whole procedure and to see how long time the test required. This was done in the usability lab mentioned earlier and a schematic layout of this lab can be seen in Figure 4.7. The tests were monitored by four cameras, whereof two had the functions of panning, tilting and zooming. The angles captured by each camera are marked by the dotted lines in the figure. Behind the test participant there were screens in order to augment the contrast of the recorded material. All the packages and trays were kept at a separate table in order to prevent any confusion for the test participant, but also to increase the workspace area and to improve the visibility. The tests were conducted with one *test monitor* accompanying the participant in the test room and one test supervisor in the observation room.

Appendix F presents the guidelines for the *test monitor*. As seen, there were several tasks to be performed by the test participant. In short, the test was comprised by two separate tasks, each consisting of two subtasks, as briefly described in Section 4.1.1. The pouring test was about opening the package, pouring the liquid up to a predefined mark in two glasses and finally closing the package. The subtasks regarding this particular test were to do it partly as accurately as possible and partly as quickly as possible, with the former disregarding the time taken and the latter with no regard to spillage or accuracy. The drinking test, on its hand, considered opening the package and taking a sip of the liquid it contained, before closing the package. This task was divided into two different subtasks, namely to perform it while sitting by the table and to perform it while walking around the table[106], as indicated by the *walking test route* in Figure 4.7.

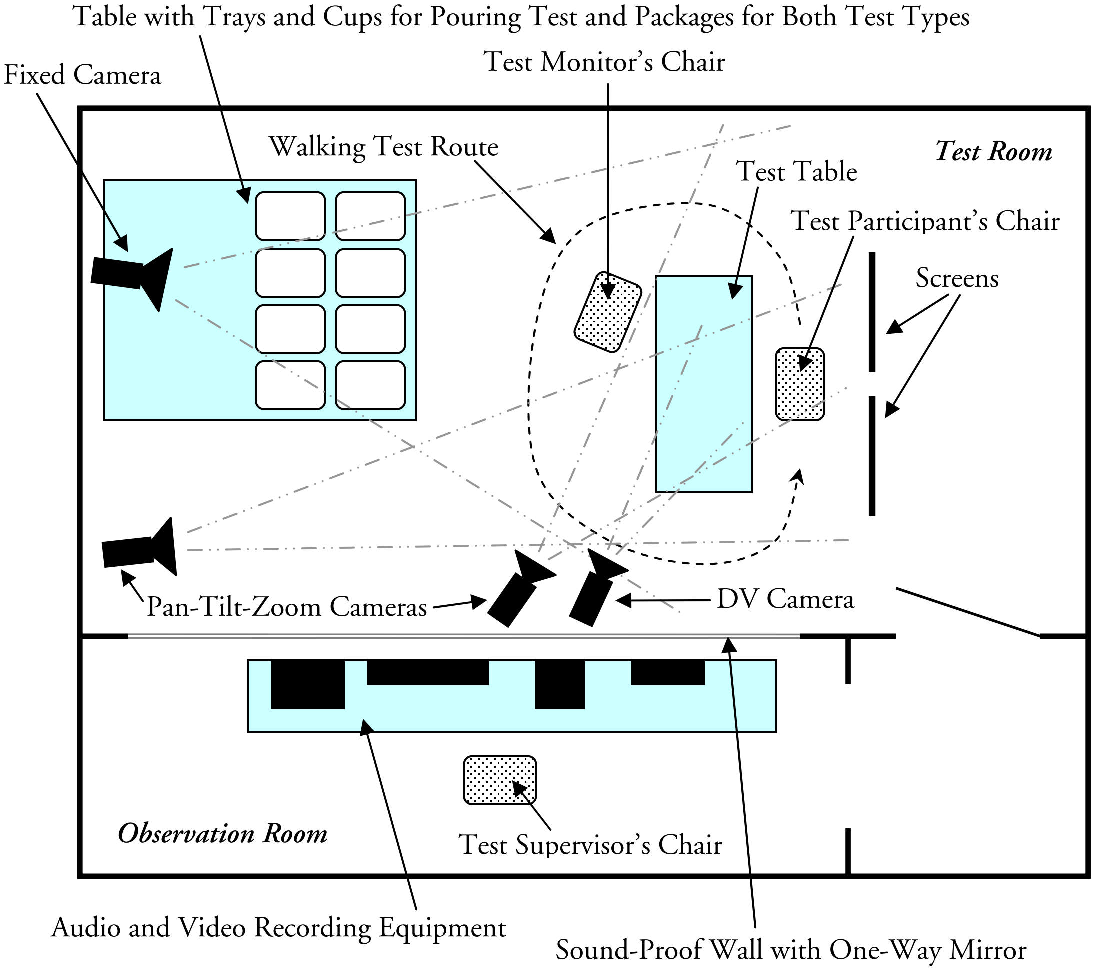


Figure 4.7 *Schematic Layout of the Usability Lab*

Shortly after conducting the test on ourselves, a pilot test was performed on five students with supervisors from the Package Company present. As everything went fine, we continued with the regular tests on 30 other students. After concluding these tests and the analysis of the data gathered during them, we went on with trying to find participants for the following tests. Based on the results of the student tests, we figured that around 20 test participants of every consumer segment would be suitable. The possible segments were, apart from students; children, adults and pensioners. However, due to time restrictions we confined ourselves with sticking to just the pensioners, as we found this segment the most dissimilar to the students.

[106] [redacted]

The problem was that we did not know any pensioners at all living within a reasonable distance. Following this, we contacted PRO, the *Swedish National Pensioners' Organization* and their Lund branch, where we were met by a benevolent attitude to our request. Despite this, they only managed to provide us with a small amount of participants. Allegedly, there had been a large pensioner gathering where our proposal was supposed to be discussed, although the band that played at this meeting was so terrible the attendees were in an extremely bad mood and, accordingly, it was not suitable to raise the question.[107] Nevertheless, with the joint effort of ourselves, PRO and the Lund branch of SKPF, the *Swedish Pensioner Association of Local Government Employees*, we managed to acquire a decent number of pensioners. By recruiting some more test participants that were close to retirement, although technically still not there, we felt that it would be more suitable to rename our user group to *elderly people* instead of pensioners. Due to the at-the-moment upcoming Christmas, it was still difficult to reach the desired amount of 20 participants and we had to settle for 14 elderly participants in the end. We performed the test on these persons, mainly as a field study at the PRO association premises, in order to see if they could manage to undergo it at all and, if so, if their results behaved in the same way as for the students.

4.2.2 Data Collection Before, During and After the Test

The *test monitor's* task was not only to conduct the test, but also to gather a lot of information about how well the test participant managed to use the packages. Before the test commenced, the participant was asked to fill out a questionnaire regarding his or hers thoughts about the usability of the package. During the test itself the *test monitor* noted, apart from age and gender, if the participant

- Had difficulties opening the package
- Did not manage to open the package at all
- Spilled while opening the package
- Spilled while pouring/drinking
- Caused dripping from/alongside the package
- Had problems closing the package

The above-mentioned problems apply to both the drinking and the pouring test. After the pouring tests, there were some additional measurements the *test monitor* had to perform, namely:

- The weight of the two glasses (including the liquid in them)
- The total area of the spilled liquid
- The total weight of the spilled liquid

These notations were made in the sheets seen in Appendices D and E. After the test, the participants were asked to fill out another questionnaire, similar to the one filled out before the test, in order to see if the intuitive feeling they had before the test was correct (the questions regarded exactly the same things as the questionnaire that can be seen in Appendix A, and therefore, these two questionnaires are combined into one in the appendix). The test participants also had the chance to express their feelings orally during a post-test debriefing. In addition to this data, the time taken by the participant to perform the tasks was measured using the aid of the video recordings of the tests.

[107] Persson, B (2006-12-11)

Using the data from the student tests, *dashboards* were created for the eight packages for:

- All 35 participants
- Pilot test (TP1-TP5)
- Test 1 (TP6-TP20)
- Test 2 (TP21-TP35)
- Erik as *test monitor*
- Tomasz as *test monitor*
- All male participants
- All female participants
- Erik's male participants
- Erik's female participants
- Tomasz' male participants
- Tomasz' female participants
- Five males & five females from Erik's participants
- Five males & five females from Tomasz' participants

An example of such *dashboards* can be found in Appendix H and the usability values obtained are given in Appendix K.

4.2.3 Consistency between Theory and Course of Action

This subchapter is meant to examine how well our work procedure concurs with the usability testing theory presented in Section 3.2.

This report contains a problem and purpose statement, fully in agreement with the existing theory. A user profile has not been made, as the task was to perform the study initially on students and then on other consumer segments, rather than investigating all possible segments. What tasks are to be performed by the participants, how this is supposed to be done and under what circumstances are all described in the preceding sections of the report. The *test monitor's* role and what data should be gathered during and after the tests are also presented, just like the usability theory suggests.

The question of an adequate number of test participants is not an issue in this case as it is one of the main tasks of the thesis to determine this appropriate amount. When it comes to test design, we have essentially been using the within-subjects design, although a lot of participants have been used. The counter-balancing technique has also been applied. We did not have much to choose from regarding test monitors, however, we find ourselves to fit the requirement description more than well.

All four kinds of tests described in the theory section have been used. Small elements of *exploratory testing* are present as we were interested in the reasons to the behavior of the participants. Elements of *assessment testing* are as well detectable due to the fact that the participants were given specific tasks and that the measurements were of a quantitative nature. Still, our tests have perhaps most in common with *validation testing*, as eight different packages were tested and compared to benchmark values. In addition, the tests were quite strict and formalized and the amount of interaction between the *test monitor* and the participant was very limited. Obviously, there are elements of *comparison testing* present as well. Additionally, the usability lab at the university fulfils all the prerequisites of such premises, as described in Section 3.2.6.

Overall, our conditions were most definitely sufficient to perform a proper usability study according to the given theory and as this theoretical frame of reference is widely recognized, we see no reason to question the credibility of our course of action.

4.3 Presentation of the Usability Score

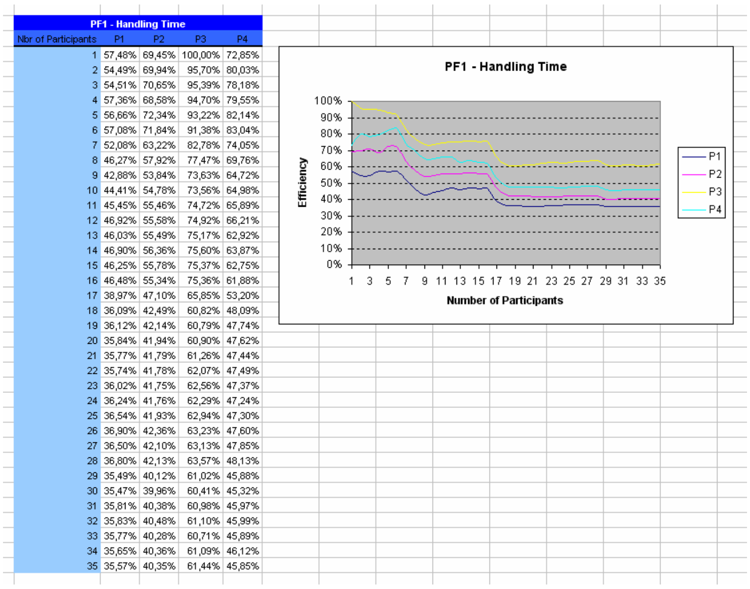
All the usability scores are compiled to the *dashboard report sheet*, as seen in Appendix G. This is the Package Company’s modified and improved version of the present layout, which was designed to contain all the necessary information about the package, including the name and a picture of it and essential design information like cap style and opening mechanisms. The usability scoring is presented in the form of speedometers. The components of the overall usability value, namely *accuracy*, *efficiency* and *satisfaction*, are thoroughly specified to provide a clearer view of the package’s usability. The representation of the *word choices* contains competitor comparison values as well, in order to determine the relative usability of the package. The competitor values for *accuracy*, *efficiency* and *satisfaction* are also given for the same reason.

4.4 Data Processing

As a part of our study, we were also supposed to determine the required amount of test participants. Two different methods were used, chart exploration and statistical calculations, as well as the Wilcoxon signed rank sum test was used as an additional supporting method. This section will explain how these methods were used.

4.4.1 Chart Exploration

This method was used just the way it is presented in Section 3.3.1. By linking the values of the *dashboard* files to a spreadsheet, values for the *accuracy*, *efficiency*, *satisfaction* and *cognitive load* were generated (due to the iterative process of the thesis, the concept of *cognitive load* is mentioned here, although it will not be presented and evaluated until Section 5.2.4). The data in these tables was then made into a chart for improved visualization. An example of this is presented in Figure 4.8.



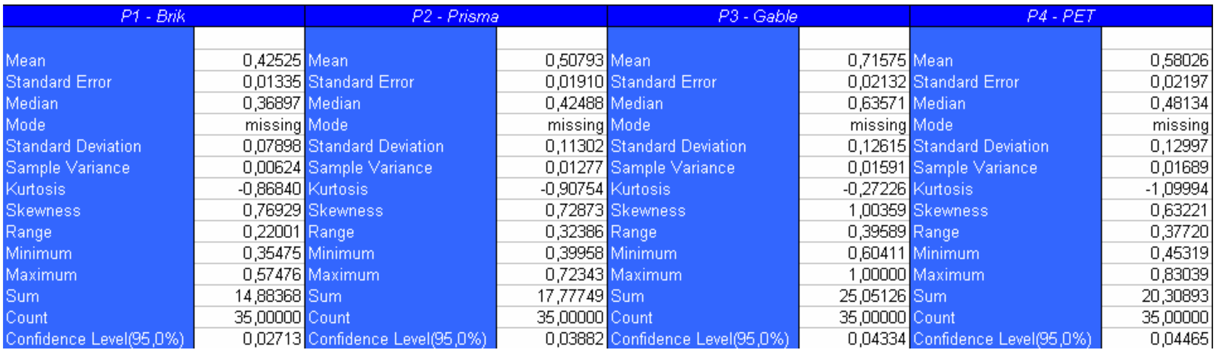
PF1 - Handling Time				
Nbr of Participants	P1	P2	P3	P4
1	57,48%	69,45%	100,00%	72,85%
2	54,49%	69,94%	95,70%	80,03%
3	54,51%	70,65%	95,39%	78,18%
4	57,36%	68,58%	94,70%	79,55%
5	56,66%	72,34%	93,22%	82,14%
6	57,08%	71,84%	91,38%	83,04%
7	52,08%	63,22%	82,78%	74,05%
8	46,27%	57,92%	77,47%	69,76%
9	42,88%	53,84%	73,63%	64,72%
10	44,41%	54,78%	73,56%	64,98%
11	45,45%	55,46%	74,72%	65,89%
12	46,92%	55,58%	74,92%	66,21%
13	46,03%	55,49%	75,17%	62,92%
14	46,90%	56,36%	75,60%	63,87%
15	46,25%	55,78%	75,37%	62,75%
16	46,48%	55,34%	75,36%	61,88%
17	38,97%	47,10%	65,85%	53,20%
18	36,09%	42,49%	60,82%	48,09%
19	36,12%	42,14%	60,79%	47,74%
20	35,84%	41,94%	60,90%	47,62%
21	35,77%	41,79%	61,26%	47,44%
22	35,74%	41,78%	62,07%	47,49%
23	36,02%	41,75%	62,56%	47,37%
24	36,24%	41,76%	62,29%	47,24%
25	36,54%	41,93%	62,94%	47,30%
26	36,90%	42,36%	63,23%	47,60%
27	36,50%	42,10%	63,13%	47,85%
28	36,80%	42,13%	63,57%	48,13%
29	35,49%	40,12%	61,02%	45,88%
30	35,47%	39,96%	60,41%	45,32%
31	35,81%	40,38%	60,98%	45,97%
32	35,83%	40,48%	61,10%	45,99%
33	35,77%	40,28%	60,71%	45,89%
34	35,65%	40,36%	61,09%	46,12%
35	35,57%	40,35%	61,44%	45,85%

Figure 4.8 *Efficiency for the Pouring Test by Chart Exploration*

The resulting charts achieved with the aid of this method can be found in Appendix K.

4.4.2 Statistical Calculation

As the test participants in our study do not represent the total population, the unbiased method described in Section 3.3.2 was used. Instead of calculating the standard deviation ourselves, we used the *Descriptive statistics* tool in Excel, which also generates several other statistical values that may be of interest (see Figure 4.9). The standard deviation generated by this function is calculated using the unbiased method.



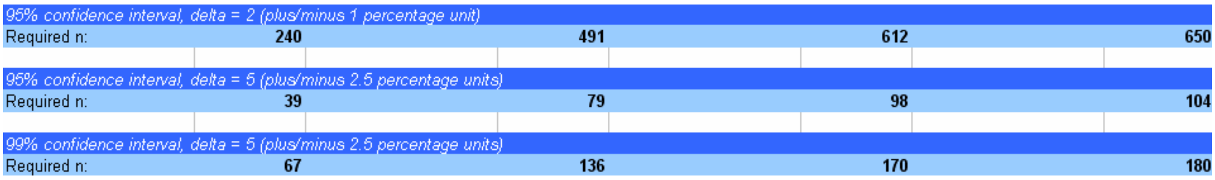
P1 - Brik		P2 - Prisma		P3 - Gable		P4 - PET	
Mean	0,42525	Mean	0,50793	Mean	0,71575	Mean	0,58026
Standard Error	0,01335	Standard Error	0,01910	Standard Error	0,02132	Standard Error	0,02197
Median	0,36897	Median	0,42488	Median	0,63571	Median	0,48134
Mode	missing	Mode	missing	Mode	missing	Mode	missing
Standard Deviation	0,07898	Standard Deviation	0,11302	Standard Deviation	0,12615	Standard Deviation	0,12997
Sample Variance	0,00624	Sample Variance	0,01277	Sample Variance	0,01591	Sample Variance	0,01689
Kurtosis	-0,86840	Kurtosis	-0,90754	Kurtosis	-0,27226	Kurtosis	-1,09994
Skewness	0,76929	Skewness	0,72873	Skewness	1,00359	Skewness	0,63221
Range	0,22001	Range	0,32386	Range	0,39589	Range	0,37720
Minimum	0,35475	Minimum	0,39958	Minimum	0,60411	Minimum	0,45319
Maximum	0,57476	Maximum	0,72343	Maximum	1,00000	Maximum	0,83039
Sum	14,88368	Sum	17,77749	Sum	25,05126	Sum	20,30893
Count	35,00000	Count	35,00000	Count	35,00000	Count	35,00000
Confidence Level(95,0%)	0,02713	Confidence Level(95,0%)	0,03882	Confidence Level(95,0%)	0,04334	Confidence Level(95,0%)	0,04465

Figure 4.9 *Descriptive Statistics for the Efficiency Values (Pilot and Student Test)*

The next step was to calculate the required amount of test participants based on the standard deviation with three different approaches;

1. A confidence level of 95% and a delta value of 2 percentage units
2. A confidence level of 95% and a delta value of 5 percentage units
3. A confidence level of 99% and a delta value of 5 percentage units

The resulting amounts of required test participants, calculated from the statistics in Figure 4.9, are seen in Figure 4.10.



95% confidence interval, delta = 2 (plus/minus 1 percentage unit)							
Required n:	240		491		612		650
95% confidence interval, delta = 5 (plus/minus 2.5 percentage units)							
Required n:	39		79		98		104
99% confidence interval, delta = 5 (plus/minus 2.5 percentage units)							
Required n:	67		136		170		180

Figure 4.10 *Required Amounts of Test Participants Using the Standard Deviation Method*

After investigating the results it became clear that the most reasonable choice was to use the second alternative (level of confidence: 95%; delta value: 5 percentage units) as the other two were far too strict.

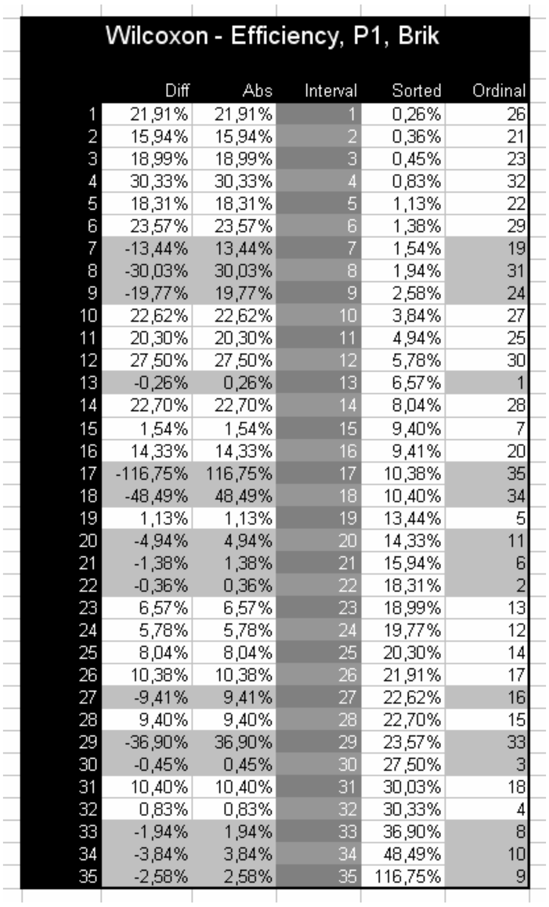
The biased method was also examined, although not for the elderly people data as there were too few participants in that test to get a correct value. The method gave practically the same results as the unbiased one, just as predicted.

The results yielded by this method are shown in Section 5.3.2.

4.4.3 Wilcoxon Signed Rank Sum Test

The Wilcoxon test was used in order to see if there were reasons to doubt the credibility of the results calculated using the standard deviation formula. Due to the fact that we wanted to see if the results of a few individuals had a major impact on the outcome, we had to use the participants' individual contribution to the cumulative average. This is a consequence of the fact that the standard deviation method used the cumulative average and not the individual result of each participant. However, this only applies to the *accuracy* and *efficiency* measures as these are constantly recalculated; the participants' individual results for the remaining measures can easily be obtained by backward calculation.

After having that part done, the cumulative average was chosen as the hypothesized value. Just like earlier mentioned in Section 3.3.3, there is nothing that implies that this is the correct value, but we were interested in investigating possible reasons to strange results and not the question of asymmetry itself. Put shortly, a large *u* value was an indicator of asymmetric tendencies and could be seen as a waiver for the possibly large quantities of test participants recommended by the statistical calculations.



Wilcoxon - Efficiency, P1, Brik

	Diff	Abs	Interval	Sorted	Ordinal
1	21,91%	21,91%	1	0,26%	26
2	15,94%	15,94%	2	0,36%	21
3	18,99%	18,99%	3	0,45%	23
4	30,33%	30,33%	4	0,83%	32
5	18,31%	18,31%	5	1,13%	22
6	23,57%	23,57%	6	1,38%	29
7	-13,44%	13,44%	7	1,54%	19
8	-30,03%	30,03%	8	1,94%	31
9	-19,77%	19,77%	9	2,58%	24
10	22,62%	22,62%	10	3,84%	27
11	20,30%	20,30%	11	4,94%	25
12	27,50%	27,50%	12	5,78%	30
13	-0,26%	0,26%	13	6,57%	1
14	22,70%	22,70%	14	8,04%	28
15	1,54%	1,54%	15	9,40%	7
16	14,33%	14,33%	16	9,41%	20
17	-116,75%	116,75%	17	10,38%	35
18	-48,49%	48,49%	18	10,40%	34
19	1,13%	1,13%	19	13,44%	5
20	-4,94%	4,94%	20	14,33%	11
21	-1,38%	1,38%	21	15,94%	6
22	-0,36%	0,36%	22	18,31%	2
23	6,57%	6,57%	23	18,99%	13
24	5,78%	5,78%	24	19,77%	12
25	8,04%	8,04%	25	20,30%	14
26	10,38%	10,38%	26	21,91%	17
27	-9,41%	9,41%	27	22,62%	16
28	9,40%	9,40%	28	22,70%	15
29	-36,90%	36,90%	29	23,57%	33
30	-0,45%	0,45%	30	27,50%	3
31	10,40%	10,40%	31	30,03%	18
32	0,83%	0,83%	32	30,33%	4
33	-1,94%	1,94%	33	36,90%	8
34	-3,84%	3,84%	34	48,49%	10
35	-2,58%	2,58%	35	116,75%	9

Figure 4.11 *Table Used to Make a Wilcoxon Test*

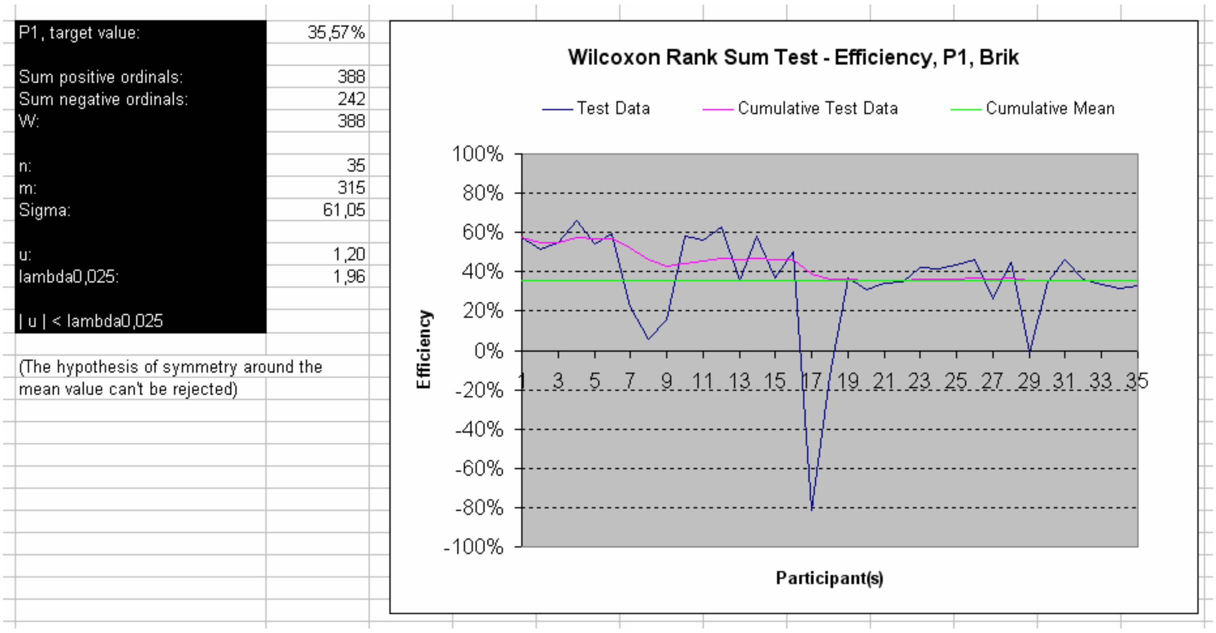
Figure 4.11 shows one of our tables used for the Wilcoxon test. The columns indicate, from left to right;

- Participant identification number
- Difference between individual contribution to the cumulative average (or in the cases of *satisfaction* and *cognitive load*; the individual result) and the target value

- Absolute value of the difference
- Rank interval (if there are multiples of a difference they share the same rank interval)
- Sorted absolute values of the differences, from the smallest to the largest
- Ordinal number

The ordinal number is the key to the whole test. In order to determine it, we took the participant's difference value (column 2) before searching for it in the list of absolute values (column 5). After finding it, we took one step to the left (to column 4, same row) to find out the rank interval. This value is also the participant's ordinal number and is inserted in the corresponding cell in column 6 (the same row as the participant's identification number).

The light gray cells in column 6 indicate negative differences and the sum of ordinal numbers in these constitute the W^- value. The sum of the rest of the values in the ordinal number column is, consequently, equal to W^+. The rest of the calculations simply follow the formulas in Section 3.3.3. To visualize the results, we made a graph showing the individual participant's contribution to the cumulative average, the cumulative average and the cumulative mean (equal to the cumulative average for all participants). An example of such a graph is seen in Figure 4.12. The graph is based on the data in Figure 4.8.



P1, target value:	35,57%
Sum positive ordinals:	388
Sum negative ordinals:	242
W:	388
n:	35
m:	315
Sigma:	61,05
u:	1,20
lambda0,025:	1,96
| u | < lambda0,025	

(The hypothesis of symmetry around the mean value can't be rejected)

Figure 4.12 *Graph Representing the Wilcoxon Test Result for a Pouring Test Package*

In the specific example shown above, it is clear that the very large amount of participants recommended by the statistical calculation (39 if looking only at the P1 package; over 100 participants if taking all the four packages of the test into consideration) is to a large extent caused by the three deep dips clearly visible in the graph in Figure 4.12. The effects of these dips are also visible in the chart exploration graph (see Figure 4.8). The reason for these dips in the *efficiency* is that a new best time has been set, causing a recalculation of all the *efficiency* values which, in turn, makes the previous times significantly worse in comparison to the leader than before. As a consequence, all *efficiency* values are lowered and this appears in the form of such dips in the graphs.

The *u* value in the example above is 1.20, as can be seen in Figure 4.12. This is rather high and indicates that the large amounts of recommended participants are due to a few participants' deviating results. This conclusion is strengthened by the charts in Figures 4.8 and 4.12 and the deep dips therein.

The results of the Wilcoxon tests are presented in Section 5.3.3.

5 Analysis and Results

The analysis has the intention of recognizing difficulties in the present method and figure out possible ways to overcome them. This chapter will explain alternative ways of calculating the different parts of the usability value, so that a comparison can be made against the present methods explained in the empirics chapter. In order to enhance the visibility in the presentation of the results, there will also be a discussion about how this can be accomplished. At the end, the necessary number of test participants will be established. In addition to this, the impact of the test monitor and the participants' characteristics will be investigated.

5.1 The Present Method

This chapter will explain the major shortcomings of the present method. This is useful for determining what should be improved in the final model.

5.1.1 Accuracy

This subchapter will deal with the pouring test *accuracy* for the present method. As earlier described, the formula for the *accuracy* of the drinking test is a bit different and takes fewer aspects into consideration. Because of this, we will focus on the pouring test from now on, as we want to bring about a method that can be applied to both test types.

The present method to measure the *accuracy* provides a misleading view of the usability *effectiveness* of the package in question. There are only two alternatives (yes/no) to answer the question if the participant spills. This means that the participant will get the same *accuracy* result irrespective of if spillage occurs in one of the tests or in both of them.

When the test is constructed the way it is in the present method, the ability to pour a requested amount of liquid is tested by measuring by how much the requested amount of poured liquid is missed. The only way to make sensible comparisons with this data is to compare it to the total requested amount of poured liquid, as this value can certainly vary between 0 and 100%. This can be interpreted as a gauge for how easy it is to regulate the flow of the liquid. The problem with this is that it generates quite high *accuracy* values with diminutive differences between packages.

Another problem with the method is that it is mathematically incorrect as it divides the number of participants that spills with the amount of poured liquid. A better way to calculate the *accuracy* value would be:

$$Accuracy = \left(\left(1 - \frac{|\Delta TPC| + |\Delta TPQ|}{VP} \right) + \left(1 - \frac{TPS}{TNP} \right) \right) \cdot \frac{1}{2}$$

where

$|\Delta TPC| =$ absolute Total amount of liquid by which the marked line was missed when Pouring Carefully

$|\Delta TPQ|$ = absolute Total amount of liquid by which the marked line was missed when Pouring Quickly

VP = Volume to Pour

TPS = Total number of Participants Spilling

TNP = Total Number of Participants

Problems that arise when using this method are quite a few;

- As mentioned before, the divergence in ΔTPC and ΔTPQ between packages is hardly recognizable. The variation is rather due to chance, and is more significant between different test participants than between different packages.

- As the formula was stated in the present method, the result for the pouring *accuracy* is only affected if the marked line is missed or if the liquid misses the glass. The *accuracy* is, thus, unaffected by the amount of spilled liquid or how the spillage occurred, as long as the participant spills. This can also be questioned as it is indisputably worse to spill a lot than to spill just a few drops. Promising solutions for this problem could be to measure the amount of spilled liquid or to further investigate in what way, or why, the spillage occurs.

- The formula does not include other errors made in addition to spillage in the *accuracy* measurement, such as not perfectly resealed caps and difficulties in opening the package.

- It is not applicable on the drinking tests because ΔTPC and ΔTPQ cannot be quantified in these tests.

A conclusion of this is that ΔTPC and ΔTPQ can be eliminated from the test method without any loss of valuable data.

5.1.2 Efficiency

The present method of calculating the *efficiency* is very promising, but it has some minor shortcomings:

- It has no way of telling how the time is distributed between the different phases of the test. It may be of interest to know which part of the test is the most and the least time consuming.

- A package that has no cap, and therefore cannot be resealed, has a big advantage over other packages, as the total handling time will inevitably be shorter.

5.1.3 Satisfaction

The users' attitude towards the package, or satisfaction thereof, is sometimes thought of as less reliable than other usability measures, perhaps because it is impossible to measure using quantifiable methods. We believe that a package's usability is just as much about the *satisfaction* of the user as the *efficiency* and *effectiveness* are. In the existing test method, *effectiveness-efficiency-satisfaction* are weighted 40-40-20, but according to us, a more fair weighting would be to give all parts equal importance.

A minor predicament with the present formula is that the *satisfaction* value varies between 20-100%. This is because there are only four “grade levels” between 1 and 5, which makes 1 point equal to 20%. This can easily be dealt with by dividing the average sum of points with the amount of possible grade levels. The formula for *satisfaction*, or *attitude*, will then look as follows:

$$Satisfaction = \frac{MV_{hold} + MV_{open} + MV_{drink} + MV_{close} + MV_{easy\ to\ use}}{5 \cdot 4}$$

where

$MV =$ Mean Value

The *word choice* chart seen in Figure 4.3 is not easily interpreted because it contains vast amounts of information. This is normally desirable, but the inclusion of benchmark values in this particular case makes the chart messy and reduces the visibility.

5.2 New Approaches

The present approach for determining the usability has some shortcomings. To assure that the methods underlying the calculation of the *overall benchmark* will be well proven we have constructed and evaluated twenty competitive alternatives. Due to the desire of keeping the attention of the reader throughout the report (and not exceed 500 pages) we have chosen to only include the most interesting alternatives below.

5.2.1 Accuracy

One of the main objectives of this thesis was to figure out the most appropriate way to calculate the *accuracy*. This has proven to be very challenging as the method needs to take a lot of different aspects in consideration. Some of the tested packages may have very good pouring capacity but are almost impossible to open, while other may be easy to open but impossible to reseal. All of this must be incorporated in the solution according to their relative importance. During the tests, we have gained a lot of understanding of what problems are worse than others and which packages we believe deserve a higher usability score relative the other ones. To be able to perform a valid comparison between different solutions we have made a chart that represents the four packages of the pouring test (P1 – P4) and their *accuracy* score for all the twenty developed methods, see Figure 5.1. Some of the methods do not reflect the usability of the package properly. As seen in the chart, one of the biggest problems has been how to separate P2, P3 and P4 as their performance is similar, although they have some significant differences. These differences have to be evaluated and graded according to their usability impact in order to obtain a sensible *accuracy* calculation method.

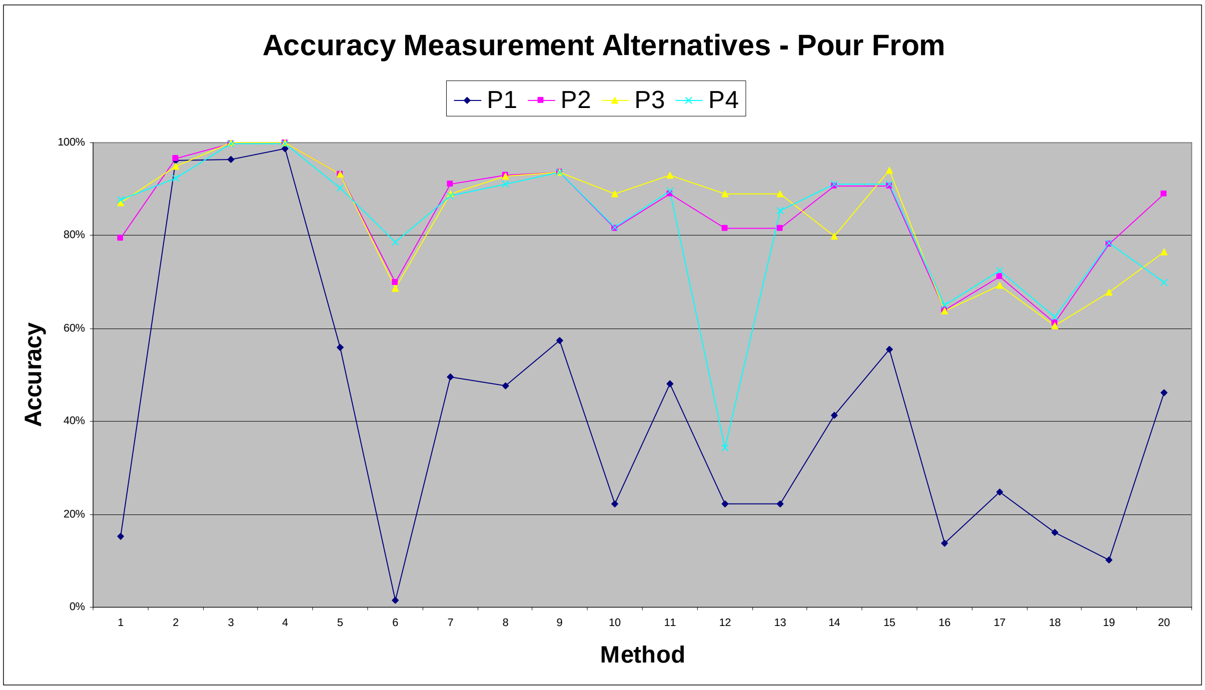


Figure 5.1 *Accuracy Measurement Alternatives, Pouring Test*

In the case of calculating the *accuracy* for the packages of the drinking test, the alternative methods are fewer. This is due to the fact that there are not as many ways of measuring the *accuracy* when the test participant drinks from the package, as it is practically impossible to measure the amount of spillage (or the equivalents to Δ*TPC* and Δ*TPQ*). The results of the alternative methods for packages to drink from are presented in Figure 5.2 below.

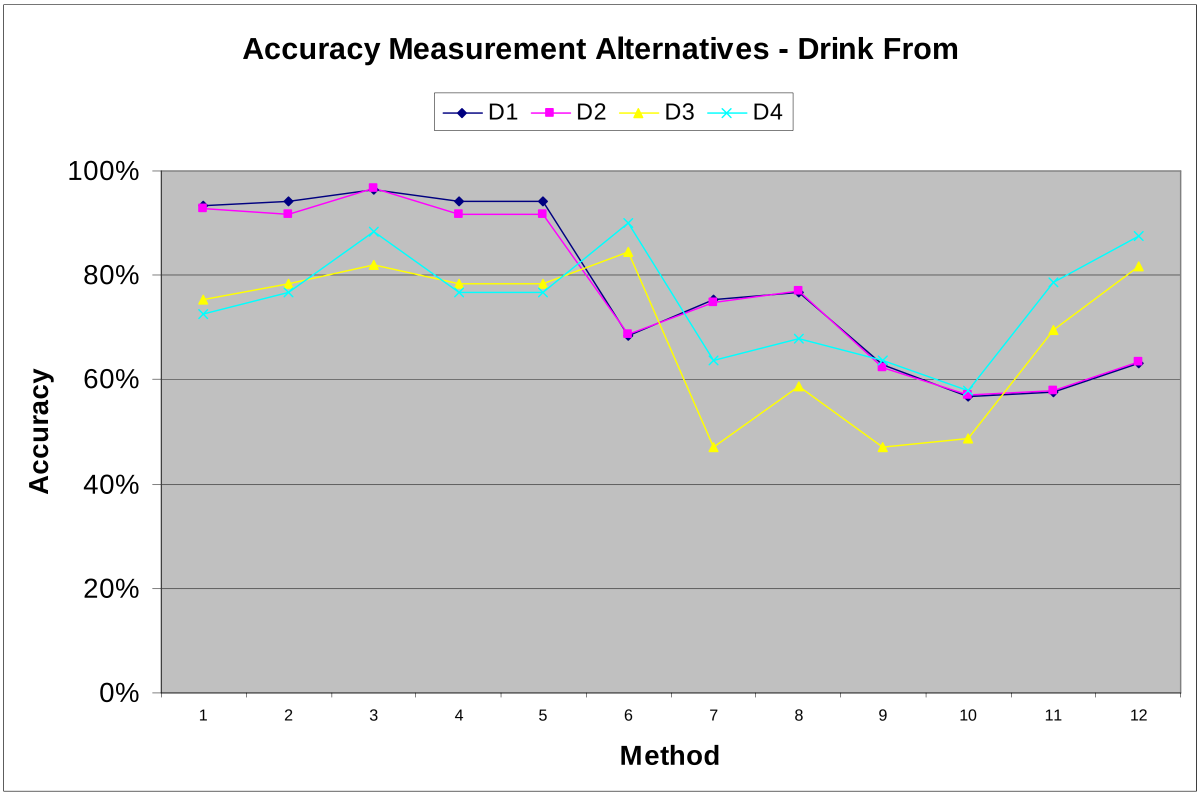


Figure 5.2 *Accuracy Measurement Alternatives, Drinking Test*

We have found some of the alternatives mentioned above more promising or interesting in one way or another than others. Some of the alternatives were further explored upon request by the Package Company, whereas others are simply the most interesting generic models. These alternatives for calculating the pouring test *accuracy* are further described and presented graphically below. The drinking test *accuracy* values are not presented as most of the methods are not applicable on that test.

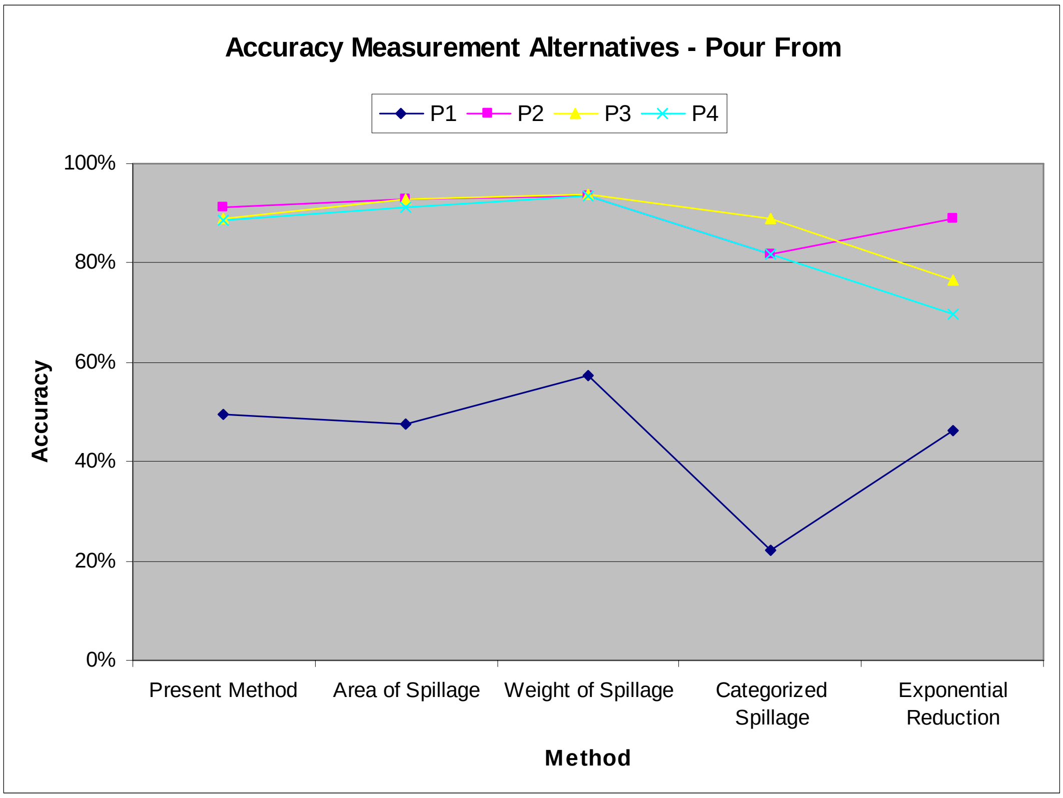


Figure 5.3 *Accuracy Values for Present Method and New Approaches*

5.2.1.1 Area of Spillage

One way of calculating the *accuracy* is by measuring the area of the spillage. There is no natural quantity of the same dimension to compare the area with, and a comparison with the lowest measured value would inevitably mean comparing to a spillage area of zero, which does not make any sense. It would be possible to compare the spillage area with the area of the tray, but in most cases, this area is so much larger so that even a relatively large spillage would have a low impact on the *accuracy* score. We believe that even a relatively small spillage should lead to a poor *accuracy* value, and therefore, the *accuracy* cannot be a linear function of the spillage area. Instead, we have decided to find an exponential function that generates results that are more corresponding to the experienced inconvenience. The following formula has proven to return acceptable values.

$$Accuracy = \left(e^{-AS}\right)^{\frac{1}{10}}$$

where

$AS =$ Area of Spillage

This method has, however, several shortcomings;

- It is impossible to measure the exact spillage area as the degree of surface tension makes the area vary between two different spillages of the same volume.

- Knowing the amount of spillage does not facilitate for the design team to understand the problem and help them to improve the package. The results from the usability tests should be at least as informative for the development team as usable for determining a proper *accuracy* value.

- It is only applicable on packages which are meant to pour from.

5.2.1.2 Weight of Spillage

Another approach for calculating the *accuracy* is by performing the calculations with the weight of the spilled liquid as the basis. As was the case in the previous alternative, there is no natural comparing factor, except perhaps for the weight of the requested poured liquid. This is however a forced solution and does not provide a good base for determining the *accuracy*.

Using the same argument as for the spillage area, the *accuracy* can be calculated using a more reality based exponential function, according to the experienced inconvenience for the user of the package. The exponent has been changed compared to preceding approach in order to generate more accurate results.

$$Accuracy = \left(e^{-(WSN-WDN)}\right)^{\frac{1}{5}}$$

where

WSN	=	Weight of Soaked Napkin
WDN	=	Weight of Dry Napkin

The negative aspects of this method are:

- Using the spillage weight when calculating the *accuracy* is more exact than using spillage area. At the same time, though, it is more time consuming to soak up the spillage in a napkin and weigh it.

- The weight of the spillage does not provide any information concerning the root of the usability problem.

- The method cannot be used in usability tests where the participants are supposed to drink from the package.

A conclusion from this method, and the previous, is that it is unnecessary to spend resources on measuring quantities such as spillage area and weight. These measurements do not provide any value for the design team and are not suitable for calculating the *accuracy*.

5.2.1.3 Categorized Spillage

Instead of measuring the amount of spillage, the *test monitor* can easily notice in what way the spillage occurs. The three most probable types are *spillage while opening*, *spillage while* *pouring/drinking* and *dripping alongside the package*. This information is far more important as it helps the developers in understanding the cause, instead of the effect, of the design. To further investigate plausible sources of error it is also a good idea to include *difficulties opening* and *difficulties closing* of the package. For every source of error, the sum of all participants' mistakes is calculated. The formula is constructed in such a way that it only considers columns that are "activated", that is, with a least one error. In practice, this means that errors that are not applicable (such as *difficulties closing* in the case of the aluminum can for instance) will not be in favor for the overall *accuracy* score. The values for every column are then used for calculating the *accuracy* formula, which provides illusory good results:

$$Accuracy = 1 - \left(\left(\frac{\sum_{j=1}^{5} TNE_j}{TNP} \right) \cdot \left(\frac{1}{\sum_{j=1}^{5} TNE_j > 0} \right) \right)$$

where

TNP = Total Number of Participants
TNE_j = Total Number of Error j (1 = Open & spill, 2 = Pour & spill, 3 = Drip, 4 = Difficulties closing (*misthreaded* cap), 5 = Difficulties opening)

Difficulties that arise when using this method are:

- When measuring the *accuracy* in this way, the method actually generates better values when the number of errors increases, as long as the new errors are of a rare kind and the total sum of errors in the newly "activated" column is at a lower level than that of the columns that are already active. In the same way, the method distorts the individual's contribution to the cumulative average because not all error types are included at all times. This can generate "virtual" individual *accuracy* results above 100%. A way of omitting this problem is to include all the columns in the calculation at all times, even if "errors" of every kind have not occurred, but this method tends to generate unreasonably good results, even for poor packages.

- The method does not take into consideration if the package could not be opened or fully resealed.

A conclusion from this method is that it is seldom beneficial to complicate matters more than necessary. A keyword for creating successful models is to keep it as simple as possible.

5.2.1.4 Exponential Reduction

The final method for the calculation of the *accuracy* is a revision of *Categorized Spillage*, where all possible errors always are included in the calculation.

One of the worst predicaments test participants can face is packages that they cannot break open. The occurrence of this phenomenon is very grave from a usability point of view and cannot be tolerated. The problem is, however, to make this error stand out and have a greater impact than the other possible errors. Our solution to this is, once again, to use the exponential function, as we believe that it is intolerable, and should lead to a greatly reduced usability score, even if only a few test participants cannot open the package. How to solve this problem can be discussed in eternity as there are no correct answers, just a lot of more or less well thought-out suggestions. Our belief is, nevertheless, that using the formula below will generate a somewhat fair reduction value for packages that could not be opened. The formula implies that if 20% of the test participants fail to open the package, the *accuracy* reduction will be of a magnitude of 45%. This reduction rate can be seen in Figure 5.4 below.

$$Exponential\ Reduction = \left(\frac{TNE_7}{TNP}\right)^{\frac{1}{2}}$$

where

$TNP =$ Total Number of Participants
$TNE_7 =$ Total Number of Error 7 (Did not open)

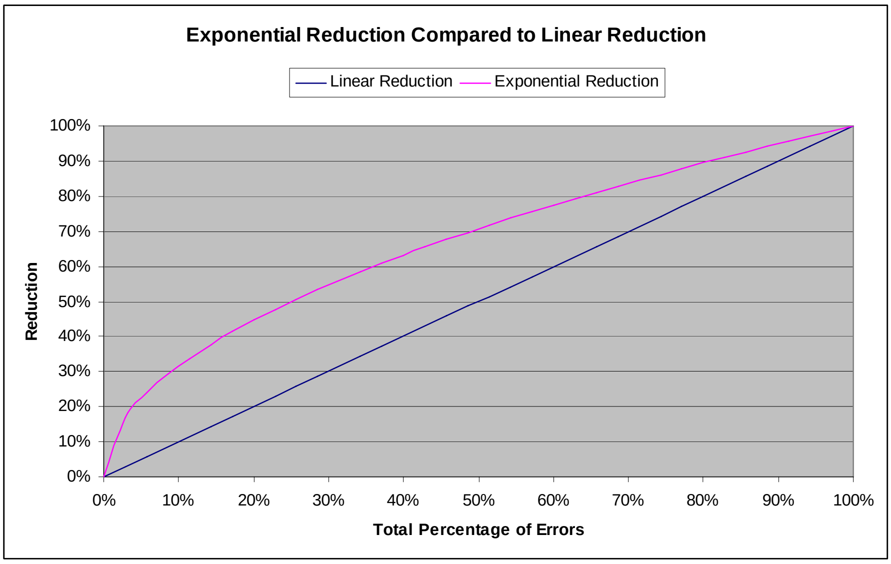


Figure 5.4 *Exponential Reduction*

The decision of using ½ as the exponential growth rate is made because it generates a reduction rate that we believe is fairly appropriate.

Another big issue that we believe should be made clearly visible in the overall *accuracy* score is if the package cannot be resealed properly. A question that arises is how this should be weighted in the solution in order to make a fair judgment against packages that can be closed and therefore may have reduced usability scores caused by difficulties closing. Our solution to this is that if there is any question about whether or not the package is fully resealed, this should be included in the usability test by turning the package upside down. These values should be treated with the same weight as the other five basic errors in order to get a fair *accuracy* calculation. When using this approach, another problem arises. How should packages that can be closed, but still not fully resealed, be judged? One idea would be to mark an error for these packages in the *did not close* column, while marking both *difficulties closing* and *did not close* for those packages that cannot even be semi-closed. These predicaments cannot be answered by reasoning; it must be determined by firmly establishing the purpose of the usability test method and how it is supposed to be used. Our suggestion is, though, to use the approach explained above. The formula will look as follows:

$$Accuracy = \max\left(0; 1 - \left(\frac{\sum_{j=1}^{6} TNE_j}{6 \cdot TNP} + \left(\frac{TNE_7}{TNP}\right)^{\frac{1}{2}}\right)\right)$$

where

TNP = Total Number of Participants

TNE$_j$ = Total Number of Error *j* (1 = Open & spill, 2 = Pour & spill, 3 = Drip, 4 = Difficulties closing (*misthreaded* cap), 5 = Difficulties opening, 6 = Did not close, 7 = Did not open)

This solution should be complemented with a *traffic light* which clearly shows if a low *accuracy* value is due to bad opening or closing mechanism (see Figure 5.5). Green light indicates no problems at all, yellow light means that at least one of the participants had some difficulties with opening or closing and red light means that at least one participant did not manage to open or close the package. This provides the *dashboard report sheet* with enhanced visibility as the reader of the report will be attentive of possible difficulties.

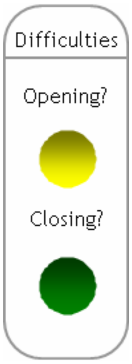


Figure 5.5 *Traffic Light for Indication of Opening or Closing Difficulties*

5.2.2 Efficiency

Even though we believe that the present method of calculating the *efficiency* provides values that describe the package's *efficiency* in a good way, we have a suggestion as how to improve the method further. This can be made by clarification of the time distribution between different tasks.

When studying the participants' behavior during the usability tests, we noticed that the time distribution varied between different packages. Some of the packages were difficult to open or close, while others were difficult to pour or drink from. Therefore, we decided to add another dimension for measuring the *efficiency*, namely the drinking and pouring time. These values where handled in the same way as the total handling time in the present method, and the *efficiency* values where calculated accordingly. This has, however, proven to be more usable in the case of pouring tests rather than drinking tests because of the very small variation between the drinking times for different packages.

Our solution is therefore to use this data only for comparison in the *handling time* chart, seen in Figure 5.6 for a drinking test package, and not to include it in the *efficiency* value.

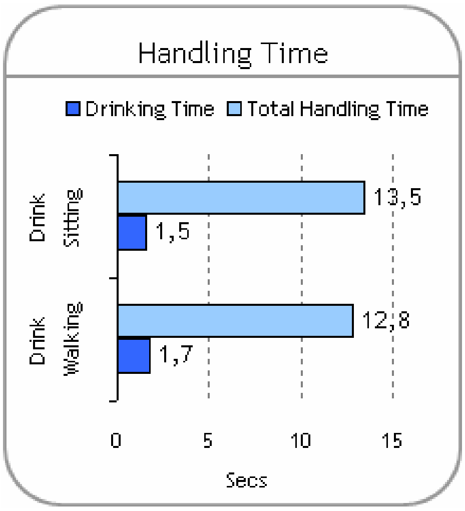


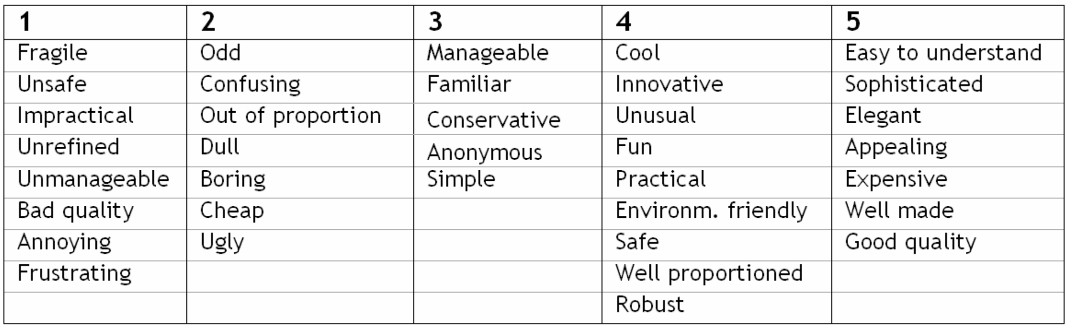
***Figure** 5.6 Drinking Time and Total Handling Time*

5.2.3 Satisfaction

The usability *satisfaction* is closely linked to the *word choices* made by the participants. We believe that it would be interesting to calculate the *satisfaction* with data from the *word choice* list as the basis, in order to examine if the two methods will render totally different results or not.

5.2.3.1 Word Choice List

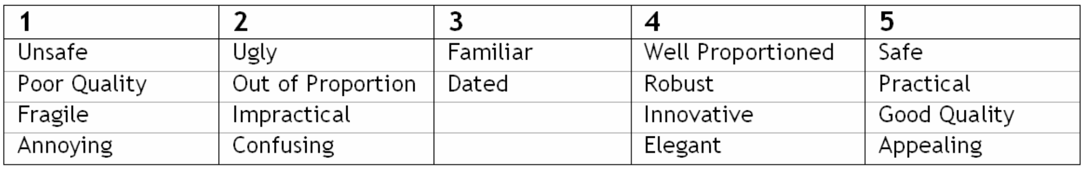
In the present method, the *word choices* made by the participants are not in any way included in the calculation of the *satisfaction*. These words are only mentioned as statistics at the bottom of the *dashboard report sheet*. In order to be able to transform the words into a *satisfaction* score, we have divided them into five groups, with every word being worth one to five points. One point equals a low usability score and five points equals a high usability score. When the participants choose a word from the randomized word list, they are unknowingly grading their attitude towards the product. The original *word choice list* had 36 words to choose from, many of which had the same essence. When grading these words, the scores got unevenly distributed as seen in Figure 5.7 (the grading is done by the authors).



1	2	3	4	5
Fragile	Odd	Manageable	Cool	Easy to understand
Unsafe	Confusing	Familiar	Innovative	Sophisticated
Impractical	Out of proportion	Conservative	Unusual	Elegant
Unrefined	Dull	Anonymous	Fun	Appealing
Unmanageable	Boring	Simple	Practical	Expensive
Bad quality	Cheap		Environm. friendly	Well made
Annoying	Ugly		Safe	Good quality
Frustrating			Well proportioned	
			Robust	

***Figure** 5.7 Original Word Choice List with Grades*

It is far easier for the test participants if the *word choice list* is reduced by half and words that are difficult to separate from one another are removed. The number of words to choose is reduced to three, in order to simplify for the participants. In the usability tests conducted by us, we used the following words which were given a score according to Figure 5.8. The words were obviously scrambled in the *word choice list* handed out to the participants (seen in Appendix C) so they would not be aware of the scoring.



1	2	3	4	5
Unsafe	Ugly	Familiar	Well Proportioned	Safe
Poor Quality	Out of Proportion	Dated	Robust	Practical
Fragile	Impractical		Innovative	Good Quality
Annoying	Confusing		Elegant	Appealing

Figure 5.8 *Reduced Word Choice List with Grades*

The total *word choice* score is then compared to the highest possible *word choice* score in order to receive a *satisfaction* value. This is a promising solution, but as the chosen words already are presented in the *word choice chart,* using the present method when calculating the *satisfaction* would bring yet another dimension to the usability results. Nevertheless, the *word choice list* method can be used in order to generate a *word choice* mean value that facilitates comparison with competitors. Due to the lack of space to present this data separately in the *dashboard,* a simplified table, as seen in Figure 5.9, can be inserted in the *word choice chart*.

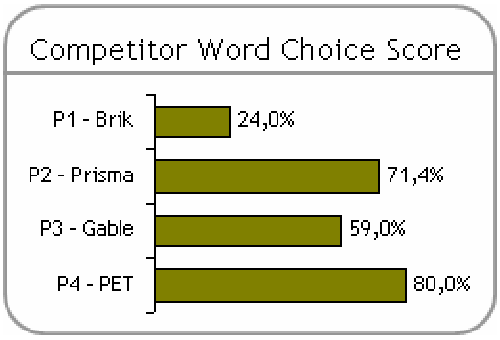


Figure 5.9 *Simplified Table for Presentation of Word Choice Competitor Comparison*

The actual scoring of the words can be discussed further as they can be interpreted differently depending on in what language they are presented. Our scoring of the words is based on the corresponding Swedish interpretation. Moreover, the available words to choose from can also be discussed. Most test participants expressed that the words found in the reduced *word choice list* were sufficient to describe their conception of the package and that they did not miss any particular word. The few deviant comments concerned the lack of words connected to the issue of environmental friendliness and some elderly test participants expressed difficulties understanding the meaning of certain words (for instance *innovative*).

5.2.3.2 Word Choice Chart

We believe that the present *word choice chart* could be greatly improved and easier to read and interpret. Our solution is to use the *word choice scores* (as described in Section 5.2.3.1) and color code them according to their positive (green), neutral (yellow) or negative (red) significance.

The revised *word choice list* contains 18 words and, as stated before, the participants are asked to choose the three most descriptive words. This way of quantifying the *satisfaction* encompasses the exclusion of the benchmarking feature of the present *dashboard* from the *word choice chart*, for greater visibility, as seen in Figure 5.10 (compared to Figure 4.3).

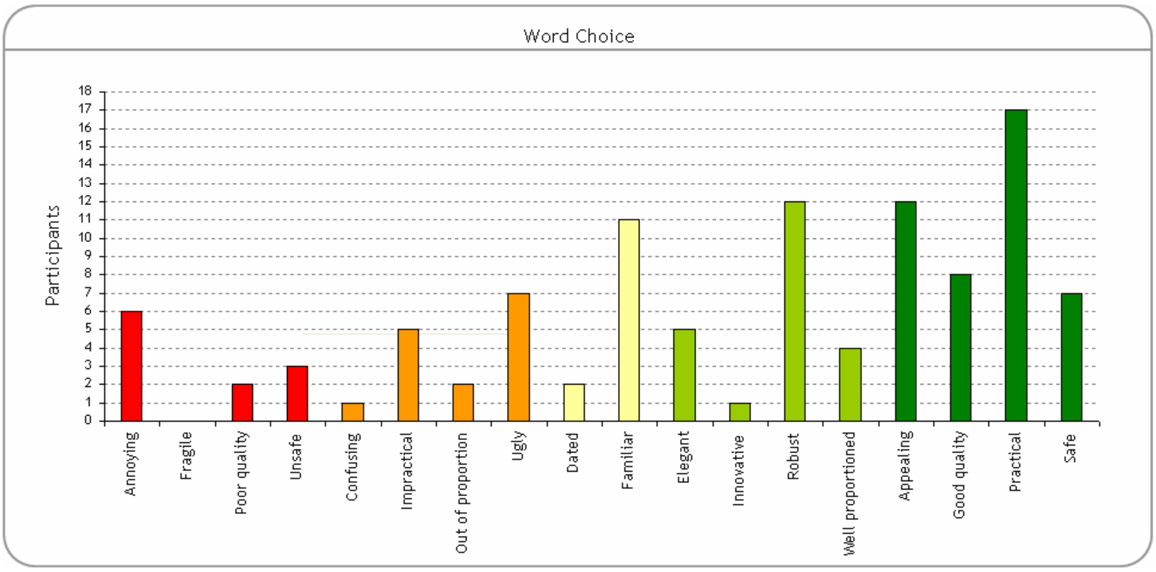


Figure 5.10 *Color Coded Word Choices*

5.2.4 Cognitive Load

The test method from the Package Company is based on the ISO definition of usability. In order to expand the usability measure, the REAL model can be used, which includes the *learnability* measure. We have chosen to call this new measure *cognitive load,* as seen in Figure 5.11, because this name is more descriptive of its purpose. The REAL model has the same basic features as the ISO definition but it also measures how the usability result is influenced by the conceptual design of the product. The model was further described in Section 3.1.3.

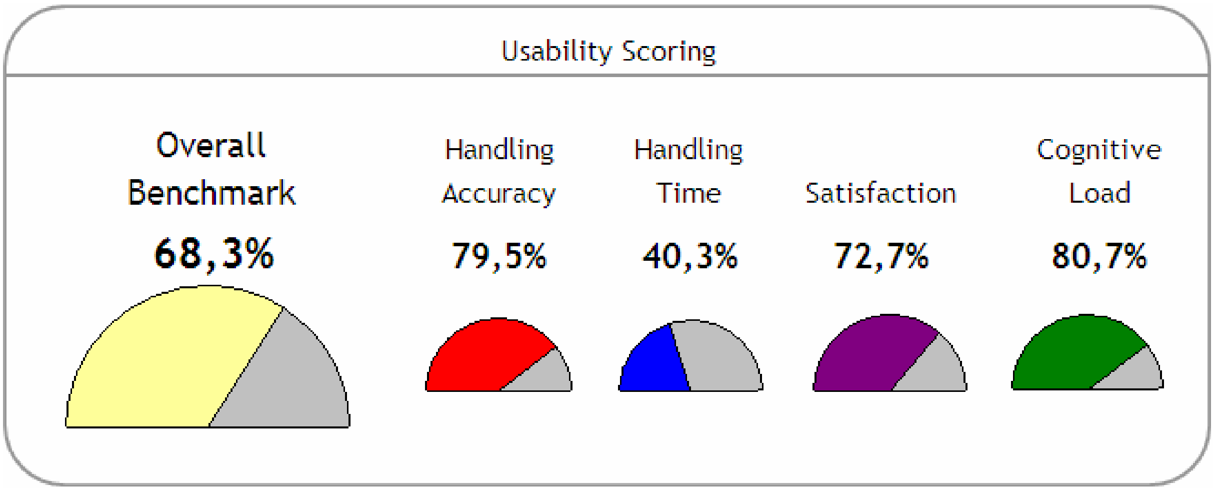


Figure 5.11 *Usability Scoring According to the REAL Model*

It is important to measure the *cognitive load* of a product in order to understand if the product's usability score is at a high level as a result of training and repeated usage or if it is a result of an engineering and design achievement.

The *cognitive load* of a package can be quantified using the comparison between the *first impression questionnaire* and *handling questionnaire* (see Appendix A). This comparison measures the package's conceptual design, which is a description of how well the appearance of the package matches how it actually functions. If the difference between the *first impression questionnaire* and the *handling questionnaire* is negligible, it proves that the package's conceptual design is excellent and that the package therefore has a high level of *cognitive load*.

Our formula for determining the *cognitive load* value looks as follows:

$$\textit{Cognitive load} = \sum_{Q=1}^{5}\left(\frac{MD\cdot TNP - \sum_{PIN=1}^{TNP}\left|FIQ_{PIN} - HQ_{PIN}\right|}{MD\cdot TNP}\right)\cdot\frac{1}{5}$$

where

Q =	Question 1-5 (1=Hold, 2=Open, 3=Drink/Pour, 4=Close, 5=Use)
MD =	Maximum Difference between the questionnaire answers (= 5-1 = 4)
TNP =	Total Number of Participants
PIN =	Participant Identification Number
FIQ =	First Impression Questionnaire
HQ =	Handling Questionnaire

However, we do not believe that this value should be included in the *overall benchmark* as its precision is correlated with how much experience the user has from the package in the past, and accordingly, at least in our tests, all the packages had high *cognitive load* values, even the poor ones. Adding this value to the *overall benchmark* may distort the usability score rather than add any value to it. Nevertheless it has big advantages when trying out new designs and should therefore be present in the *dashboard report sheet* under *comparison with competitors,* as seen in Figure 5.12.

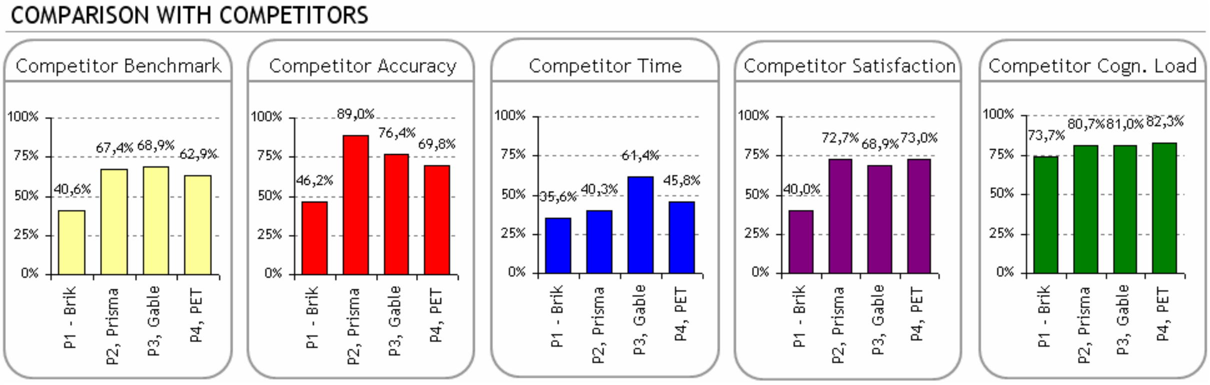


Figure 5.12 *Competitor Benchmark*

5.3 Data Processing

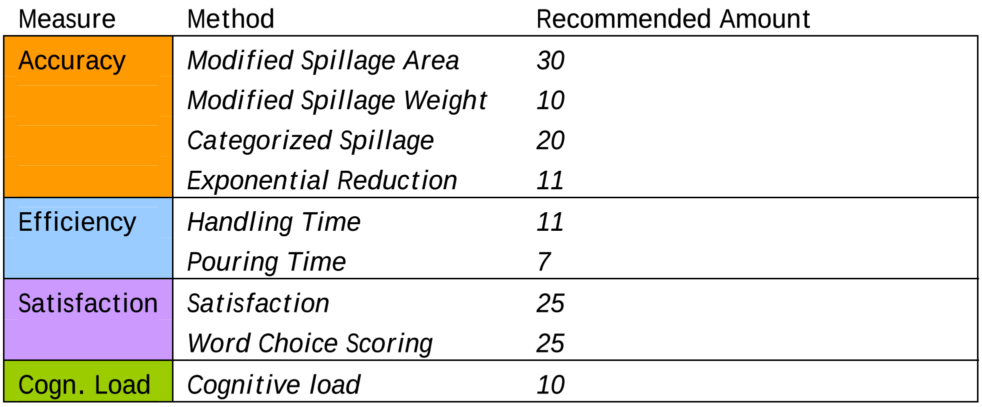
Using the methods described in Section 4.4, we processed the data gathered during the usability tests with the intention to determine how many participants are required to get reliable results. The following subchapters will present what these methods led to and how we interpret the findings.

5.3.1 Chart Exploration

Firstly, we used the simple chart exploration technique. Charts showing the outcome for all the different usability measures can be found in Appendix I, due to the huge amount. In this chapter we will confine ourselves to only present what these charts recommend when it comes to the amount of test participants.

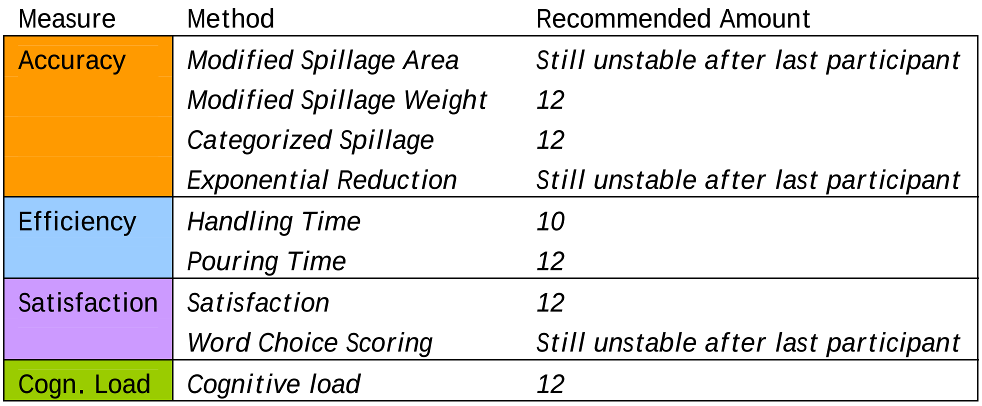
5.3.1.1 Recommended Amounts

As seen in the charts, the curves sometimes oscillate heavily and a result we call stable may very well be considered unstable by others and vice versa. Further, curves that seem to be stable may begin to oscillate again when adding more test participants to the input data of the charts. The recommended amounts, shown in Figures 5.13 to 5.16, are an indication of when all four curves in the charts (representing the different packages) appear to have stabilized according to us.



Measure	Method	Recommended Amount
Accuracy	*Modified Spillage Area*	*30*
	Modified Spillage Weight	*10*
	Categorized Spillage	*20*
	Exponential Reduction	*11*
Efficiency	*Handling Time*	*11*
	Pouring Time	*7*
Satisfaction	*Satisfaction*	*25*
	Word Choice Scoring	*25*
Cogn. Load	*Cognitive load*	*10*

Figure 5.13 *Recommendations, Chart Exploration, Pouring Test, Students*



Measure	Method	Recommended Amount
Accuracy	*Modified Spillage Area*	*Still unstable after last participant*
	Modified Spillage Weight	*12*
	Categorized Spillage	*12*
	Exponential Reduction	*Still unstable after last participant*
Efficiency	*Handling Time*	*10*
	Pouring Time	*12*
Satisfaction	*Satisfaction*	*12*
	Word Choice Scoring	*Still unstable after last participant*
Cogn. Load	*Cognitive load*	*12*

Figure 5.14 *Recommendations, Chart Exploration, Pouring Test, Elderly People*

Measure	Method	Recommended Amount
Accuracy	*Present Method*	*25*
	Modified Present Method	*30*
	Exponential Reduction	*6*
Efficiency	*Handling Time*	*19*
	Drinking Time	*15*
Satisfaction	*Satisfaction*	*15*
	Word Choice Scoring	*20*
Cogn. Load	*Cognitive load*	*10*

***Figure* 5.15** *Recommendations, Chart Exploration, Drinking Test, Students*

Measure	Method	Recommended Amount
Accuracy	*Present Method*	*12*
	Modified Present Method	*Still unstable after last participant*
	Exponential Reduction	*10*
Efficiency	*Handling Time*	*13*
	Drinking Time	*12*
Satisfaction	*Satisfaction*	*10*
	Word Choice Scoring	*13*
Cogn. Load	*Cognitive load*	*10*

***Figure* 5.16** *Recommendations, Chart Exploration, Drinking Test, Elderly People*

5.3.1.2 Comments

Using the calculation methods proposed by the authors and the results for the students in the summary above, it seems like 20 to 25 participants definitely is enough to get fair results. One must bear in mind that the *Exponential Reduction* method implies severe reductions in the *accuracy* value if a participant cannot manage to open the package at all. As a consequence, the resulting curve in the chart may take a large leap. However, if disregarding these leaps for a second, it is most obvious that the curves are very straight from the very beginning.

In a similar manner, the curves in the *efficiency* charts tend to take quite big leaps at times, although they are rather straight otherwise. This is caused by the fact that the calculations are based on different benchmark values as different amounts of participants are included and the benchmark value will inevitably go down with more participants. This seemingly big problem is however only artificial, as it does not have any impact on the overall *efficiency* value due to the fact that this value is calculated using the same benchmark value for all participants. In the future, when more tests have been undertaken, the benchmark value will be more accurate and this effect will not be as evident in the statistical calculations as in the present case. Consequently, when analyzing the *efficiency* curves it is important to bear this complex of problems in mind and disregard the big leaps in the charts.

The *satisfaction* value seems to be the critical factor regarding the recommended amount of test participants, because it stabilizes quite slowly as seen in Appendix I. This is due to the fact that the *satisfaction* is the participant’s subjective notion and the large deviation is therefore not surprising. The *cognitive load*, on the other hand, does not oscillate very much, as all the participants are familiar with most of the packages and the properties of these do not surprise the participants. This leads to similar answers in both the *first impression* and the *handling questionnaire*, and consequently in high *cognitive load* values for all participants.

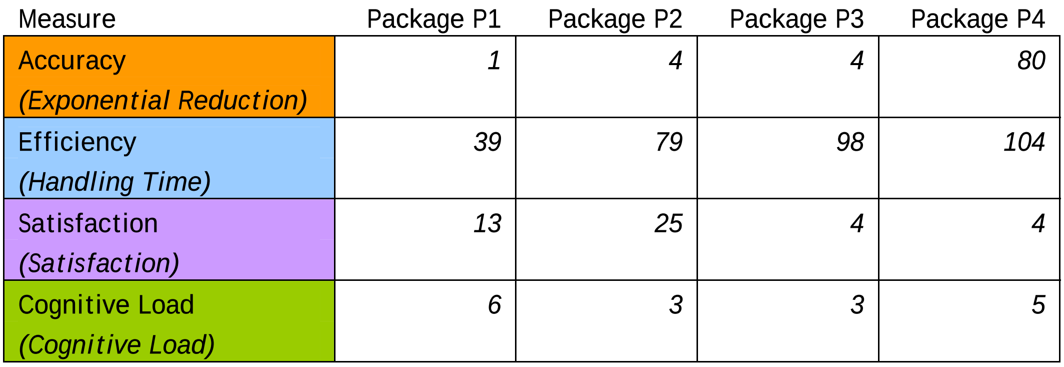
The results of the elderly people indicate that in some of the cases it is not necessary to use much more than 10 participants. However, throwing an eye at the corresponding charts for the students raises a warning flag, because even if all the curves are straight for a while, some of them may still start to oscillate later on when additional participants are added. This is most likely what would happen to the results of the elderly people.

5.3.2 Statistical Calculation

Following the chart exploration, we applied the statistical calculation method. We emphasize that this research was only made for the methods of usability measurement calculations we found the most interesting and suitable. The reason to this is the iterative process of this thesis; even though this subchapter is immediately after the previous one, they were written with several weeks in between. During this time, a selection (which was mentioned in Section 5.2.1) was made regarding which methods to go further with.

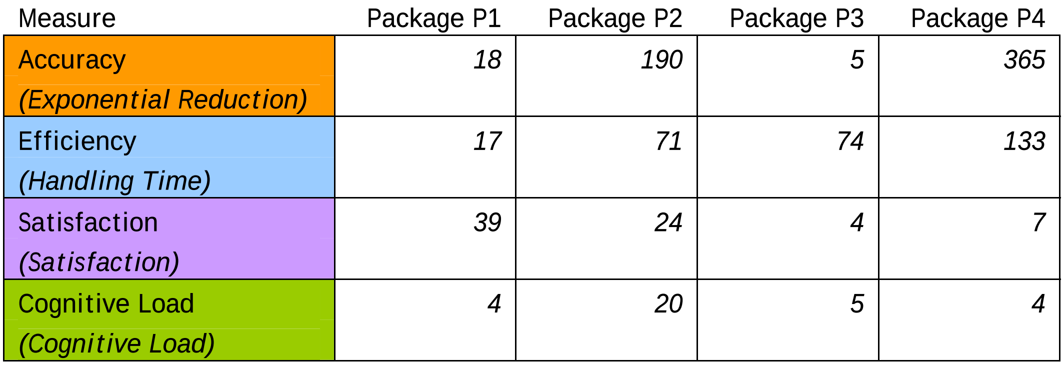
5.3.2.1 Recommended Amounts

The method was thoroughly explained in Section 4.4.2 and the results it gave, presuming we are interested in a level of confidence of 95% and a delta value of 5 percentage units, are shown in Figures 5.17 to 5.20:



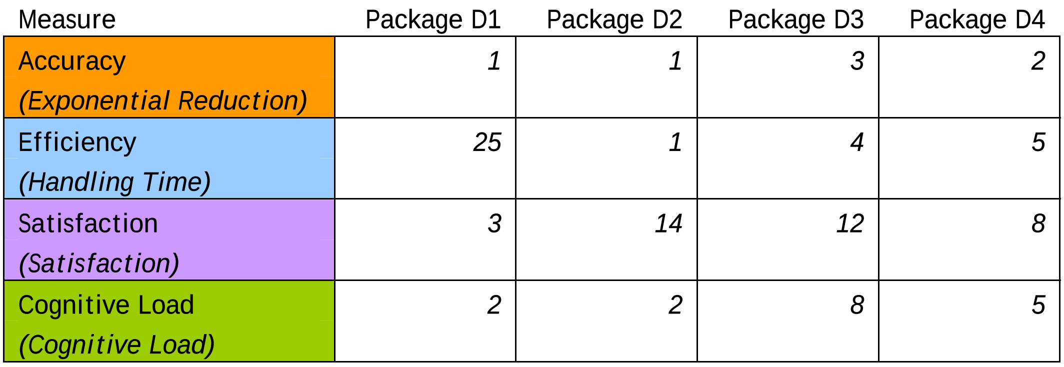
Measure	Package P1	Package P2	Package P3	Package P4
Accuracy *(Exponential Reduction)*	*1*	*4*	*4*	*80*
Efficiency *(Handling Time)*	*39*	*79*	*98*	*104*
Satisfaction *(Satisfaction)*	*13*	*25*	*4*	*4*
Cognitive Load *(Cognitive Load)*	*6*	*3*	*3*	*5*

Figure 5.17 *Recommendations, Statistical Calculation, Pouring Test, Students*



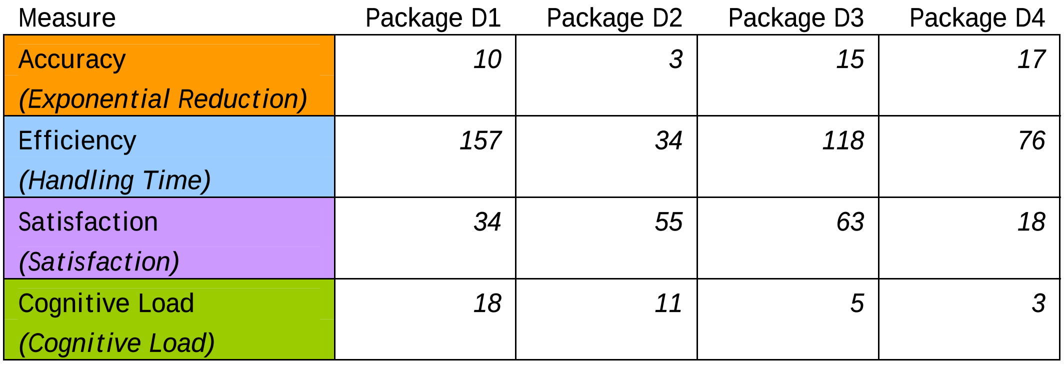
Measure	Package P1	Package P2	Package P3	Package P4
Accuracy *(Exponential Reduction)*	*18*	*190*	*5*	*365*
Efficiency *(Handling Time)*	*17*	*71*	*74*	*133*
Satisfaction *(Satisfaction)*	*39*	*24*	*4*	*7*
Cognitive Load *(Cognitive Load)*	*4*	*20*	*5*	*4*

Figure 5.18 *Recommendations, Statistical Calculation, Pouring Test, Elderly People*



Measure	Package D1	Package D2	Package D3	Package D4
Accuracy *(Exponential Reduction)*	*1*	*1*	*3*	*2*
Efficiency *(Handling Time)*	*25*	*1*	*4*	*5*
Satisfaction *(Satisfaction)*	*3*	*14*	*12*	*8*
Cognitive Load *(Cognitive Load)*	*2*	*2*	*8*	*5*

Figure 5.19 *Recommendations, Statistical Calculation, Drinking Test, Students*



Measure	Package D1	Package D2	Package D3	Package D4
Accuracy *(Exponential Reduction)*	*10*	*3*	*15*	*17*
Efficiency *(Handling Time)*	*157*	*34*	*118*	*76*
Satisfaction *(Satisfaction)*	*34*	*55*	*63*	*18*
Cognitive Load *(Cognitive Load)*	*18*	*11*	*5*	*3*

Figure 5.20 *Recommendations, Statistical Calculation, Drinking Test, Elderly People*

5.3.2.2 Comments

In general, the recommendation of using 20-25 test participants from the chart exploration method seems to coincide fairly well with a majority of the values presented above, although with some exceptions. It is mainly the *efficiency* values, and sometimes the *accuracy* values as well, that differ from the earlier recommendation.

The high values in some cases of the *accuracy* method depend on the very large reduction in the score whenever *did not open* occurs. This in turn generates a much higher standard deviation and following that, the large amount of recommended participants. As earlier mentioned and seen in Appendix I, the *accuracy* curves are rather straight apart from these major leaps and this is well reflected by the remaining values in the summary above.

The underlying reason to the rather large amounts of recommended participants for the *efficiency* method is, just like for the chart exploration technique, that the value is constantly recalculated as a new best time is set and, thus, the benchmark value is changed. This leads to more volatility, causing a higher standard deviation and, in the end, to the vast amounts of participants recommended.

5.3.3 Wilcoxon Signed Rank Sum Test

A way to make sure whether the results from the statistical calculations are reasonable or not is, as mentioned in Section 3.3.3, to perform a Wilcoxon signed rank sum test. This is done in order to determine whether the large standard deviations are caused by deep dips in the cumulative average that sometimes appear or if there actually is a large natural volatility responsible for the outcome.

5.3.3.1 Results

The *u* values (derived in Section 3.3.3) for the tests are presented in Figures 5.21 to 5.24 below, while the data used to calculate these values and graphs showing the relation between the cumulative averages and the individual contributions to the cumulative averages can be found in Appendix J. Similarly to the preceding subchapter, we only provide the *u* values for the single most interesting calculation method of every usability measure.

Measure	Package P1	Package P2	Package P3	Package P4
Accuracy	*0.15*	*0.54*	*0.25*	*3.91*
Efficiency	*1.20*	*1.39*	*1.56*	*1.05*
Satisfaction	*0.01*	*0.36*	*0.08*	*0.61*
Cognitive Load	*0.28*	*0.15*	*1.23*	*0.90*

Figure 5.21 *U Values for Pouring Test, Students*

Measure	Package P1	Package P2	Package P3	Package P4
Accuracy	*0.66*	*2.17*	*0.18*	*0.72*
Efficiency	*0.35*	*0.35*	*0.47*	*0.16*
Satisfaction	*0.28*	*0.09*	*0.38*	*0.09*
Cognitive Load	*0.41*	*0.16*	*0.47*	*0.13*

Figure 5.22 *U Values for Pouring Test, Elderly People*

Measure	Package D1	Package D2	Package D3	Package D4
Accuracy	*1.02*	*1.62*	*0.02*	*0.11*
Efficiency	*0.25*	*0.45*	*0.50*	*0.70*
Satisfaction	*0.54*	*0.39*	*0.14*	*0.38*
Cognitive Load	*0.15*	*0.57*	*0.54*	*0.66*

Figure 5.23 *U Values for Drinking Test, Students*

Measure	Package D1	Package D2	Package D3	Package D4
Accuracy	*0.22*	*0.22*	*0.16*	*1.85*
Efficiency	*0.03*	*0.28*	*0.09*	*0.03*
Satisfaction	*0.66*	*0.47*	*0.16*	*0.22*
Cognitive Load	*0.09*	*0.09*	*0.22*	*0.28*

Figure 5.24 *U Values for Drinking Test, Elderly People*

5.3.3.2 Comments

The *u* values exceeding 1.96 actually indicate that the hypothesis of symmetry around the cumulative mean can be rejected in the corresponding cases with a level of confidence of 95%. Yet, we are merely interested in whether the *u* value is high or not and how this relates to the amounts of participants recommended by the statistical calculations. The above-mentioned limit is, however, a good objective of what values are to be regarded as very high.

We start by analyzing the values for the *pouring test* with students. There are five recommended amounts that exceed 25, ranging from 39 to 104. Their corresponding *u* values in the tables above are all well above 1.00, reaching all the way up to 3.91, far beyond the limit of possible asymmetry. This implies that the large recommended amounts can be neglected.

Regarding the drinking test, there are no recommendations exceeding 25 participants, but some of the *u* values are still quite large. The reason for this can be found by studying the table used to determine the *u* value in the first place. Take the *accuracy* for the D2 package as an example. As seen in Appendix J (page XXVII), many of the differences are the same and accordingly, there are not more than five rank intervals, which in turn makes them very wide. A *u* value calculated with no more than five wide intervals should not be taken too seriously as it is just as hard to get a reasonable value that way as it is to drive a car smoothly using nothing but full throttle and full brake.

When it comes to the elderly people tests, it seems like there are several factors causing the high recommendations. First of all, there was a larger span in the results, as the group of elderly people is less homogeneous than the student group. Some elderly people managed to perform the tasks just as good as the students, whereas other had severe difficulties. There are many possible reasons to this, but the large variation between the results remains nonetheless a fact, although the curves in the charts seen in Appendix I actually seem to stabilize rather quickly after a lot of initial oscillation. Another reason for the recommendations is naturally the small amount of test participants, making the prognostication less reliable.

5.3.4 Influence of the Test Monitor

The gathered usability data for the tests can be found in Appendix K. With the assistance of these tables, it is easy to compare the difference in results for the participants of the two *test monitors*.

Significant differences in the outcome would be an indication of the fact that the *test monitor* actually has an impact, which of course is most unwanted. However, following the *test monitor* guidelines that were set up guaranteed that all tests were conducted under equal circumstances, even if the personalities of the *tests monitors* differ quite a lot in some aspects.

Comparing the *overall benchmark* values for all 35 test participants makes it clear that the difference is negligible, which is shown in Figure 5.25 below (values in percent).

Test Monitor	P1	P2	P3	P4	D1	D2	D3	D4
Erik	*38.7*	*64.9*	*66.5*	*64.8*	*63.0*	*47.4*	*62.0*	*72.1*
Tomasz	*42.6*	*69.9*	*71.5*	*62.7*	*67.3*	*53.3*	*67.7*	*75.4*

Figure 5.25 *Overall Benchmarks for the Two Test Monitors*

If we investigate the *accuracy* value (calculated using the *Exponential Reduction* method), which after all is the single most interesting measure of the test method, we see that the differences between the corresponding values are very small. This is presented in Figure 5.26.

Test Monitor	P1	P2	P3	P4	D1	D2	D3	D4
Erik	*46.3*	*89.4*	*75.9*	*77.1*	*62.0*	*62.3*	*82.1*	*88.3*
Tomasz	*46.1*	*88.7*	*77.0*	*67.4*	*64.2*	*64.4*	*81.0*	*86.9*

Figure 5.26 *Accuracy Values for the Two Test Monitors*

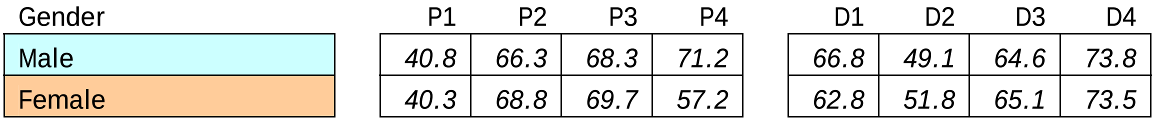
The only major difference apply for the P4 package and the reason to this is that the *accuracy* method used, namely the *Exponential Reduction*, lowers the value significantly whenever a participant cannot manage to open a package. This is exactly what happened twice to Tomasz' participants and only once to Erik's, hence the large gap in the *accuracy* value.

The *efficiency* values are not presented as they give unfair results. This is because of the fact that in order to calculate the values for one of the *test monitors*, all the input data for the participants of the other *test monitor* was removed from the data material by the authors. Consequently, the benchmark value for the *efficiency* measure was unexceptionally taken from the group of participants being investigated and not the whole population of 35 participants (or 14 for the elderly people test). On the other hand, the *efficiency* values calculated this way provide an understanding of the time distribution within the investigated group.

Patterns similar to that of the *accuracy* are found if investigating the *cognitive load* measure. They will not be presented here, but the reader can easily verify it by looking at Appendix K. There are some differences in the *satisfaction* values between the participants of the two *test monitors*, even though they are not large enough to suspect any influence of the *test monitor's* behavior.

5.3.5 Influence of the Test Participants

A possible reason to the fact that Tomasz' participants performed slightly better in several of the cases might be the differences in demographics, as a majority of Erik's participants were male, whereas a majority of Tomasz' participants were female. In order to determine whether the gender of the participant was of any relevance, comparisons equal to those for the influence of the *test monitor* were made. The one in Figure 5.27 compares the *overall benchmark* for all male participants with those for the opposite gender.



Gender	P1	P2	P3	P4	D1	D2	D3	D4
Male	*40.8*	*66.3*	*68.3*	*71.2*	*66.8*	*49.1*	*64.6*	*73.8*
Female	*40.3*	*68.8*	*69.7*	*57.2*	*62.8*	*51.8*	*65.1*	*73.5*

Figure 5.27 *Overall Benchmarks for Male and Female Student Participants*

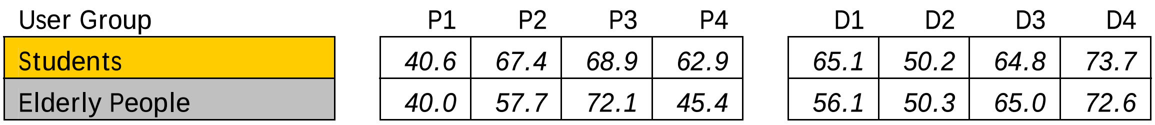
As seen in the figure above, the results are very similar between the genders, except for the P4 package. This is, once again, due to the *accuracy* measure (*Exponential Reduction*) and the large reduction for packages that could not be opened. As all three cases where this occurred involved this particular package and a female participant, the large gap is not surprising at all. If only considering the *accuracy* values for the P4 package, they were 90.8% for the men and 58.3% for the women.

The amount of male participants exceeded that of the female ones, although we find this of lesser importance. It is still fairly obvious that the gender of the test participant has no substantial impact on the outcome. However, there are some interesting things that can be seen if studying the data material thoroughly, like the fact that, generally, the male participants drank and poured with better accuracy, yet the female participants still expressed a greater satisfaction with the products. Likewise, the women had higher *cognitive load* values than the men.

Unfortunately, both test monitors had a skew mix of male and female test participants. We have already made clear that this did not affect the result, but we can try to determine if there were any differences between the male participants of the two test monitors, and correspondingly for the female ones. Comparison of the results indicate that there was no significant diversity between the male participants of the two test monitors, although the equivalent gap is definitely larger for the female participants, with Tomasz' female participants achieving substantially better values, both regarding *accuracy* and *satisfaction*. This may very well explain why Tomasz' test participants had slightly higher *overall benchmark* values than Erik's.

The last thing we investigated was equal amounts of male and female participants for both the test monitors. This yielded slightly better values for Tomasz' participants, but the fact remains that the gender does not affect the result in any particular way. It is still important to remember that we only conducted these tests on 35 participants and this may be a too small basis to draw any conclusions from.

The 35 student participants were all within the age range of 22 to 29 years old. We find this far too narrow to start analyzing the data material for possible age-related tendencies. Instead of this, it may be of some interest to compare the group of students to the elderly people that also underwent the test. This group is even smaller, consisting of only 14 persons between the age of 54 and 80, but a comparison still gives a hint on whether all age categories manage to operate the packages equally well. Figure 5.28 shows how the *overall benchmark* results look like:



User Group	P1	P2	P3	P4	D1	D2	D3	D4
Students	40.6	67.4	68.9	62.9	65.1	50.2	64.8	73.7
Elderly People	40.0	57.7	72.1	45.4	56.1	50.3	65.0	72.6

Figure 5.28 *Overall Benchmark Values for Students and Elderly People*

As can be seen in the figure above, the elderly people managed rather well with most of the packages (we will not go into detail with what sort of packages were problematic to the elderly people or discuss possible reasons to this). The major difference between the students and the elderly is that the latter group had widespread difficulties to open the packages. However, once the packages were opened, the elderly managed to pour just as accurate as the students, if not even better, although they did it at a slower pace. Another predicament is that it proved to be very difficult to make the elderly differentiate the two pouring subtasks. The outcome of this is that results of the quick test differ insignificantly from those of the careful test.

Finally, we compared the results of the elderly male participants with those of the elderly female ones. The former group tends to achieve slightly better results, although the difference is too small to pay any particular attention to. The results for the elderly people can also be found in Appendix K.

6 Conclusions and Recommendations

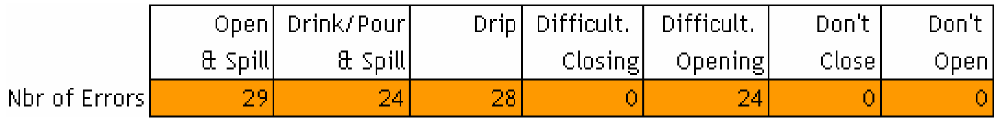
This chapter propounds the conclusions made in the analysis and stipulates the most appropriate method of calculating the usability value. Furthermore, the findings of the statistical examinations will be summarized and clarified.

6.1 Package Usability Test Method

A problem with the present usability test method is that it has different calculation methods depending on what kind of package is tested. Our objective has been to develop a test method that can be implemented on all kinds of packages in both drinking and pouring usability tests, in order to simplify the learning process for the *test monitor*. Some of the data collected during the usability tests has proven to be rather unimportant and should preferably be excluded from the test method. The methods we have chosen to use in the usability test method and their required data are explained below.

6.1.1 Accuracy

We have chosen to use a usability test method that calculates the *accuracy* based on how many errors of each of the seven error types shown in Figure 6.1 are made by the participants. This method is called *Exponential Reduction* in the analysis.



	Open & Spill	Drink/Pour & Spill	Drip	Difficult. Closing	Difficult. Opening	Don't Close	Don't Open
Nbr of Errors	29	24	28	0	24	0	0

Figure 6.1 *The Seven Possible Accuracy Errors*

The *accuracy* is calculated by comparing the total amount of errors in the first six columns with the maximum amount of possible errors. The error called *difficulties closing* also includes occurrence of *misthreading* of the cap. We believe that *did not open* is a far more crucial error than the other ones, which is why we have chosen to make an exponential reduction for every error of that kind. The formula looks as follows:

$$Accuracy = \max\left(0; 1 - \left(\frac{\sum_{j=1}^{6} TNE_j}{6 \cdot TNP} + \left(\frac{TNE_7}{TNP}\right)^{\frac{1}{2}}\right)\right)$$

where

TNP =	Total Number of Participants
TNE_j =	Total Number of Error j (1 = Open & spill, 2 = Pour & spill, 3 = Drip, 4 = Difficulties closing (*misthreaded* cap), 5 = Difficulties opening, 6 = Did not close, 7 = Did not open)

We have discovered that post-test activities like measuring and weighing the spillage and measuring values like Δ*TPC* and Δ*TPQ* (mentioned in Section 5.1.1) are extremely time-consuming with little or no gain in the final result. This method makes it possible to exclude these measurements from the test method without any loss of valuable information.

In order to include a fair judgment of packages that cannot be fully resealed, we have decided that all packages should be turned upside down and the occurrence of leakage should be noted by the *test monitor*. If the leakage is obvious this is naturally not necessary. A package that that can be closed but is not resealed properly should be marked with an error in the *did not close* column for all participants. A package that cannot be closed at all should be marked with errors in both the *difficulties closing* and the *did not close* column for all participants. This manner provides a fair solution and prevents any package from getting advantages by not applying to a certain type of error.

We have chosen to include a *traffic light* in the *dashboard report sheet*, as seen in Appendix H, to indicate difficulties with the opening or closing mechanisms. Green light indicates no problems at all, yellow light means that at least one of the participants had some difficulties with opening or closing and red light means that at least one participant did not manage to open or close the package. This *traffic light* highlights the errors in an explicit way and will immediately catch the attention of the reader.

6.1.2 Efficiency

The *efficiency* measurement is not changed from the present method. The participant's total handling times are compared to the fastest participant in the test of the package with the shortest mean time. In the case of packages to drink from this is calculated according to the following formula (obviously, a similar corresponding formula applies to the packages of the pouring test):

$$Efficiency = \frac{\left(\frac{fastest\ handling\ time\ sitting}{mean\ handling\ time\ sitting}\right) + \left(\frac{fastest\ handling\ time\ walking}{mean\ handling\ time\ walking}\right)}{2}$$

Packages that cannot be closed have a great advantage, opposed to packages that have a cap, as their total handling times will inevitable be shorter. This is problematic as the fastest times are used as benchmarks for the other packages and this leads to inequitable conditions between different packages. However, the possibility of closing the package is not a prerequisite for using the usability test method and this advantage is taken care of with a lowered *accuracy* score.

We have chosen to include the pouring/drinking time in the handling time graphs (as a complement to the total handling time), as seen in Figure 5.6, in order to give a better overview of the time distribution of the different parts of the pouring/drinking process.

A predicament with the *efficiency* calculation method is that the new way of calculating the *accuracy* has made the *efficiency* value less correlated with the *accuracy*. The current *efficiency* value is based on a comparison between the fastest handling time, for the package with fastest mean time, and all other participants' times. A problem with this approach is however that an unnaturally competitive and fast participant can spoil all other packages' *efficiency* values. This might be due to excessive neglect of the amount of spillage, ignoring to drink a sufficient amount of liquid or not filling up the glasses properly. The measures of the area and the weight of the spillage, as well as ΔTPC and ΔTPQ, have, as already mentioned, been excluded in the new *accuracy* calculations because they have proven to be of none or limited interest. However, this makes it possible to bring forth a high *efficiency* value without any reprimands in the other usability scores. We have not yet experienced this possible weakness in the method, but as more tests are performed, this is increasingly likely to occur. This must therefore be an issue for further investigation and should be seriously considered. An alternative solution to evade this problem would be to compare all participants' handling times with the fastest package's mean handling time, although this will always generate an *efficiency* value of 100 % for the package with the best *efficiency*.

6.1.3 Satisfaction

We believe that the *satisfaction* should be based on the *handling questionnaire*, seen in Appendix A, according to the present method. Each participant's experienced *satisfaction* is compared to the maximum possible score. The formula looks as follows:

$$Satisfaction = \frac{MV_{hold} + MV_{open} + MV_{drink} + MV_{close} + MV_{easy\ to\ use}}{5 \cdot 4}$$

where

$MV =$ Mean Value

We have chosen to keep this method as it generates fair results and the questions in the questionnaire are straightforward and can hardly be misinterpreted.

The *word choice list* has been reduced to include only 18 words, and the participants are asked to chose the three most descriptive. As we believe that the word choices tell much about how the participant experience the usability, we have chosen to give the words a score from one to five (the participant is not aware of the fact that the words are graded) and then presented them in a color-coded chart for better visualization, as seen in Figure 5.10. This solution includes a competitor comparison value, based on the *word choice scoring*, according to Figure 5.9. The scoring of the words may be revised, depending on how they are chosen to be interpreted.

6.1.4 Cognitive Load

We have decided to add a measure called to the usability test. The measurement of this new aspect explores if the package's physical appearance reflects its actual usability. The calculations are based on the participants' answers in the *first impression* and *handling questionnaires*, seen in Appendix A.

The formula for determining the *cognitive load* looks as follows:

$$Cognitive\ load = \sum_{Q=1}^{5} \left(\frac{MD \cdot TNP - \sum_{PIN=1}^{TNP} \left| FIQ_{PIN} - HQ_{PIN} \right|}{MD \cdot TNP} \right) \cdot \frac{1}{5}$$

where

Q =	Question 1-5 (1=Hold, 2=Open, 3=Drink/Pour, 4=Close, 5=Use)
MD =	Maximum Difference between the questionnaires (= 5-1 = 4)
TNP =	Total Number of Participants
PIN =	Participant Identification Number
FIQ =	First Impression Questionnaire
HQ =	Handling Questionnaire

We have chosen to exclude the *cognitive load* measure from the calculation of the *overall benchmark* because it had too much positive impact, as even poor packages had a high *cognitive load* value. However, this measurement can still prove to be interesting when examining a product that never has been used before and we have decided to still include the *cognitive load* value in the *competitor comparison* part of the *dashboard report sheet*.

6.2 Presentation of the Usability Score

The presentation of the usability score will be somewhat different compared to the present method, as seen if comparing Appendix G (a slightly modified version) and Appendix H. The package information box has been extended with details about the ability to reclose the package, a traffic light regarding opening and closing problems has been added and the handling time graph has been equipped with additional information in the form of the drinking/pouring time. In excess of all these changes, the competitor comparison now also includes a presentation of the *cognitive load* value for all the packages in the test. The comparison of the *word choices* has been altered so it is now inserted in the top corner and the chosen words for the current package are graded and color-coded.

In addition, the *overall benchmark*, or in other words the usability score of the package, is computed by adding up the equally weighted *accuracy*, *efficiency* and *satisfaction* values instead of the former uneven weighting.

6.3 Repeatability and Reproducibility

Another part of the authors' task was to determine the amount of test participants required to attain reliable results. Our tests show that, for the test method we propose, it should be sufficient to use 25 test participants. Overall, the results generated by the arrangement just described give a good view of the package's usability properties. Concerning the *accuracy* measure, perhaps the most important of those included in the test method, the chosen method (*Exponential Reduction*) actually requires far less participants than 25 in most cases. However, one must remember that this study has only taken eight certain packages into consideration. It is not necessary that the test results, graphs and recommendations would have looked much the same if the study would have been undertaken using a completely different set of packages.

Further, the recommendation just given is mainly based on a study using a rather narrow foundation consisting of 35 engineering students of similar ages and 14 elderly persons. This is indeed a group of people not fully representative of the entire population of beverage package consumers. Our research also shows that despite the fact that the difference between young and elderly people is not as large as one might have predicted, this difference still

exists and must not be overlooked. On the other hand, the test method is seemingly unaffected by the characteristics of the *test monitor* or the gender of the participant.
Regarding the group of elderly consumers, most of them can manage to perform the tasks of the tests, even though there is a significantly larger need to assist the participants of this segment. Further, they are often not familiar with modern packages (like the aluminum can for instance) and many packages are too difficult to open due to physical weakness, especially for elderly female participants. It also requires much more time and patience from the *test monitor* compared to students or other consumer categories. One interesting finding that appeared in the elderly segment tests is the difficulties in opening the package. Some of the packages were clearly more difficult to open than others, but the students could always, with just a few exceptions, manage to do it (even though some of them needed to use brute force). This is not the case with the elderly, as they are physically more restricted. This restriction also leads to reduced mobility, forcing the test team to perform field studies on the elderly people's terms. A consequence of this is effort and time consuming transportation of packages, cameras and all the other equipment as well as deteriorated test conditions.

It is difficult to decide what is the most important, to produce a package of poor quality that practically everyone is able to open or one that guarantees a proper sealing but may cause inconvenience to some users. Although most people would probably agree that the latter is of greater importance (in most cases, there will probably be someone able to help users with difficulties anyway), it is still interesting to perform the tests on elderly people. If nothing else, they can at least be a part of the test population as *least competent users*. The conclusion of this is that different consumer segments might be needed to fully investigate all parts of the package usability, but a more homogeneous group should be used to obtain stable results that can also be used as competitor benchmark values.

In general, it would be advisable to perform full large-scale tests on a combination of different consumer segments, particularly of different ages. Perhaps this would require more than the earlier mentioned 25 test participants, but we think that it is necessary in order to achieve an overall picture of the package usability. On the other hand, there is no need to be worried if it is hard to acquire even amounts of male and female participants, as this has not got any significant impact on the outcome.

7 Generalizations and Future Studies

This chapter is meant to give a deeper insight into the issues that we believe are not fully investigated yet and might need some further exploration or decision making. Some questions have no uniform or correct answers, but it is still important to be aware of the complexity of the problems.

7.1 Packages and Participants

The usability tests we have performed have acted as the foundation for the development of the new usability measures, but they are not necessarily representative for future package usability tests, neither at the Package Company nor in general.

Our tests were conducted with packages that had legible physical design differences. The question is how the test method will respond to packages that are very similar in design and if the method is sensitive enough to detect and propound these small differences. A related question is if the method is sufficiently sophisticated to be utilized on packages with completely different physical attributes from those we based our conclusions on. These issues can only be answered by additional usability tests.

The conclusions we have made in this thesis are based on tests performed on homogeneous groups as students, predominantly such in engineering, and elderly. A question to take in consideration is if other consumer segments would be as competitive as some of the student participants of our tests and if they would generate equal results? Our investigation shows that the variation of the test participants' results is predominantly due to personal differences rather than gender or age (the latter applies to participants within the student segment we investigated). It is still important to keep the test groups as standardized as possible when conducting tests in the future so that correct competitor comparisons can be made. A way to even out differences between participants is to add a standardized and compulsory comparison element that highlights discrepancies between different test groups. Using this complementary method, baselines for all participants will be created enabling a fair comparison of the usability score between different test sets and packages.

7.2 Package Usability Test Method

The following subchapters contain some recommendations for the future regarding the components that compose the *overall benchmark* value in the package usability test method.

7.2.1 Accuracy

When using the *accuracy* method called *Exponential Reduction,* it is very well possible to alter the exponential growth rate of *did not open* errors, if the current one proves to be deceptive. We have made the decision of using ½ as the growth rate based on our personal preferences of what could be considered as a fair reduction. If this is a reasonable reduction or not will easily be clarified by performing additional usability tests.

7.2.2 Efficiency

As explained earlier in the report, the *efficiency* measure is less correlated with the *accuracy* in the newly developed usability test method. This poses a problem as it renders possibility to get a high artificial *efficiency* value without fully completing the task. This can easily be overcome by using the fastest mean value as benchmark instead of the individually fastest time. The only problem with this solution is that it generates an *efficiency* value of 100 % for the package with the shortest mean time. If this is a problem or not must be concluded by the Package Company, but we do not believe that this is the case.

7.2.3 Satisfaction

The *word choice list* can be altered to infinity, as it can be interpreted in different ways depending on age and personal preferences of the test participants. In order to determine the most common interpretation of the word choices, a large-scale survey amongst feasible test participants can favorably be performed.

Some words that were included in an earlier version of the test method have later on been rationalized away for the benefit of increased simplicity for the test participants. However, some of these words, such as *environmental friendliness*, have actually been requested by the test participants when using the shortened version of the *word choice list*. One can always question if this phrase should be part of a usability test or not, but it has, nevertheless, an obvious relation to the users' satisfaction.

Another problematic issue is the scoring of the *word choice list*. The scorings are based on our Swedish interpretation of these words and can easily be altered without any loss of information, if our grading proves to be erroneous.

7.2.4 Cognitive Load

The *cognitive load* value and its influence on the *overall benchmark* value should be further evaluated when the package usability method has been used on newly developed packages, in order to determine if it renders any valuable information or not. Hopefully, the results will be more useful when the test participants' old experiences and opinions can be taken out of the picture.

References

Printed Sources

Arbnor, Ingeman & Bjerke, Björn (1997), *Methodology for Creating Business Knowledge*, 2nd edition, SAGE Publications, Thousand Oaks

Bell, Judith (2000), *Introduktion till forskningsmetodik*, 3rd edition, Studentlitteratur, Lund

Blom, Gunnar & Holmquist, Björn (1998), *Statistikteori med tillämpningar,* 3rd edition, Studentlitteratur, Lund

Carlshamre, Pär (2001), *A Usability Perspective on Requirements Engineering: From Methodology to Products*, Department of Computer and Information Science, Linköpings universitet, Linköping

Ejvegård, Rolf (1996), *Vetenskaplig metod*, 2nd edition, Studentlitteratur, Lund

Faulkner, Xristine (2000), *Usability Engineering*, Palgrave, Houndmills

Gulliksen, Jan & Göransson, Bengt (2002), *Användarcentrerad systemdesign – en process med fokus på användare och användbarhet*, Studentlitteratur, Lund

Holme, Idar Magne & Solvang, Bernt Krohn (1997), *Forskningsmetodik. Om kvalitativa och kvantitativa metoder,* 2nd edition, Studentlitteratur, Lund

Jacobsen, Dag Ingvar (2002), *Vad, hur och varför? Om metodval i företagsekonomi och andra samhällsvetenskapliga ämnen*, Studentlitteratur, Lund

Lundahl, Ulf & Skärvad, Per-Hugo (1999), *Utredningsmetodik för samhällsvetare och ekonomer,* 3rd edition, Studentlitteratur, Lund

Nielsen, Jakob (1993), *Usability Engineering*, AP Professional, Chestnut Hill

Rubin, Jeffrey (1994), *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests*, John Wiley & Sons, New York

Ulrich, Karl T. & Eppinger, Steven D. (2003), *Product Design and Development*, 3rd edition, McGraw-Hill, New York

Wallén, Göran (1996), *Vetenskapsteori och forskningsmetodik*, 2nd edition, Studentlitteratur, Lund

Wiedersheim-Paul, Finn & Eriksson, Lars Torsten (1991), *Att utreda, forska och rapportera*, Liber-Hermods, Malmö

Internal Sources

[redacted]

[redacted]

Internet Sources

Mathematics Learning Support Centre; Statistics: 2.2 The Wilcoxon signed rank sum test, (2006-12-04)
http://mlsc.lboro.ac.uk/documents/wsrt.pdf

Usability by Design, (2006-09-11),
http://www.usability.uk.com

Interviews

[redacted]

[redacted]

Persson, Britt, *Chairwoman*, PRO Lund, (2006-12-11)

Wiktorsson, Magnus, *Assistant Professor,* Mathematical Statistics, Lund University, (2006-11-29)

Appendix A – First Impression/Handling Questionnaire[108]

Package: □

First Impression/Handling Questionnaire

Considering the package you have just used, indicate your agreement or disagreement with each and every of the following statements by circling one of the figures 1-5:

I expect the package being **comfortable to hold** in my hand

strongly disagree 1 2 3 4 5 strongly agree

I expect the package being **easy to open**

strongly disagree 1 2 3 4 5 strongly agree

I expect the package being **easy to drink** from

strongly disagree 1 2 3 4 5 strongly agree

I expect the package being **easy to close**

strongly disagree 1 2 3 4 5 strongly agree

Overall, I expect this package being **easy to use**

strongly disagree 1 2 3 4 5 strongly agree

[108] [redacted]

Appendix B – Original Word Choice List[109]

Read the following list of words. Considering the package you just have been pouring from, mark the **five** words with a cross that you think describe the package the best.

Robust

Fun

Manageable

Familiar

Confusing

Sophisticated

Anonymous

Elegant

Unmanageable

Well Proportioned

Conservative

Frustrating

Practical

Unrefined

Unsafe

Dull

Innovative

Safe

Boring

Fragile

Appealing

Cheap

Impractical

Annoying

Ugly

Environmentally Friendly

Cool

Expensive

Poor Quality

Easy to understand

High Quality

Simple

Unusual

Well Made

Out of Proportion

Odd

[109] [redacted]

Appendix C – Modified Word Choice List

Please read through the following list and mark the **THREE** words you find the most descriptive for the package you have just used.

Practical

Dated

Innovative

Poor quality

Well proportioned

Confusing

Fragile

Safe

Ugly

Familiar

Unsafe

Robust

Appealing

Out of proportion

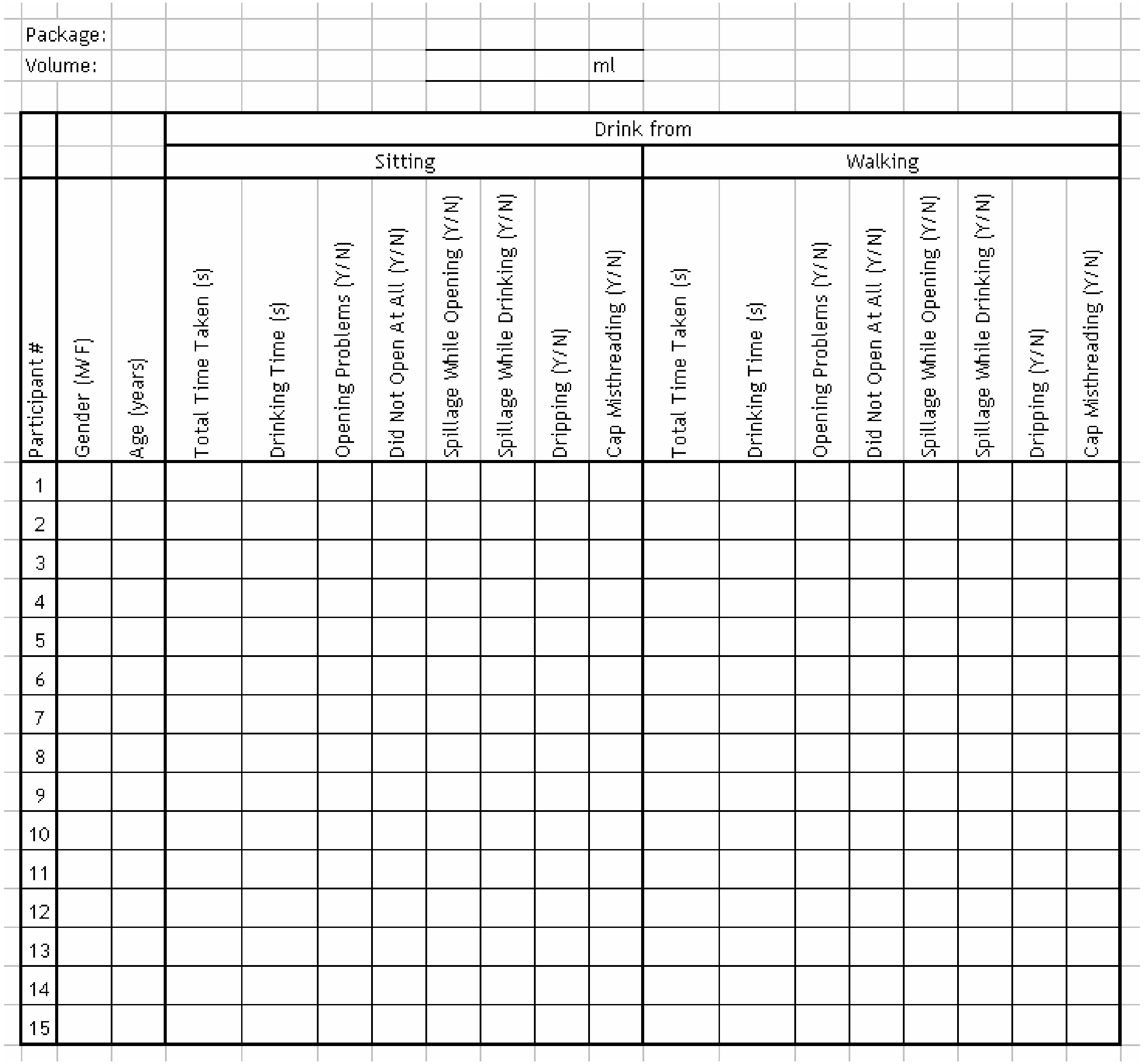
Good quality

Impractical

Elegant

Annoying

Appendix D – Test Data Sheet, Drinking



Package:

Volume: ml

			Drink from															
			Sitting								Walking							
Participant #	Gender (M/F)	Age (years)	Total Time Taken (s)	Drinking Time (s)	Opening Problems (Y/N)	Did Not Open At All (Y/N)	Spillage While Opening (Y/N)	Spillage While Drinking (Y/N)	Dripping (Y/N)	Cap Misthreading (Y/N)	Total Time Taken (s)	Drinking Time (s)	Opening Problems (Y/N)	Did Not Open At All (Y/N)	Spillage While Opening (Y/N)	Spillage While Drinking (Y/N)	Dripping (Y/N)	Cap Misthreading (Y/N)
1																		
2																		
3																		
4																		
5																		
6																		
7																		
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		

Appendix E – Test Data Sheet, Pouring

Package:

Volume: ml

Weight of liquid up to line: g

Weight of dry napkin: g

			Pour from																							
			Carefully												Quickly											
Participant #	Gender (M/F)	Age (years)	Total Time Taken (s)	Pouring Time (s)	Opening Problems (Y/N)	Did Not Open At All (Y/N)	Spillage While Opening (Y/N)	Spiallge While Pouring (Y/N)	Dripping (Y/N)	Cap Misthreading (Y/N)	Weight, Glass 1 (g)	Weight, Glass 2 (g)	Spillage Area (cm2)	Weight, Soaked Napkin (g)	Total Time Taken (s)	Pouring Time (s)	Opening Problems (Y/N)	Did Not Open At All (Y/N)	Spillage While Opening (Y/N)	Spiallge While Pouring (Y/N)	Dripping (Y/N)	Cap Misthreading (Y/N)	Weight, Glass 1 (g)	Weight, Glass 2 (g)	Spillage Area (cm2)	Weight, Soaked Napkin (g)
1																										
2																										
3																										
4																										
5																										
6																										
7																										
8																										
9																										
10																										
11																										
12																										
13																										
14																										
15																										

Appendix F – Test Monitor Guidelines

Welcome the test participant (TP) and explain in an informal manner what will happen (”you are about to test some packages and fill out some short questionnaires regarding your opinions about the packages”).

Ask the TP to sit down and make himself/herself comfortable.

Instruct the TP about the tasks that are to be done. Emphasize the seriousness – no fooling around will be accepted.

The order in which the packages are tested varies from TP to TP – see separate schedule. Some participants start with the pouring test.

Procedure, **drinking test**:

- Show the package to the TP, hand over 2 copies of it and the *pre-test questionnaire*. Tell TP to fill it out.

- Receive the filled out questionnaire.

- During a **Sitting** test, the TP should:
  1. Open the first of the two packages
  2. Take a sip
  3. Put down the package on the table again

- During a **Walking** test, the TP should:
  1. Start to walk around the table while holding the second package
  2. Open the package
  3. Take a sip
  4. Put down the package on the table

- During both these types of tests, note if the TP
  - Did not manage to open the package at all
  - Had difficulties opening the package
  - Spilled while opening the package
  - Spilled while drinking (ask as it may be hard to see)
  - Caused dripping from the package
  - Had problems closing the package

- Hand over the *handling questionnaire* and *word list*. Ask TP to fill them out.

- Receive the filled out questionnaire and word list.

Procedure, **pouring test**.

- Show the package to the TP, hand over 2 copies of it and the *pre-test questionnaire*. Tell TP to fill it out.

- Receive the filled out questionnaire.

- Place a tray with two plastic cups in front of the TP.

- During a **Careful** test, the TP should:
  1. Open the first of the two packages
  2. Fill up both the glasses as close as possible to the marked line
  3. Close the package again

     - Put away the first tray and place a second one with two new plastic cups on in front of the TP.

- During a **Quick** test, the TP should:
  1. Open the second of the two packages
  2. Fill up both the glasses as fast as possible, with the marked line as an indicator of the amount to be poured – ignore spillage and precision
  3. Close the package again

     - Put away the second tray.

- During both these types of tests, note if the TP
  - Did not manage to open the package at all
  - Had difficulties opening the package
  - Spilled while opening the package
  - Spilled while pouring (ask as it may be hard to see)
  - Caused dripping from the package
  - Had problems closing the package

- Hand over the *handling questionnaire* and *word list*. Ask the TP to fill them out.

- Receive the filled out questionnaire and word list.

Ask if the TP has anything to say about the packages (that has not been covered by the questionnaires and word lists).

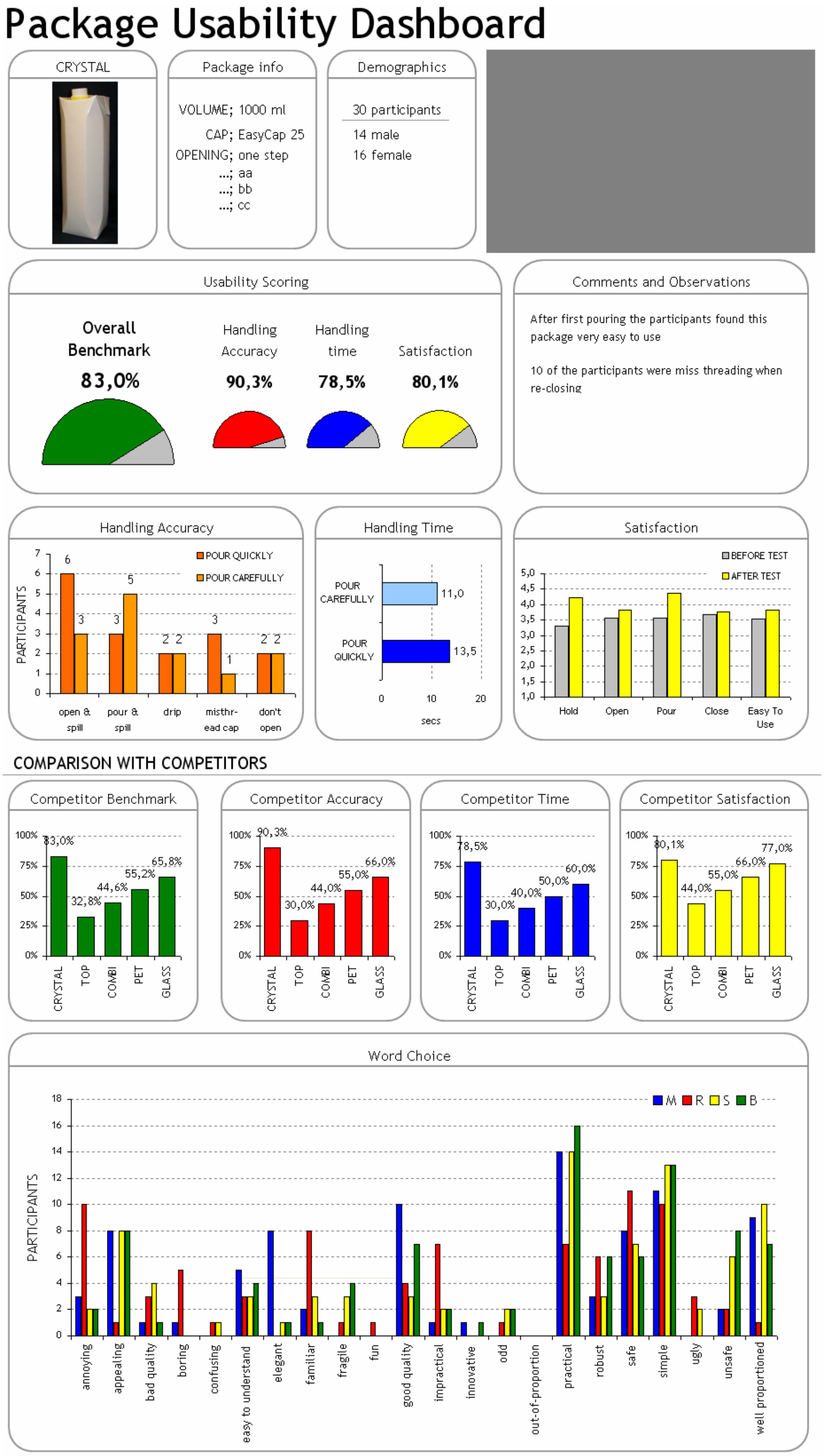
Thank the FP for his/hers participation, hand over the packages that the TP wish to take with him/her in a plastic bag and lead the TP out.

Gather the necessary **pouring** test data by:

- Determining the area of the spilled liquid from the pouring test
- Weighing the glasses of the pouring test
- Weighing the soaked napkins of the pouring test

Watch the video footage of the test in order to determine both the drinking/ pouring time and the total handling time (opening, drinking/pouring and closing if applicable).

Appendix G – Modified Present Report Sheet[110]



Package Usability Dashboard

CRYSTAL

Package info

VOLUME; 1000 ml
CAP; EasyCap 25
OPENING; one step
...; aa
...; bb
...; cc

Demographics

30 participants
14 male
16 female

[redacted]

Usability Scoring

Overall Benchmark	Handling Accuracy	Handling time	Satisfaction
83,0%	90,3%	78,5%	80,1%

Comments and Observations

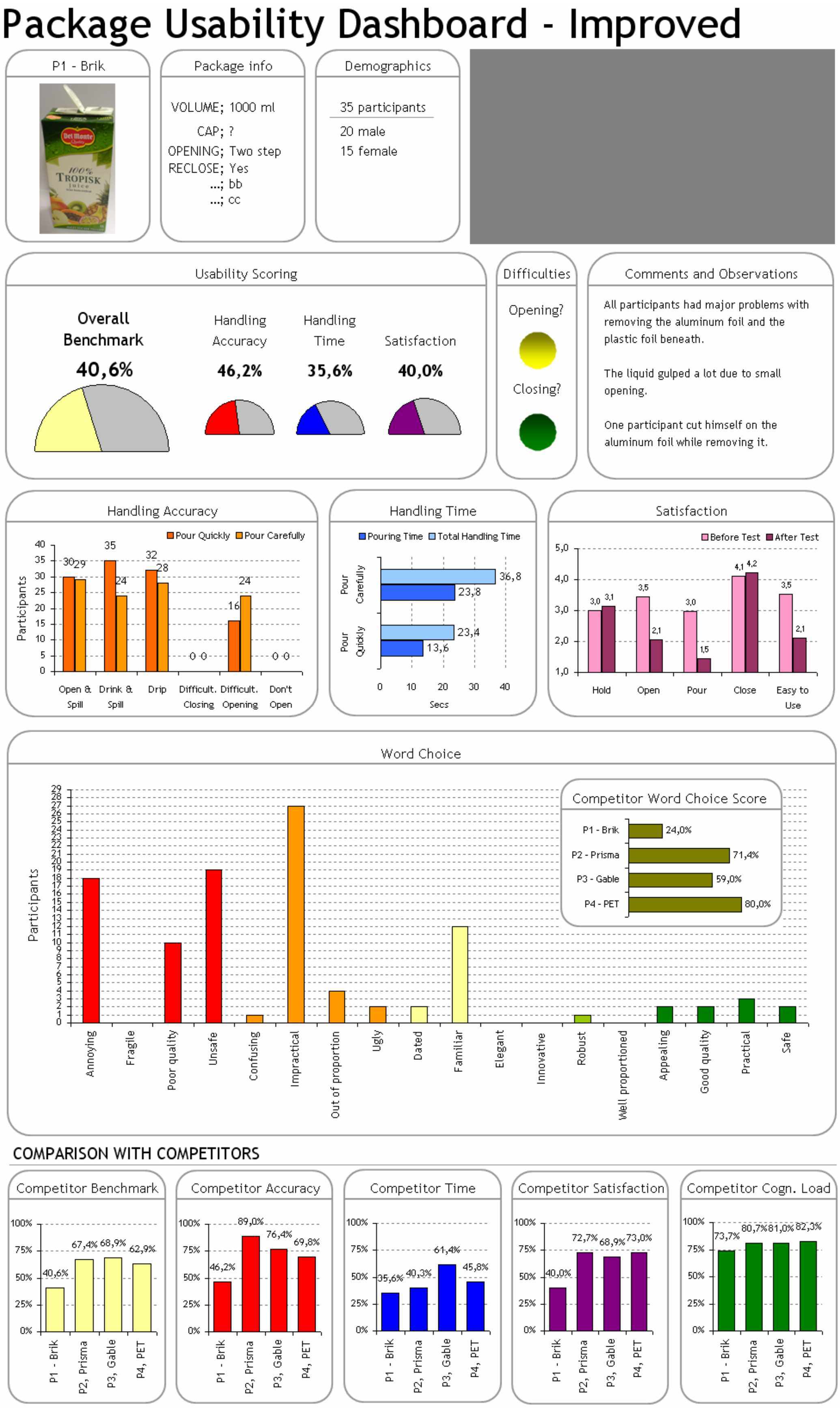
After first pouring the participants found this package very easy to use

10 of the participants were miss threading when re-closing

COMPARISON WITH COMPETITORS

[110] [redacted]

Appendix H – Improved Report Sheet



Package Usability Dashboard - Improved

P1 - Brik

Package info

VOLUME; 1000 ml
CAP; ?
OPENING; Two step
RECLOSE; Yes
...; bb
...; cc

Demographics

35 participants
20 male
15 female

Usability Scoring

Overall Benchmark 40,6%

Handling Accuracy 46,2%

Handling Time 35,6%

Satisfaction 40,0%

Difficulties

Opening?

Closing?

Comments and Observations

All participants had major problems with removing the aluminum foil and the plastic foil beneath.

The liquid gulped a lot due to small opening.

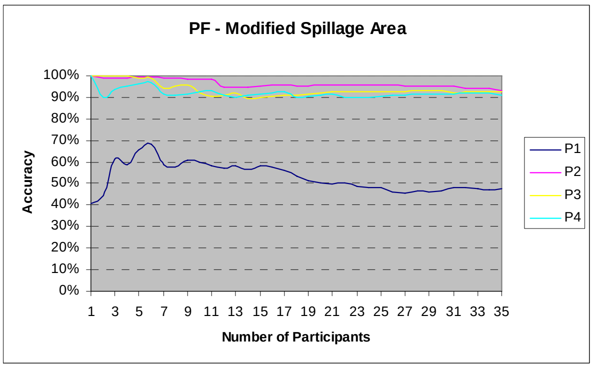
One participant cut himself on the aluminum foil while removing it.

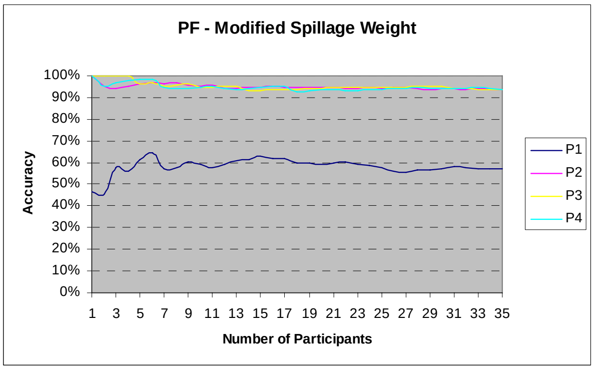
COMPARISON WITH COMPETITORS

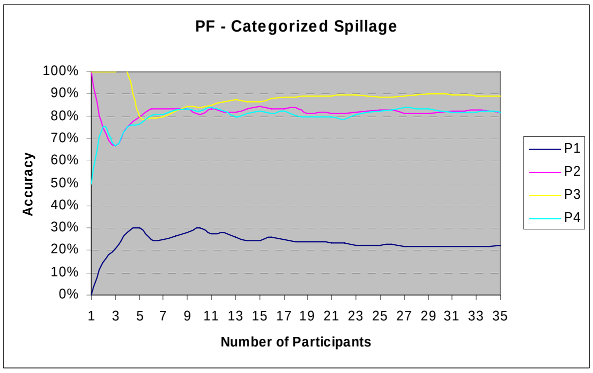
Appendix I – Chart Exploration

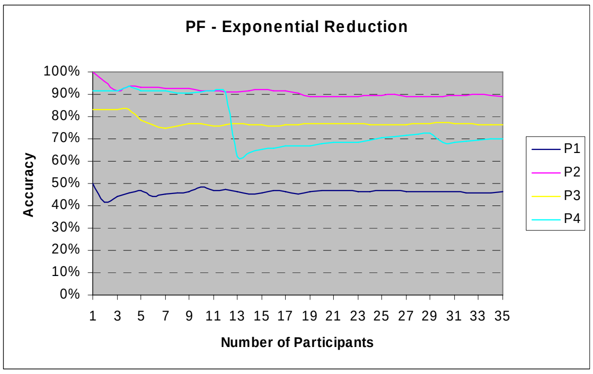
Pouring Test, Students (TP1-TP35)

Accuracy

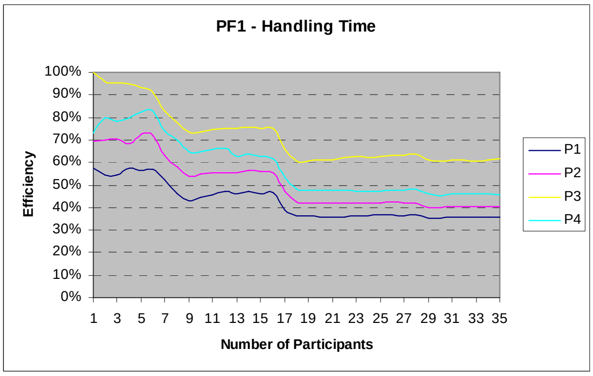


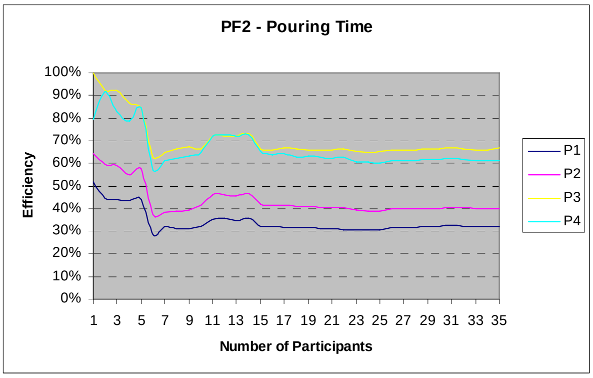




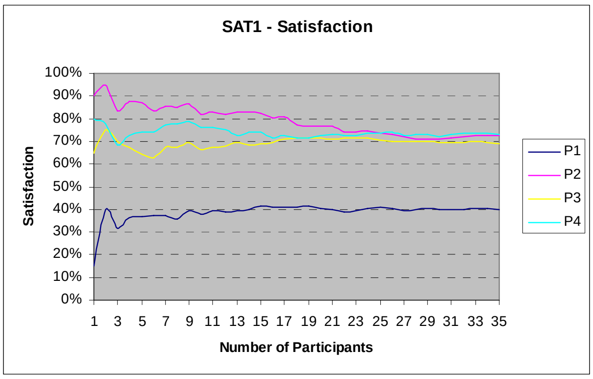


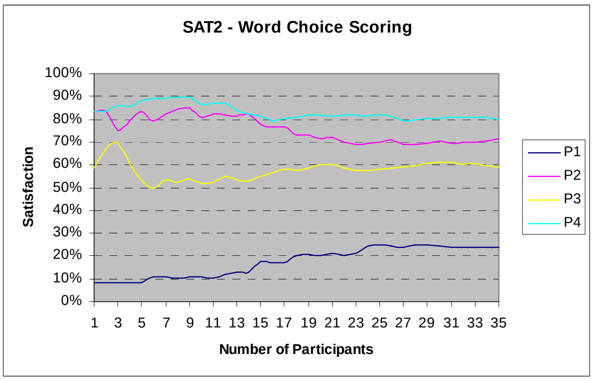
Efficiency



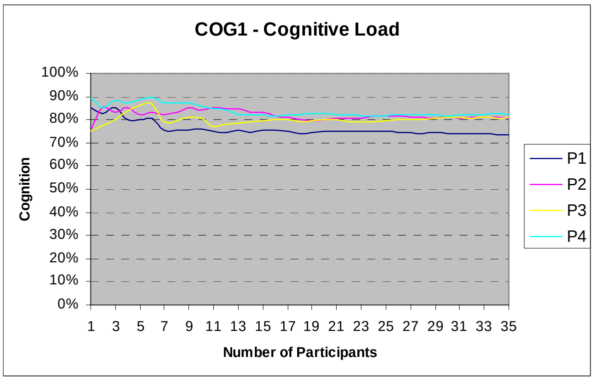


Satisfaction



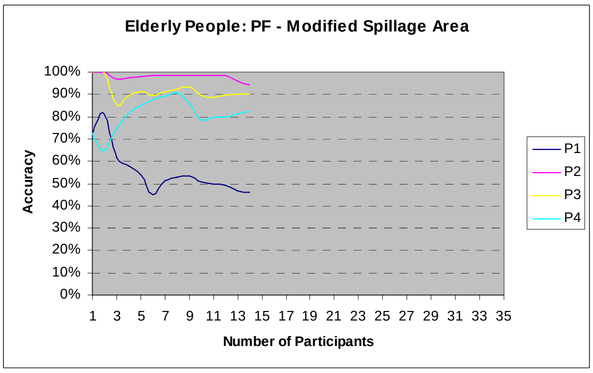


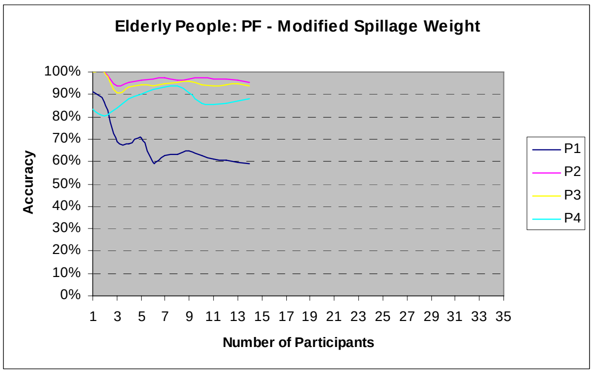
Cognitive Load

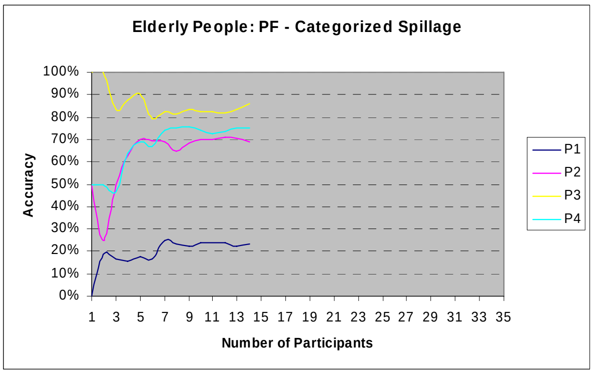


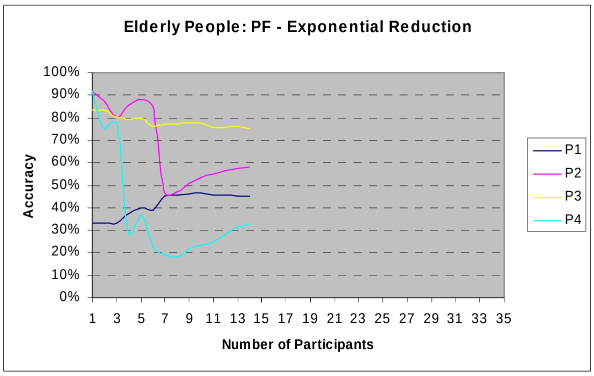
Pouring Test, Elderly People Test (TP36-TP49)

Accuracy

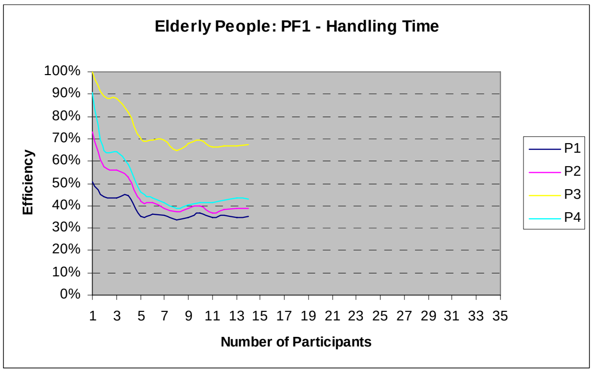


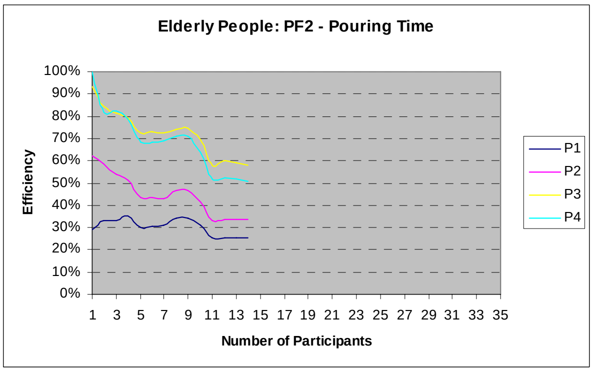




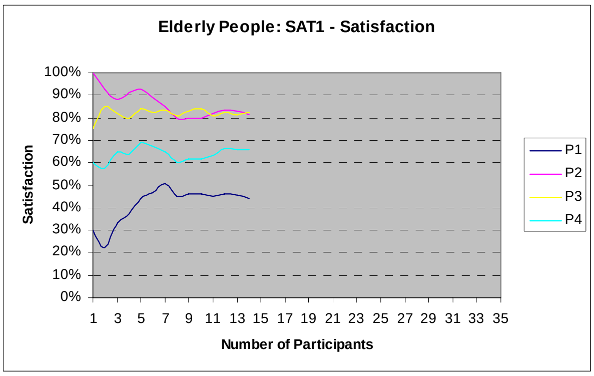


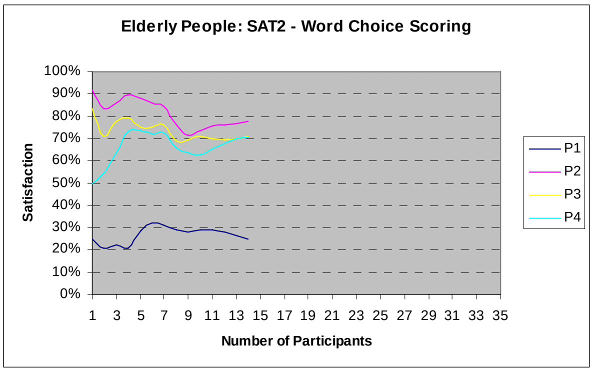
Efficiency



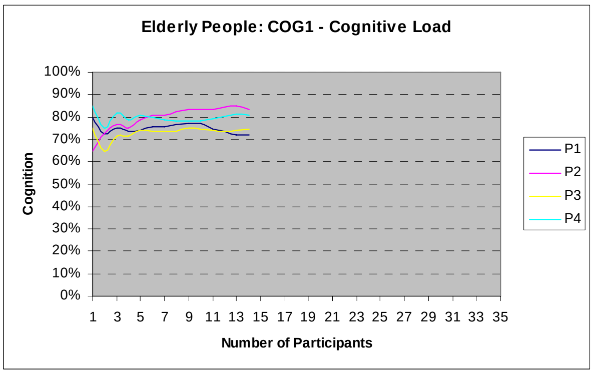


Satisfaction



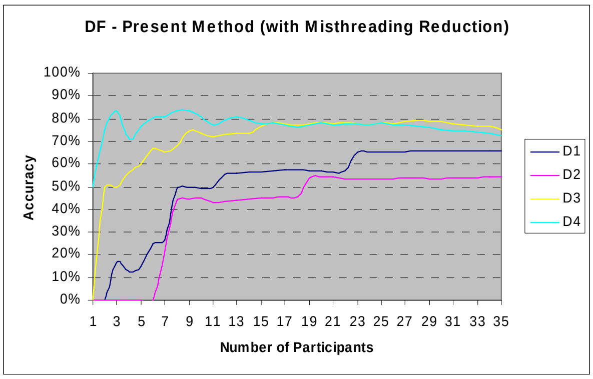


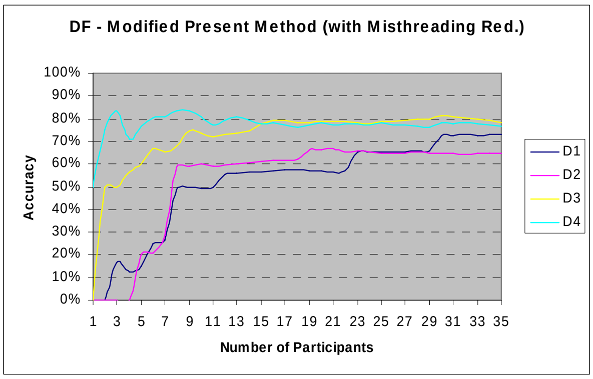
Cognitive Load

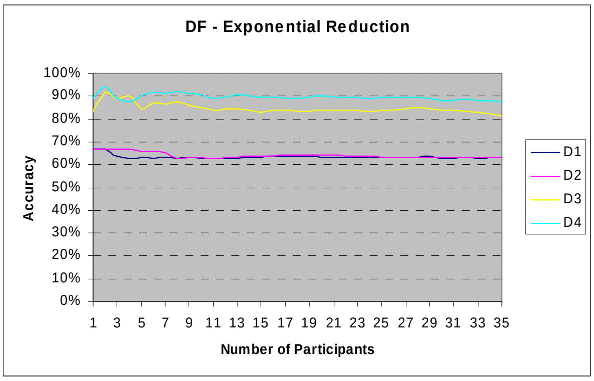


Drinking Test, Students (TP1-TP35)

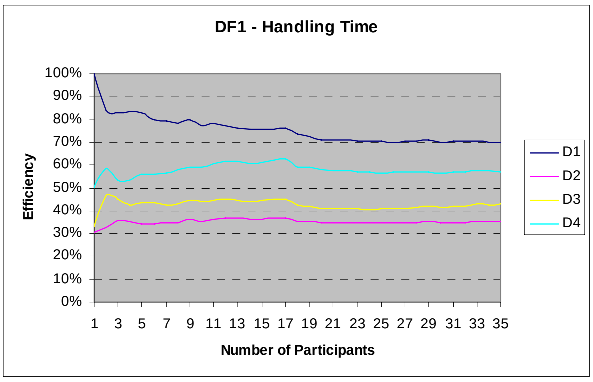
Accuracy

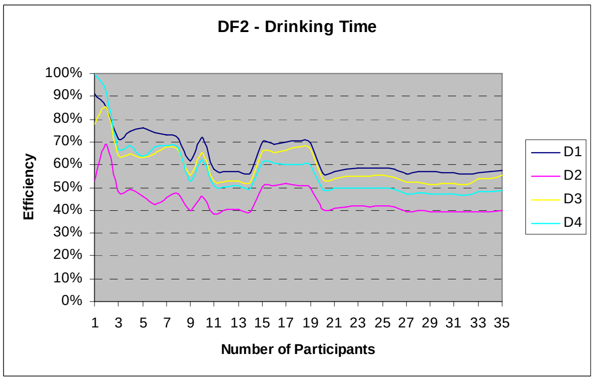




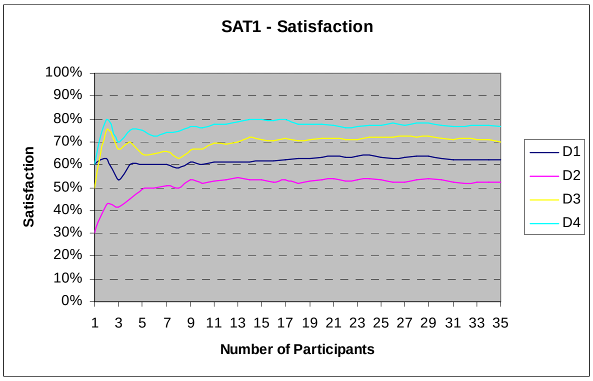


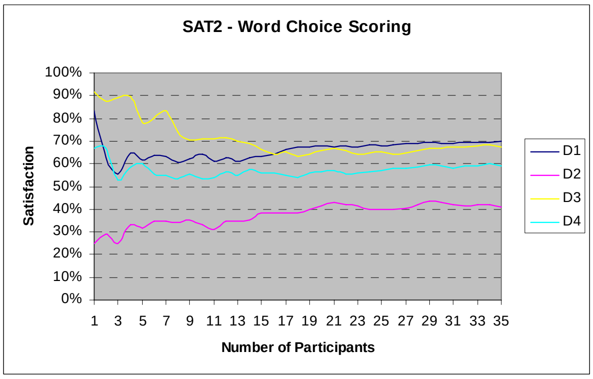
Efficiency



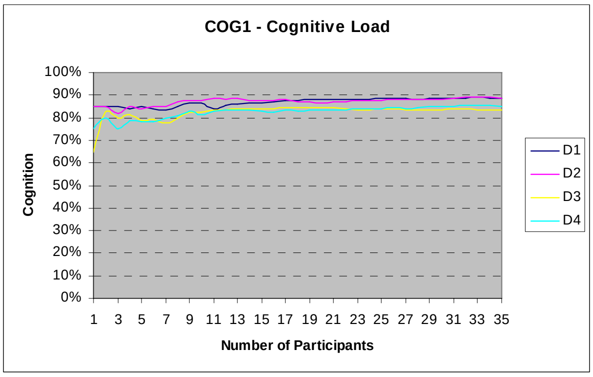


Satisfaction



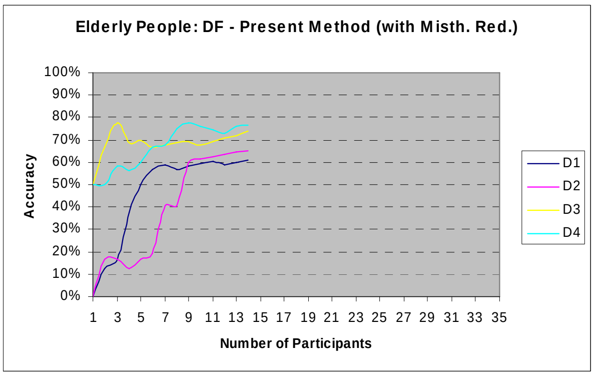


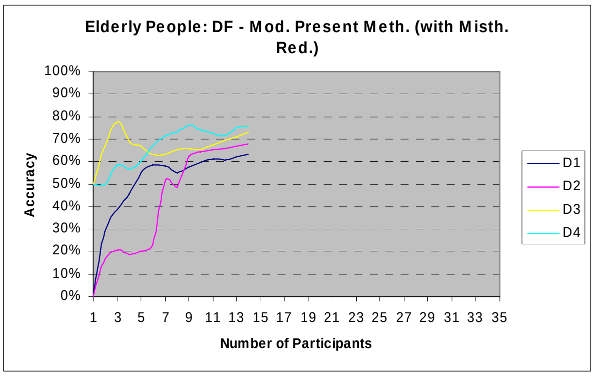
Cognitive Load

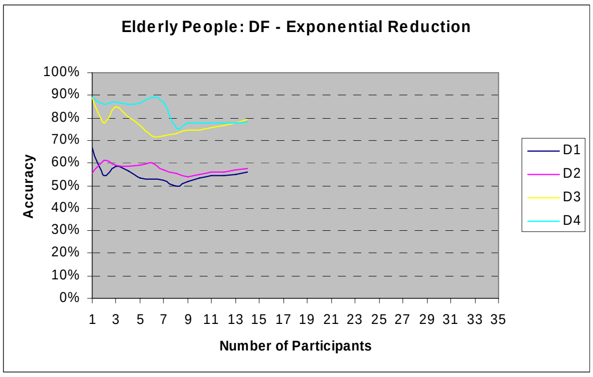


Drinking Test, Elderly People Test (TP36-TP49)

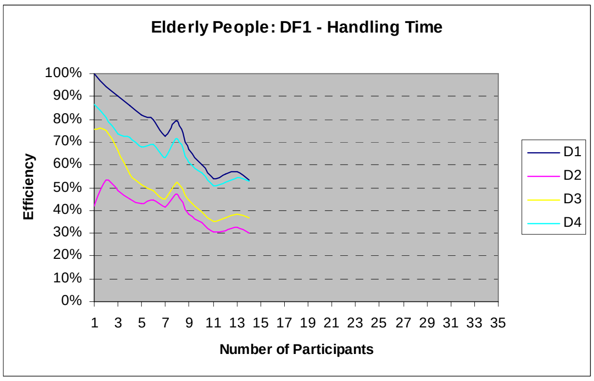
Accuracy

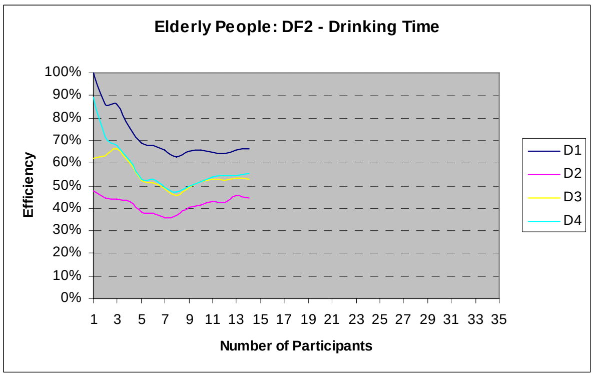




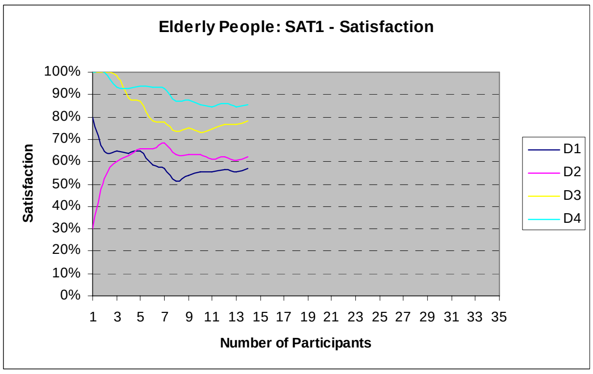


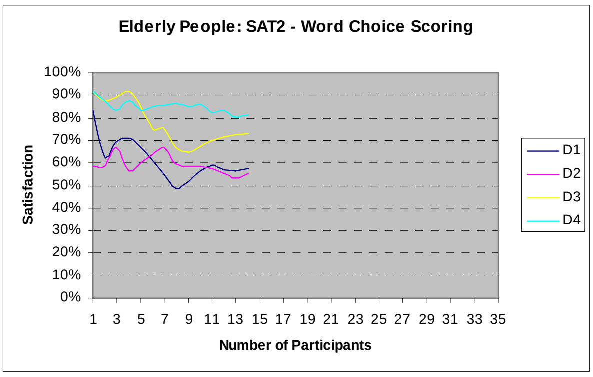
Efficiency



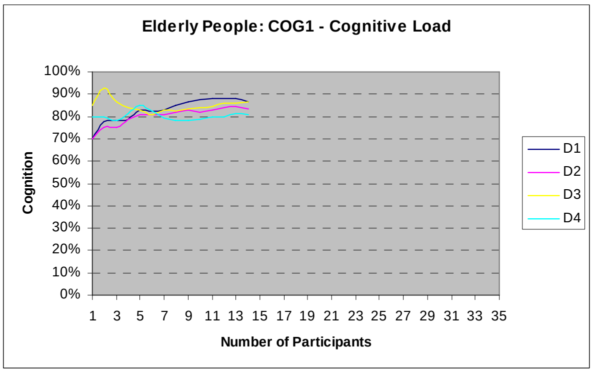


Satisfaction





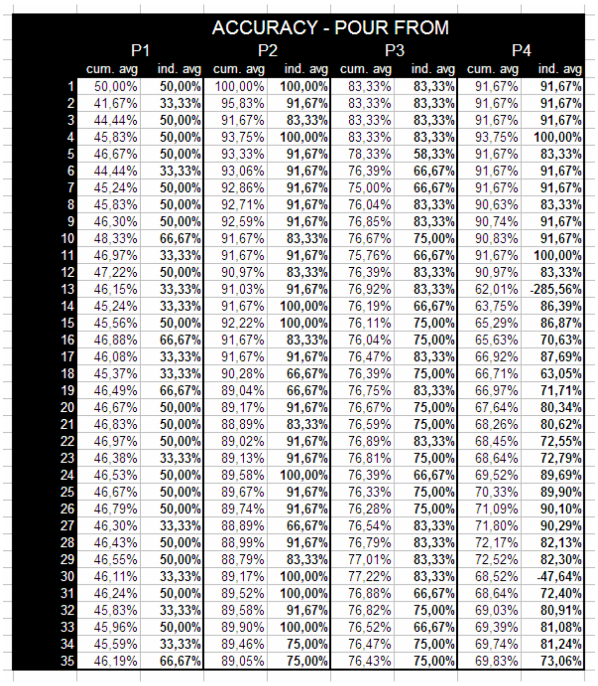
Cognitive Load



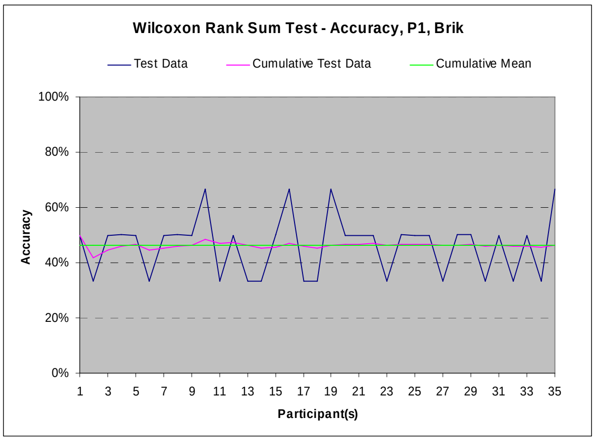
Appendix J – Wilcoxon Signed Rank Sum Test

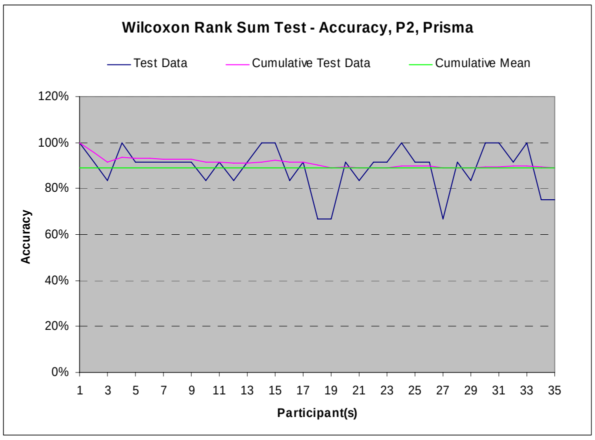
Pouring Test, Students (TP1-TP35)

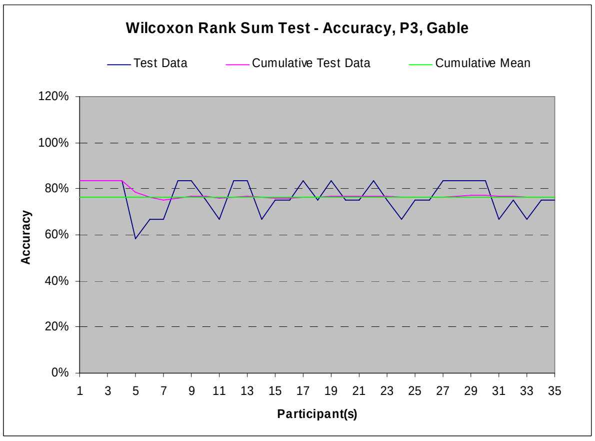
Accuracy – Exponential Reduction

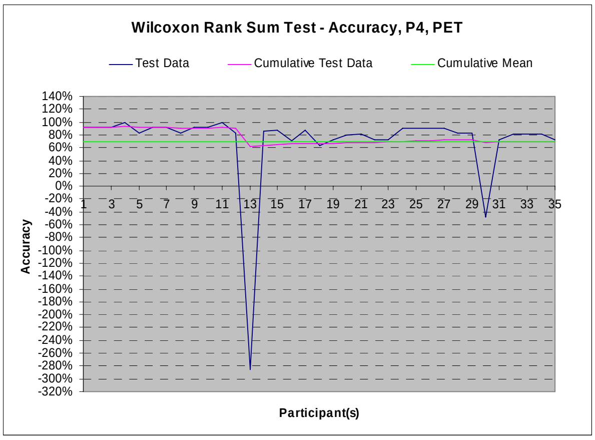


ACCURACY - POUR FROM								
	P1		P2		P3		P4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	50,00%	50,00%	100,00%	100,00%	83,33%	83,33%	91,67%	91,67%
2	41,67%	33,33%	95,83%	91,67%	83,33%	83,33%	91,67%	91,67%
3	44,44%	50,00%	91,67%	83,33%	83,33%	83,33%	91,67%	91,67%
4	45,83%	50,00%	93,75%	100,00%	83,33%	83,33%	93,75%	100,00%
5	46,67%	50,00%	93,33%	91,67%	78,33%	58,33%	91,67%	83,33%
6	44,44%	33,33%	93,06%	91,67%	76,39%	66,67%	91,67%	91,67%
7	45,24%	50,00%	92,86%	91,67%	75,00%	66,67%	91,67%	91,67%
8	45,83%	50,00%	92,71%	91,67%	76,04%	83,33%	90,63%	83,33%
9	46,30%	50,00%	92,59%	91,67%	76,85%	83,33%	90,74%	91,67%
10	48,33%	66,67%	91,67%	83,33%	76,67%	75,00%	90,83%	91,67%
11	46,97%	33,33%	91,67%	91,67%	75,76%	66,67%	91,67%	100,00%
12	47,22%	50,00%	90,97%	83,33%	76,39%	83,33%	90,97%	83,33%
13	46,15%	33,33%	91,03%	91,67%	76,92%	83,33%	62,01%	-285,56%
14	45,24%	33,33%	91,67%	100,00%	76,19%	66,67%	63,75%	86,39%
15	45,56%	50,00%	92,22%	100,00%	76,11%	75,00%	65,29%	86,87%
16	46,88%	66,67%	91,67%	83,33%	76,04%	75,00%	65,63%	70,63%
17	46,08%	33,33%	91,67%	91,67%	76,47%	83,33%	66,92%	87,69%
18	45,37%	33,33%	90,28%	66,67%	76,39%	75,00%	66,71%	63,05%
19	46,49%	66,67%	89,04%	66,67%	76,75%	83,33%	66,97%	71,71%
20	46,67%	50,00%	89,17%	91,67%	76,67%	75,00%	67,64%	80,34%
21	46,83%	50,00%	88,89%	83,33%	76,59%	75,00%	68,26%	80,62%
22	46,97%	50,00%	89,02%	91,67%	76,89%	83,33%	68,45%	72,55%
23	46,38%	33,33%	89,13%	91,67%	76,81%	75,00%	68,64%	72,79%
24	46,53%	50,00%	89,58%	100,00%	76,39%	66,67%	69,52%	89,69%
25	46,67%	50,00%	89,67%	91,67%	76,33%	75,00%	70,33%	89,90%
26	46,79%	50,00%	89,74%	91,67%	76,28%	75,00%	71,09%	90,10%
27	46,30%	33,33%	88,89%	66,67%	76,54%	83,33%	71,80%	90,29%
28	46,43%	50,00%	88,99%	91,67%	76,79%	83,33%	72,17%	82,13%
29	46,55%	50,00%	88,79%	83,33%	77,01%	83,33%	72,52%	82,30%
30	46,11%	33,33%	89,17%	100,00%	77,22%	83,33%	68,52%	-47,64%
31	46,24%	50,00%	89,52%	100,00%	76,88%	66,67%	68,64%	72,40%
32	45,83%	33,33%	89,58%	91,67%	76,82%	75,00%	69,03%	80,91%
33	45,96%	50,00%	89,90%	100,00%	76,52%	66,67%	69,39%	81,08%
34	45,59%	33,33%	89,46%	75,00%	76,47%	75,00%	69,74%	81,24%
35	46,19%	66,67%	89,05%	75,00%	76,43%	75,00%	69,83%	73,06%

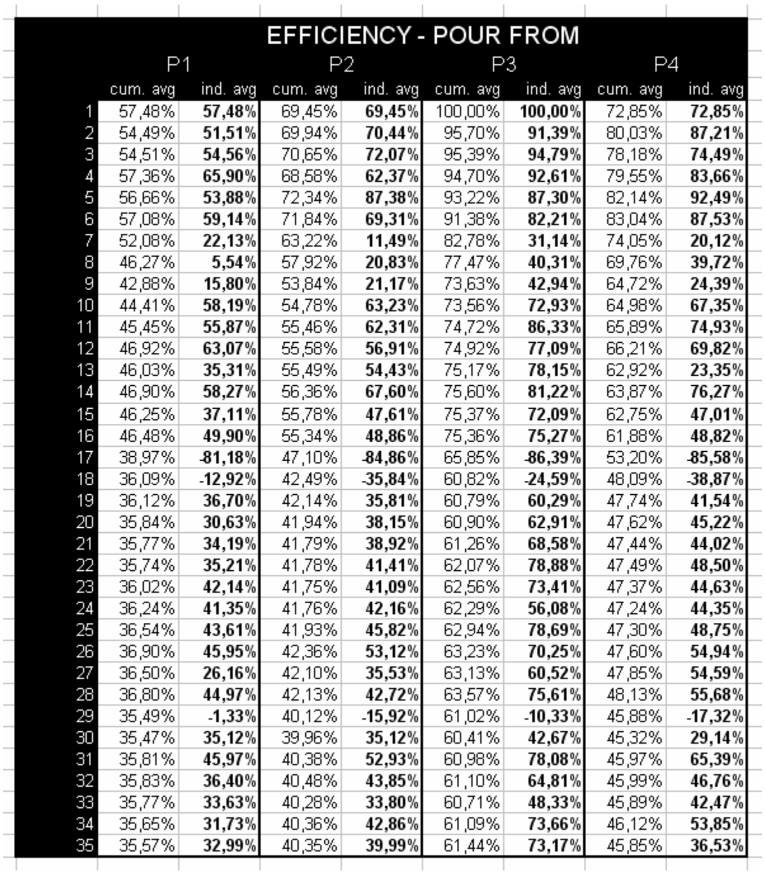




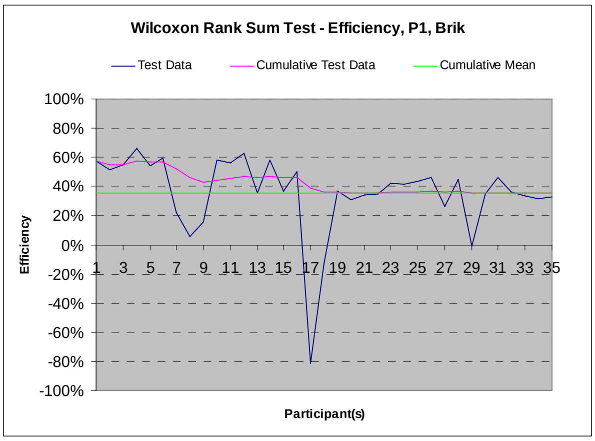


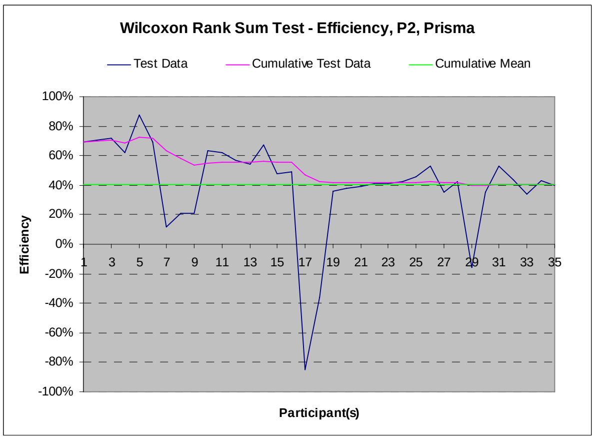


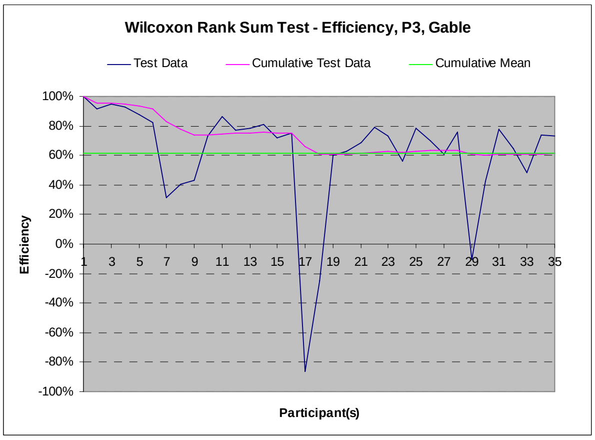
Efficiency – Handling Time

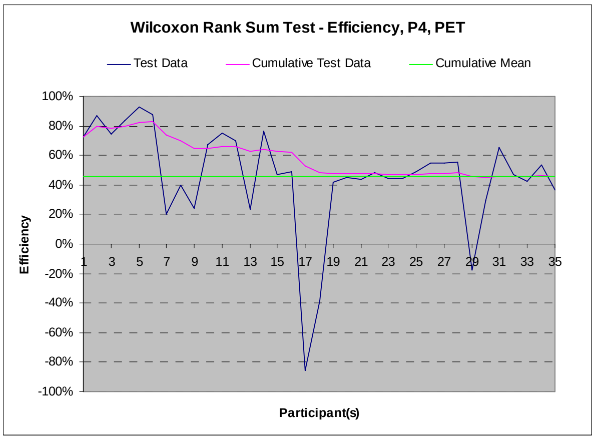


EFFICIENCY - POUR FROM								
	P1		P2		P3		P4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	57,48%	**57,48%**	69,45%	**69,45%**	100,00%	**100,00%**	72,85%	**72,85%**
2	54,49%	**51,51%**	69,94%	**70,44%**	95,70%	**91,39%**	80,03%	**87,21%**
3	54,51%	**54,56%**	70,65%	**72,07%**	95,39%	**94,79%**	78,18%	**74,49%**
4	57,36%	**65,90%**	68,58%	**62,37%**	94,70%	**92,61%**	79,55%	**83,66%**
5	56,66%	**53,88%**	72,34%	**87,38%**	93,22%	**87,30%**	82,14%	**92,49%**
6	57,08%	**59,14%**	71,84%	**69,31%**	91,38%	**82,21%**	83,04%	**87,53%**
7	52,08%	**22,13%**	63,22%	**11,49%**	82,78%	**31,14%**	74,05%	**20,12%**
8	46,27%	**5,54%**	57,92%	**20,83%**	77,47%	**40,31%**	69,76%	**39,72%**
9	42,88%	**15,80%**	53,84%	**21,17%**	73,63%	**42,94%**	64,72%	**24,39%**
10	44,41%	**58,19%**	54,78%	**63,23%**	73,56%	**72,93%**	64,98%	**67,35%**
11	45,45%	**55,87%**	55,46%	**62,31%**	74,72%	**86,33%**	65,89%	**74,93%**
12	46,92%	**63,07%**	55,58%	**56,91%**	74,92%	**77,09%**	66,21%	**69,82%**
13	46,03%	**35,31%**	55,49%	**54,43%**	75,17%	**78,15%**	62,92%	**23,35%**
14	46,90%	**58,27%**	56,36%	**67,60%**	75,60%	**81,22%**	63,87%	**76,27%**
15	46,25%	**37,11%**	55,78%	**47,61%**	75,37%	**72,09%**	62,75%	**47,01%**
16	46,48%	**49,90%**	55,34%	**48,86%**	75,36%	**75,27%**	61,88%	**48,82%**
17	38,97%	**-81,18%**	47,10%	**-84,86%**	65,85%	**-86,39%**	53,20%	**-85,58%**
18	36,09%	**-12,92%**	42,49%	**-35,84%**	60,82%	**-24,59%**	48,09%	**-38,87%**
19	36,12%	**36,70%**	42,14%	**35,81%**	60,79%	**60,29%**	47,74%	**41,54%**
20	35,84%	**30,63%**	41,94%	**38,15%**	60,90%	**62,91%**	47,62%	**45,22%**
21	35,77%	**34,19%**	41,79%	**38,92%**	61,26%	**68,58%**	47,44%	**44,02%**
22	35,74%	**35,21%**	41,78%	**41,41%**	62,07%	**78,88%**	47,49%	**48,50%**
23	36,02%	**42,14%**	41,75%	**41,09%**	62,56%	**73,41%**	47,37%	**44,63%**
24	36,24%	**41,35%**	41,76%	**42,16%**	62,29%	**56,08%**	47,24%	**44,35%**
25	36,54%	**43,61%**	41,93%	**45,82%**	62,94%	**78,69%**	47,30%	**48,75%**
26	36,90%	**45,95%**	42,36%	**53,12%**	63,23%	**70,25%**	47,60%	**54,94%**
27	36,50%	**26,16%**	42,10%	**35,53%**	63,13%	**60,52%**	47,85%	**54,59%**
28	36,80%	**44,97%**	42,13%	**42,72%**	63,57%	**75,61%**	48,13%	**55,68%**
29	35,49%	**-1,33%**	40,12%	**-15,92%**	61,02%	**-10,33%**	45,88%	**-17,32%**
30	35,47%	**35,12%**	39,96%	**35,12%**	60,41%	**42,67%**	45,32%	**29,14%**
31	35,81%	**45,97%**	40,38%	**52,93%**	60,98%	**78,08%**	45,97%	**65,39%**
32	35,83%	**36,40%**	40,48%	**43,85%**	61,10%	**64,81%**	45,99%	**46,76%**
33	35,77%	**33,63%**	40,28%	**33,80%**	60,71%	**48,33%**	45,89%	**42,47%**
34	35,65%	**31,73%**	40,36%	**42,86%**	61,09%	**73,66%**	46,12%	**53,85%**
35	35,57%	**32,99%**	40,35%	**39,99%**	61,44%	**73,17%**	45,85%	**36,53%**

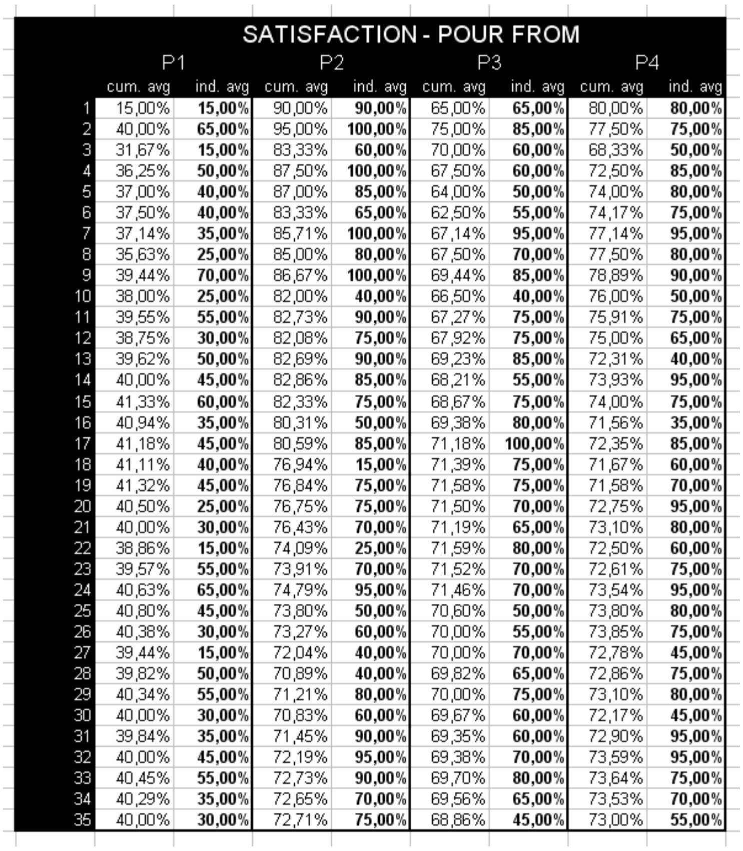




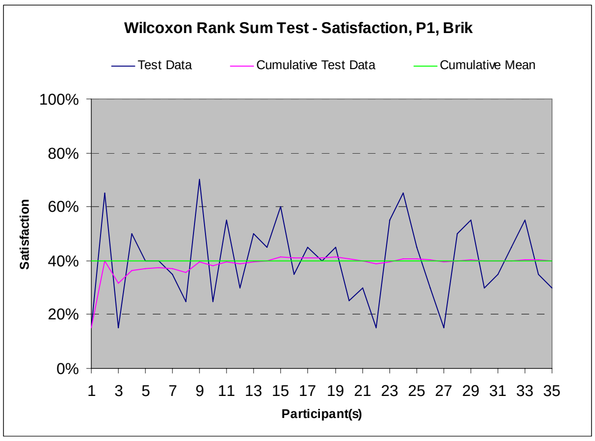


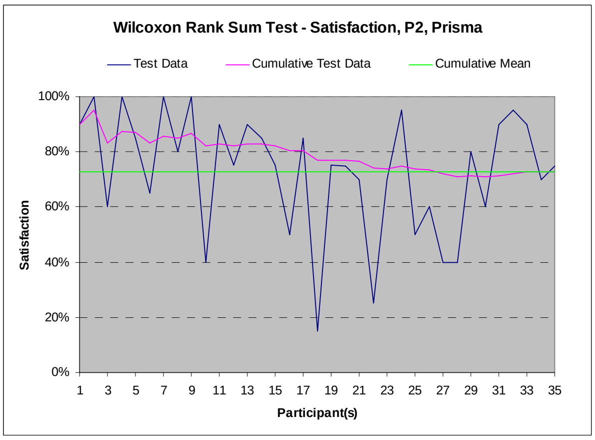


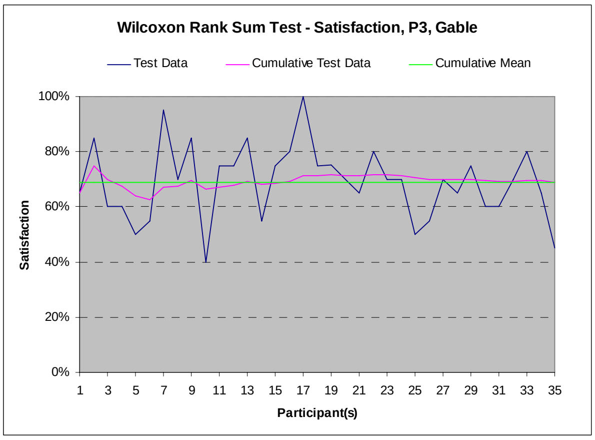
Satisfaction

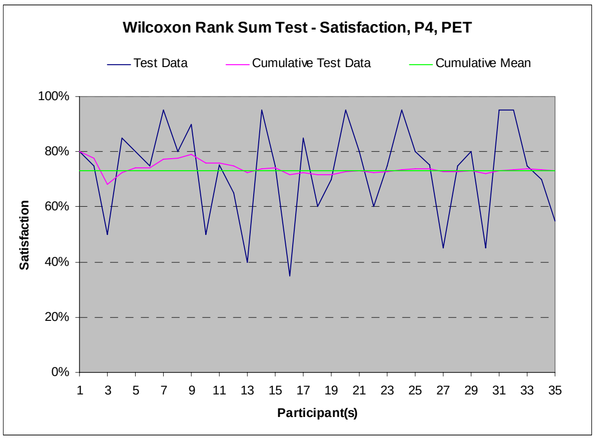


SATISFACTION - POUR FROM	P1		P2		P3		P4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	15,00%	15,00%	90,00%	90,00%	65,00%	65,00%	80,00%	80,00%
2	40,00%	65,00%	95,00%	100,00%	75,00%	85,00%	77,50%	75,00%
3	31,67%	15,00%	83,33%	60,00%	70,00%	60,00%	68,33%	50,00%
4	36,25%	50,00%	87,50%	100,00%	67,50%	60,00%	72,50%	85,00%
5	37,00%	40,00%	87,00%	85,00%	64,00%	50,00%	74,00%	80,00%
6	37,50%	40,00%	83,33%	65,00%	62,50%	55,00%	74,17%	75,00%
7	37,14%	35,00%	85,71%	100,00%	67,14%	95,00%	77,14%	95,00%
8	35,63%	25,00%	85,00%	80,00%	67,50%	70,00%	77,50%	80,00%
9	39,44%	70,00%	86,67%	100,00%	69,44%	85,00%	78,89%	90,00%
10	38,00%	25,00%	82,00%	40,00%	66,50%	40,00%	76,00%	50,00%
11	39,55%	55,00%	82,73%	90,00%	67,27%	75,00%	75,91%	75,00%
12	38,75%	30,00%	82,08%	75,00%	67,92%	75,00%	75,00%	65,00%
13	39,62%	50,00%	82,69%	90,00%	69,23%	85,00%	72,31%	40,00%
14	40,00%	45,00%	82,86%	85,00%	68,21%	55,00%	73,93%	95,00%
15	41,33%	60,00%	82,33%	75,00%	68,67%	75,00%	74,00%	75,00%
16	40,94%	35,00%	80,31%	50,00%	69,38%	80,00%	71,56%	35,00%
17	41,18%	45,00%	80,59%	85,00%	71,18%	100,00%	72,35%	85,00%
18	41,11%	40,00%	76,94%	15,00%	71,39%	75,00%	71,67%	60,00%
19	41,32%	45,00%	76,84%	75,00%	71,58%	75,00%	71,58%	70,00%
20	40,50%	25,00%	76,75%	75,00%	71,50%	70,00%	72,75%	95,00%
21	40,00%	30,00%	76,43%	70,00%	71,19%	65,00%	73,10%	80,00%
22	38,86%	15,00%	74,09%	25,00%	71,59%	80,00%	72,50%	60,00%
23	39,57%	55,00%	73,91%	70,00%	71,52%	70,00%	72,61%	75,00%
24	40,63%	65,00%	74,79%	95,00%	71,46%	70,00%	73,54%	95,00%
25	40,80%	45,00%	73,80%	50,00%	70,60%	50,00%	73,80%	80,00%
26	40,38%	30,00%	73,27%	60,00%	70,00%	55,00%	73,85%	75,00%
27	39,44%	15,00%	72,04%	40,00%	70,00%	70,00%	72,78%	45,00%
28	39,82%	50,00%	70,89%	40,00%	69,82%	65,00%	72,86%	75,00%
29	40,34%	55,00%	71,21%	80,00%	70,00%	75,00%	73,10%	80,00%
30	40,00%	30,00%	70,83%	60,00%	69,67%	60,00%	72,17%	45,00%
31	39,84%	35,00%	71,45%	90,00%	69,35%	60,00%	72,90%	95,00%
32	40,00%	45,00%	72,19%	95,00%	69,38%	70,00%	73,59%	95,00%
33	40,45%	55,00%	72,73%	90,00%	69,70%	80,00%	73,64%	75,00%
34	40,29%	35,00%	72,65%	70,00%	69,56%	65,00%	73,53%	70,00%
35	40,00%	30,00%	72,71%	75,00%	68,86%	45,00%	73,00%	55,00%

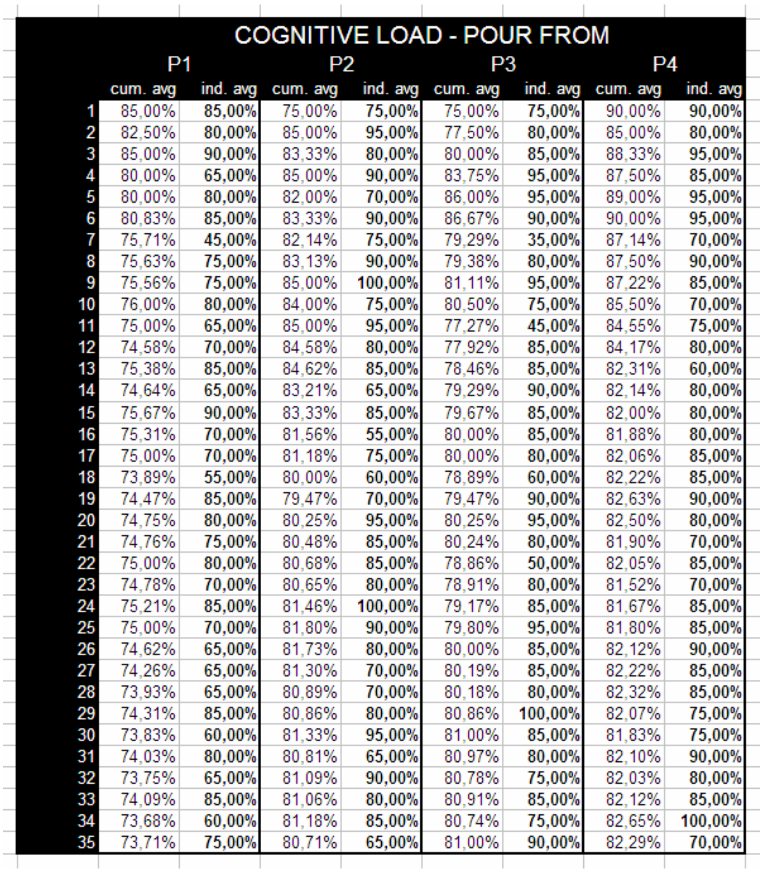

Wilcoxon Rank Sum Test - Satisfaction, P1, Brik


Wilcoxon Rank Sum Test - Satisfaction, P2, Prisma

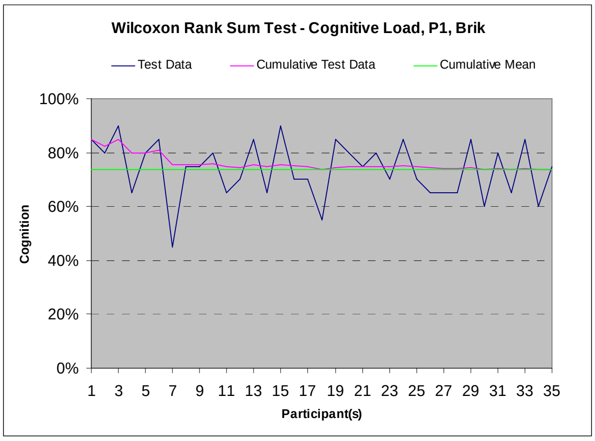

Wilcoxon Rank Sum Test - Satisfaction, P3, Gable


Wilcoxon Rank Sum Test - Satisfaction, P4, PET

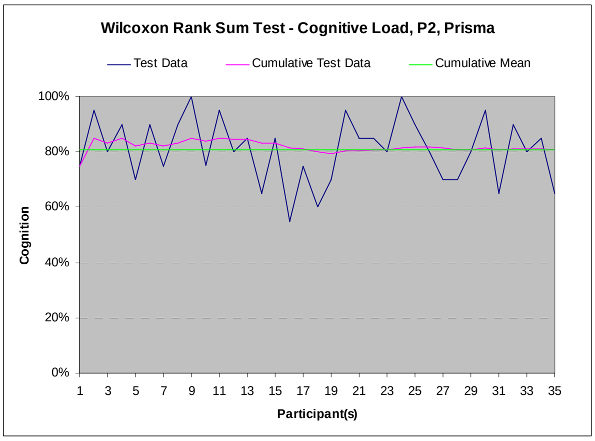
Cognitive Load



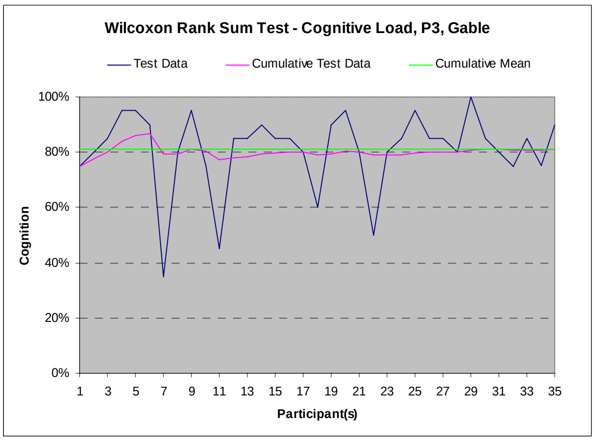
COGNITIVE LOAD - POUR FROM								
	P1		P2		P3		P4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	85,00%	85,00%	75,00%	75,00%	75,00%	75,00%	90,00%	90,00%
2	82,50%	80,00%	85,00%	95,00%	77,50%	80,00%	85,00%	80,00%
3	85,00%	90,00%	83,33%	80,00%	80,00%	85,00%	88,33%	95,00%
4	80,00%	65,00%	85,00%	90,00%	83,75%	95,00%	87,50%	85,00%
5	80,00%	80,00%	82,00%	70,00%	86,00%	95,00%	89,00%	95,00%
6	80,83%	85,00%	83,33%	90,00%	86,67%	90,00%	90,00%	95,00%
7	75,71%	45,00%	82,14%	75,00%	79,29%	35,00%	87,14%	70,00%
8	75,63%	75,00%	83,13%	90,00%	79,38%	80,00%	87,50%	90,00%
9	75,56%	75,00%	85,00%	100,00%	81,11%	95,00%	87,22%	85,00%
10	76,00%	80,00%	84,00%	75,00%	80,50%	75,00%	85,50%	70,00%
11	75,00%	65,00%	85,00%	95,00%	77,27%	45,00%	84,55%	75,00%
12	74,58%	70,00%	84,58%	80,00%	77,92%	85,00%	84,17%	80,00%
13	75,38%	85,00%	84,62%	85,00%	78,46%	85,00%	82,31%	60,00%
14	74,64%	65,00%	83,21%	65,00%	79,29%	90,00%	82,14%	80,00%
15	75,67%	90,00%	83,33%	85,00%	79,67%	85,00%	82,00%	80,00%
16	75,31%	70,00%	81,56%	55,00%	80,00%	85,00%	81,88%	80,00%
17	75,00%	70,00%	81,18%	75,00%	80,00%	80,00%	82,06%	85,00%
18	73,89%	55,00%	80,00%	60,00%	78,89%	60,00%	82,22%	85,00%
19	74,47%	85,00%	79,47%	70,00%	79,47%	90,00%	82,63%	90,00%
20	74,75%	80,00%	80,25%	95,00%	80,25%	95,00%	82,50%	80,00%
21	74,76%	75,00%	80,48%	85,00%	80,24%	80,00%	81,90%	70,00%
22	75,00%	80,00%	80,68%	85,00%	78,86%	50,00%	82,05%	85,00%
23	74,78%	70,00%	80,65%	80,00%	78,91%	80,00%	81,52%	70,00%
24	75,21%	85,00%	81,46%	100,00%	79,17%	85,00%	81,67%	85,00%
25	75,00%	70,00%	81,80%	90,00%	79,80%	95,00%	81,80%	85,00%
26	74,62%	65,00%	81,73%	80,00%	80,00%	85,00%	82,12%	90,00%
27	74,26%	65,00%	81,30%	70,00%	80,19%	85,00%	82,22%	85,00%
28	73,93%	65,00%	80,89%	70,00%	80,18%	80,00%	82,32%	85,00%
29	74,31%	85,00%	80,86%	80,00%	80,86%	100,00%	82,07%	75,00%
30	73,83%	60,00%	81,33%	95,00%	81,00%	85,00%	81,83%	75,00%
31	74,03%	80,00%	80,81%	65,00%	80,97%	80,00%	82,10%	90,00%
32	73,75%	65,00%	81,09%	90,00%	80,78%	75,00%	82,03%	80,00%
33	74,09%	85,00%	81,06%	80,00%	80,91%	85,00%	82,12%	85,00%
34	73,68%	60,00%	81,18%	85,00%	80,74%	75,00%	82,65%	100,00%
35	73,71%	75,00%	80,71%	65,00%	81,00%	90,00%	82,29%	70,00%



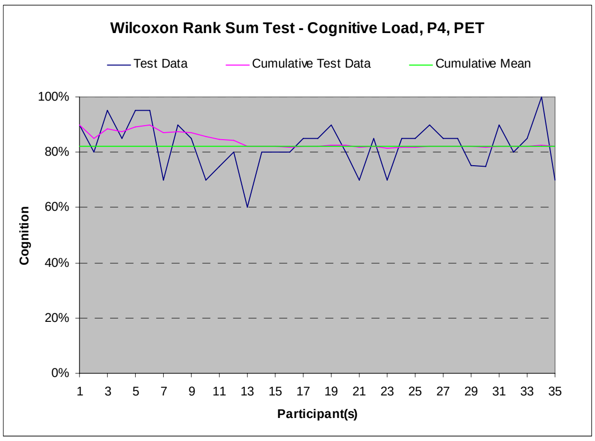
Wilcoxon Rank Sum Test - Cognitive Load, P1, Brik



Wilcoxon Rank Sum Test - Cognitive Load, P2, Prisma



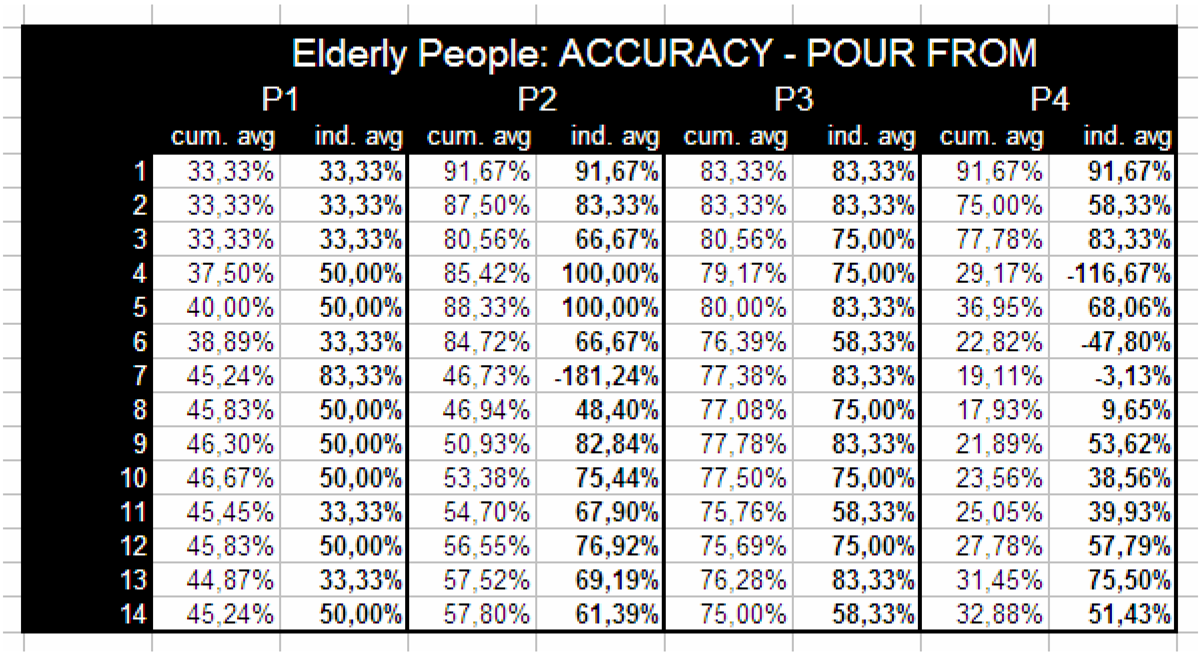
Wilcoxon Rank Sum Test - Cognitive Load, P3, Gable



Wilcoxon Rank Sum Test - Cognitive Load, P4, PET

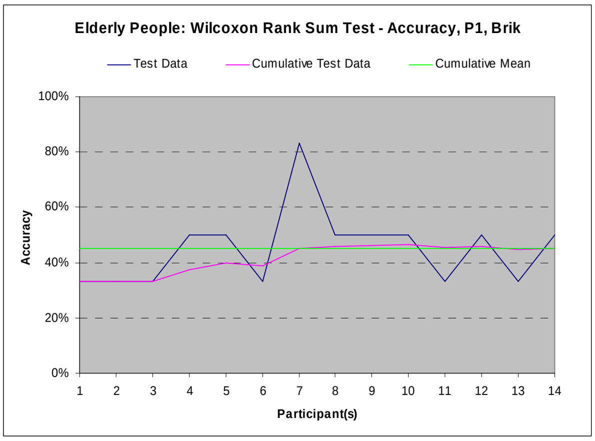
Pouring Test, Elderly People (TP36-TP49)

Accuracy – Exponential Reduction

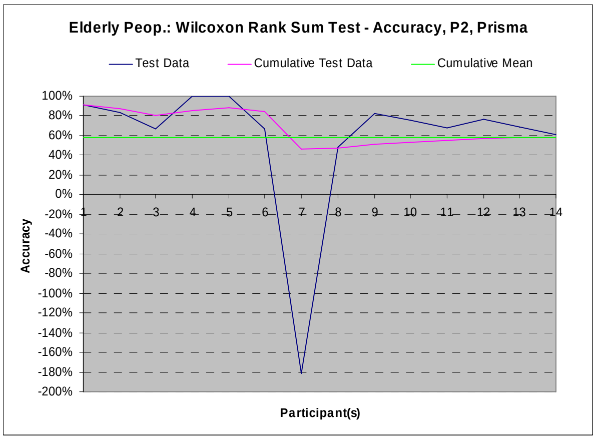


Elderly People: ACCURACY - POUR FROM

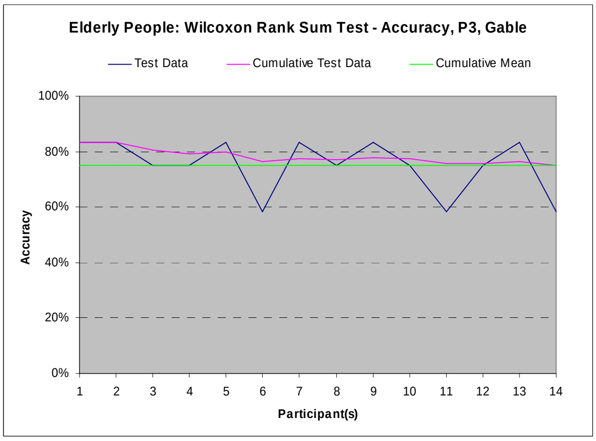
	P1		P2		P3		P4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	33,33%	33,33%	91,67%	91,67%	83,33%	83,33%	91,67%	91,67%
2	33,33%	33,33%	87,50%	83,33%	83,33%	83,33%	75,00%	58,33%
3	33,33%	33,33%	80,56%	66,67%	80,56%	75,00%	77,78%	83,33%
4	37,50%	50,00%	85,42%	100,00%	79,17%	75,00%	29,17%	-116,67%
5	40,00%	50,00%	88,33%	100,00%	80,00%	83,33%	36,95%	68,06%
6	38,89%	33,33%	84,72%	66,67%	76,39%	58,33%	22,82%	-47,80%
7	45,24%	83,33%	46,73%	-181,24%	77,38%	83,33%	19,11%	-3,13%
8	45,83%	50,00%	46,94%	48,40%	77,08%	75,00%	17,93%	9,65%
9	46,30%	50,00%	50,93%	82,84%	77,78%	83,33%	21,89%	53,62%
10	46,67%	50,00%	53,38%	75,44%	77,50%	75,00%	23,56%	38,56%
11	45,45%	33,33%	54,70%	67,90%	75,76%	58,33%	25,05%	39,93%
12	45,83%	50,00%	56,55%	76,92%	75,69%	75,00%	27,78%	57,79%
13	44,87%	33,33%	57,52%	69,19%	76,28%	83,33%	31,45%	75,50%
14	45,24%	50,00%	57,80%	61,39%	75,00%	58,33%	32,88%	51,43%



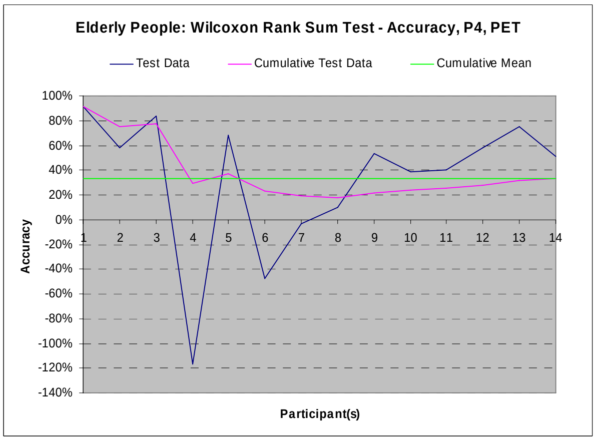
Elderly People: Wilcoxon Rank Sum Test - Accuracy, P1, Brik



Elderly Peop.: Wilcoxon Rank Sum Test - Accuracy, P2, Prisma

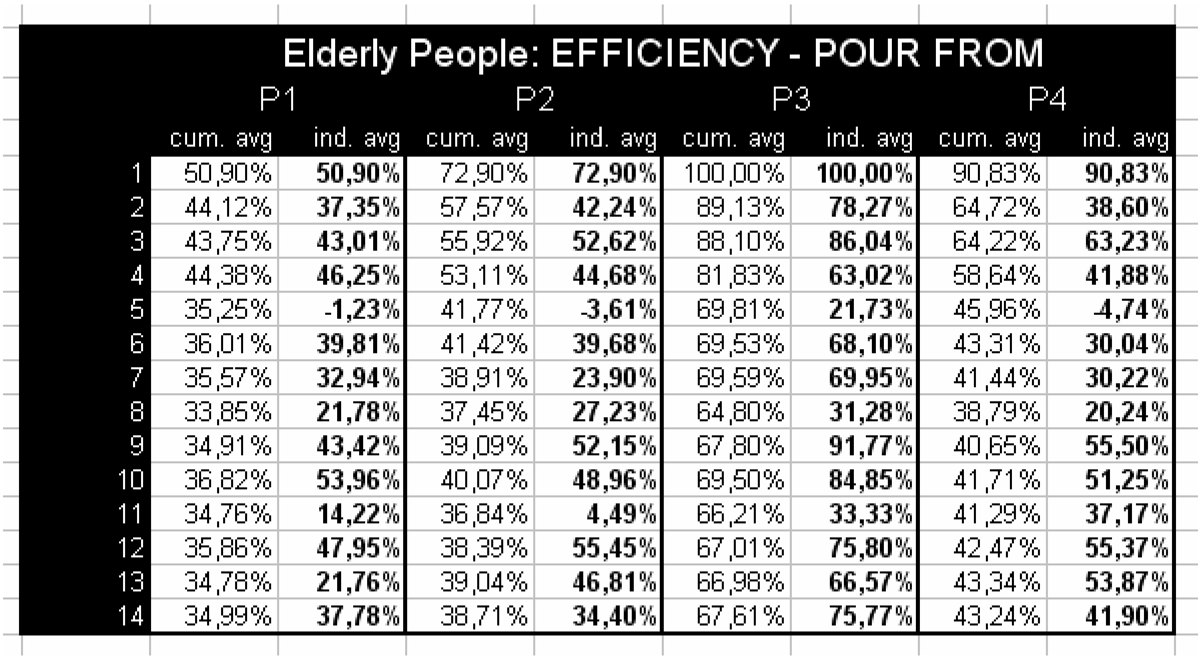


Elderly People: Wilcoxon Rank Sum Test - Accuracy, P3, Gable

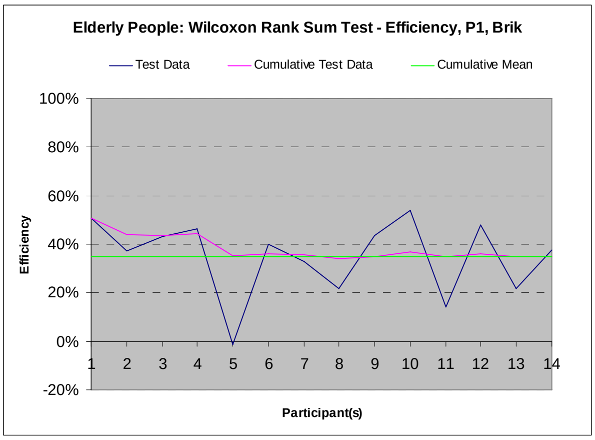


Elderly People: Wilcoxon Rank Sum Test - Accuracy, P4, PET

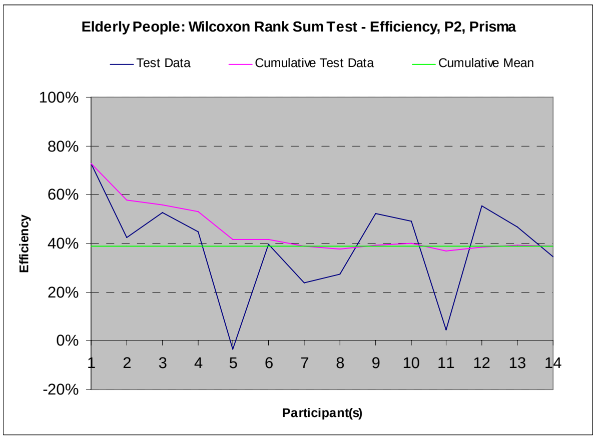
Efficiency – Handling Time



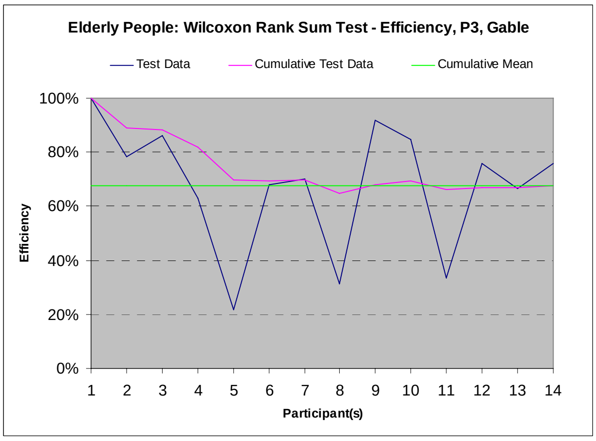
Elderly People: EFFICIENCY - POUR FROM								
	P1		P2		P3		P4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	50,90%	**50,90%**	72,90%	**72,90%**	100,00%	**100,00%**	90,83%	**90,83%**
2	44,12%	**37,35%**	57,57%	**42,24%**	89,13%	**78,27%**	64,72%	**38,60%**
3	43,75%	**43,01%**	55,92%	**52,62%**	88,10%	**86,04%**	64,22%	**63,23%**
4	44,38%	**46,25%**	53,11%	**44,68%**	81,83%	**63,02%**	58,64%	**41,88%**
5	35,25%	**-1,23%**	41,77%	**-3,61%**	69,81%	**21,73%**	45,96%	**-4,74%**
6	36,01%	**39,81%**	41,42%	**39,68%**	69,53%	**68,10%**	43,31%	**30,04%**
7	35,57%	**32,94%**	38,91%	**23,90%**	69,59%	**69,95%**	41,44%	**30,22%**
8	33,85%	**21,78%**	37,45%	**27,23%**	64,80%	**31,28%**	38,79%	**20,24%**
9	34,91%	**43,42%**	39,09%	**52,15%**	67,80%	**91,77%**	40,65%	**55,50%**
10	36,82%	**53,96%**	40,07%	**48,96%**	69,50%	**84,85%**	41,71%	**51,25%**
11	34,76%	**14,22%**	36,84%	**4,49%**	66,21%	**33,33%**	41,29%	**37,17%**
12	35,86%	**47,95%**	38,39%	**55,45%**	67,01%	**75,80%**	42,47%	**55,37%**
13	34,78%	**21,76%**	39,04%	**46,81%**	66,98%	**66,57%**	43,34%	**53,87%**
14	34,99%	**37,78%**	38,71%	**34,40%**	67,61%	**75,77%**	43,24%	**41,90%**



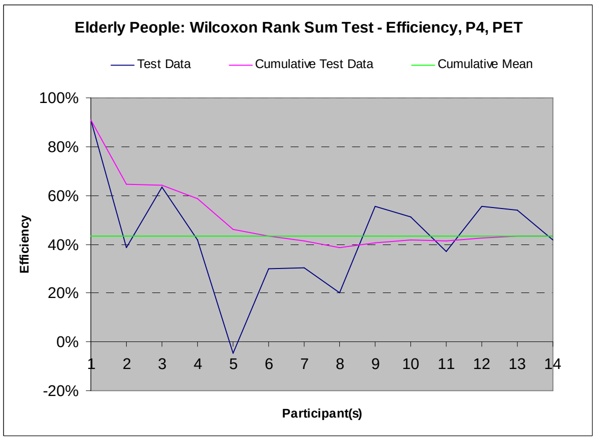
Elderly People: Wilcoxon Rank Sum Test - Efficiency, P1, Brik



Elderly People: Wilcoxon Rank Sum Test - Efficiency, P2, Prisma

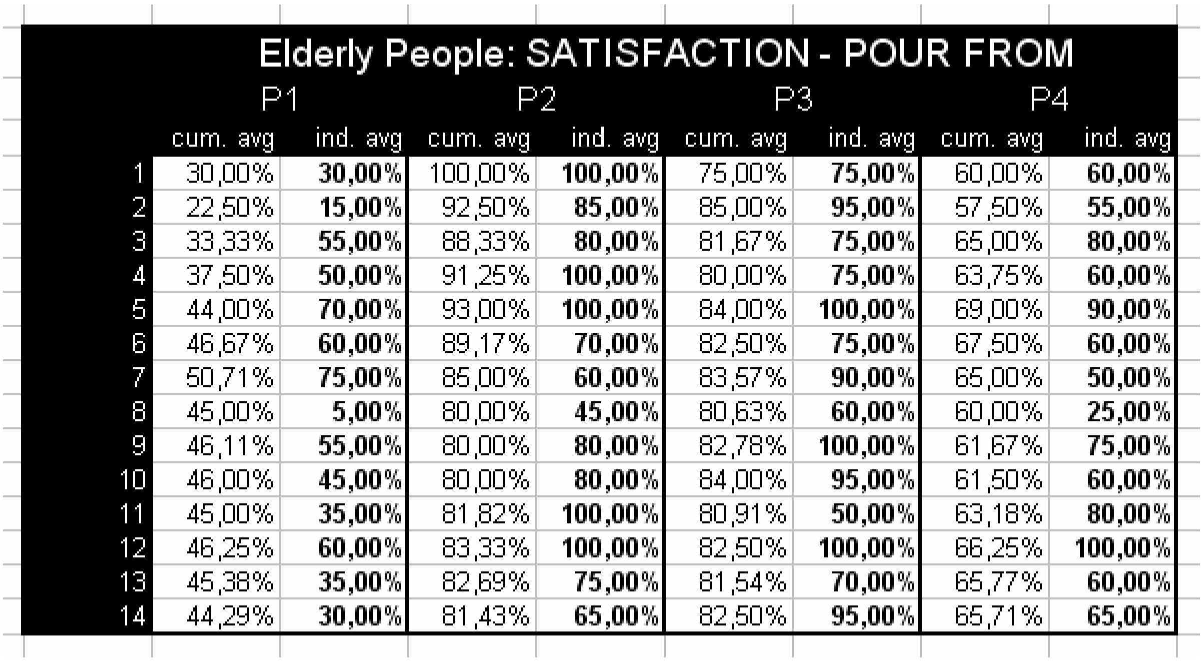


Elderly People: Wilcoxon Rank Sum Test - Efficiency, P3, Gable

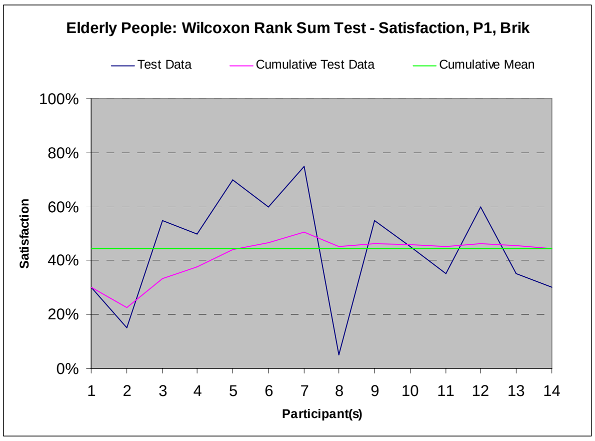


Elderly People: Wilcoxon Rank Sum Test - Efficiency, P4, PET

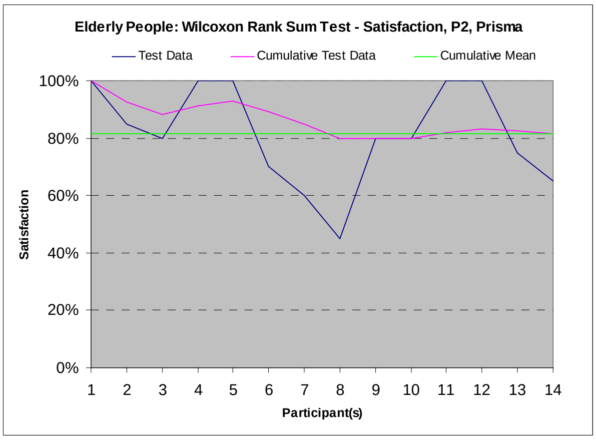
Satisfaction



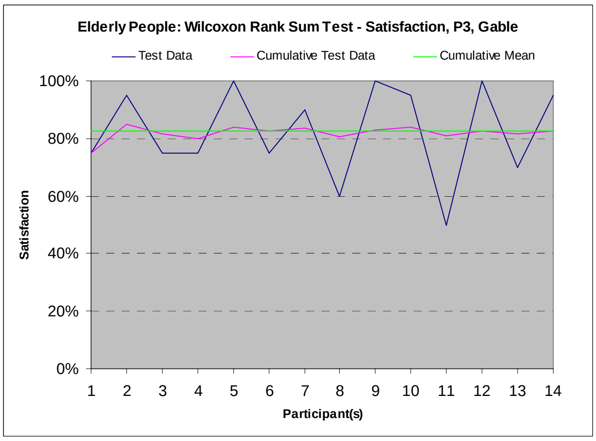
Elderly People: SATISFACTION - POUR FROM	P1		P2		P3		P4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	30,00%	**30,00%**	100,00%	**100,00%**	75,00%	**75,00%**	60,00%	**60,00%**
2	22,50%	**15,00%**	92,50%	**85,00%**	85,00%	**95,00%**	57,50%	**55,00%**
3	33,33%	**55,00%**	88,33%	**80,00%**	81,67%	**75,00%**	65,00%	**80,00%**
4	37,50%	**50,00%**	91,25%	**100,00%**	80,00%	**75,00%**	63,75%	**60,00%**
5	44,00%	**70,00%**	93,00%	**100,00%**	84,00%	**100,00%**	69,00%	**90,00%**
6	46,67%	**60,00%**	89,17%	**70,00%**	82,50%	**75,00%**	67,50%	**60,00%**
7	50,71%	**75,00%**	85,00%	**60,00%**	83,57%	**90,00%**	65,00%	**50,00%**
8	45,00%	**5,00%**	80,00%	**45,00%**	80,63%	**60,00%**	60,00%	**25,00%**
9	46,11%	**55,00%**	80,00%	**80,00%**	82,78%	**100,00%**	61,67%	**75,00%**
10	46,00%	**45,00%**	80,00%	**80,00%**	84,00%	**95,00%**	61,50%	**60,00%**
11	45,00%	**35,00%**	81,82%	**100,00%**	80,91%	**50,00%**	63,18%	**80,00%**
12	46,25%	**60,00%**	83,33%	**100,00%**	82,50%	**100,00%**	66,25%	**100,00%**
13	45,38%	**35,00%**	82,69%	**75,00%**	81,54%	**70,00%**	65,77%	**60,00%**
14	44,29%	**30,00%**	81,43%	**65,00%**	82,50%	**95,00%**	65,71%	**65,00%**



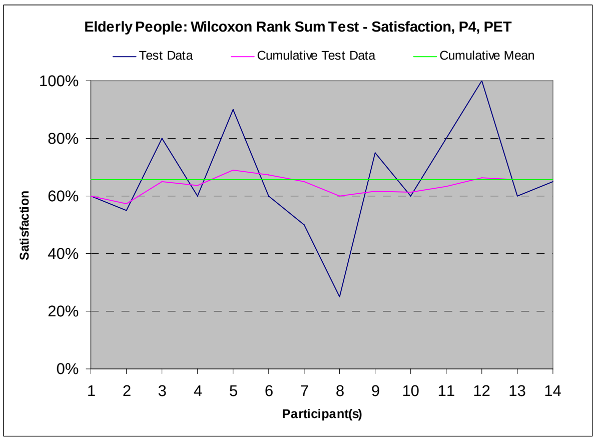
Elderly People: Wilcoxon Rank Sum Test - Satisfaction, P1, Brik



Elderly People: Wilcoxon Rank Sum Test - Satisfaction, P2, Prisma

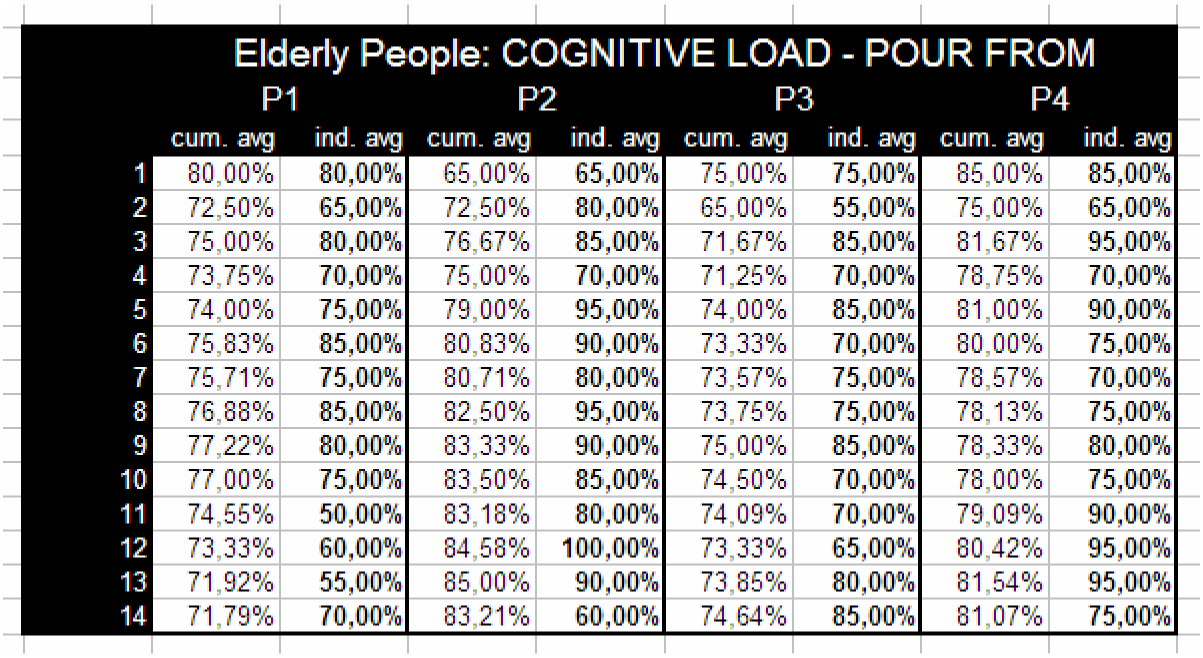


Elderly People: Wilcoxon Rank Sum Test - Satisfaction, P3, Gable



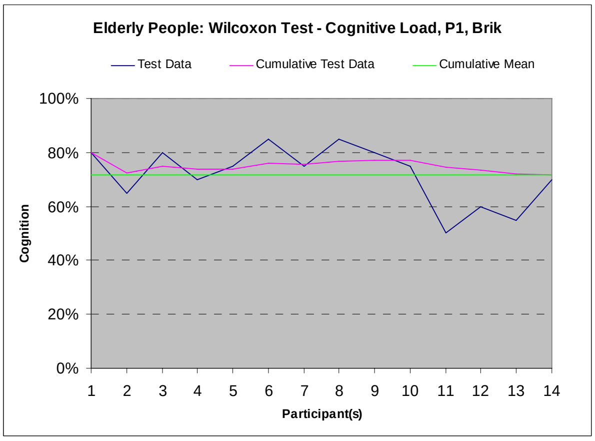
Elderly People: Wilcoxon Rank Sum Test - Satisfaction, P4, PET

Cognitive Load

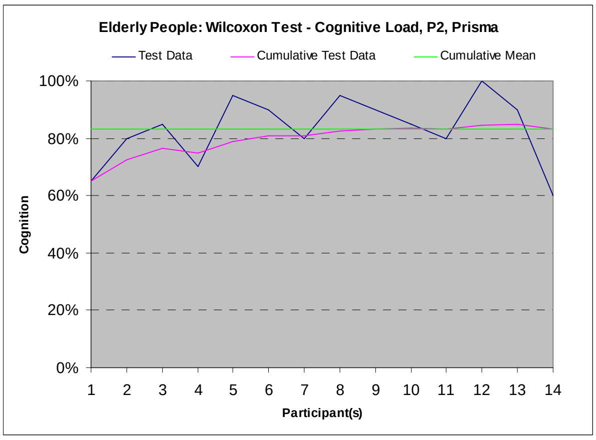


Elderly People: COGNITIVE LOAD - POUR FROM

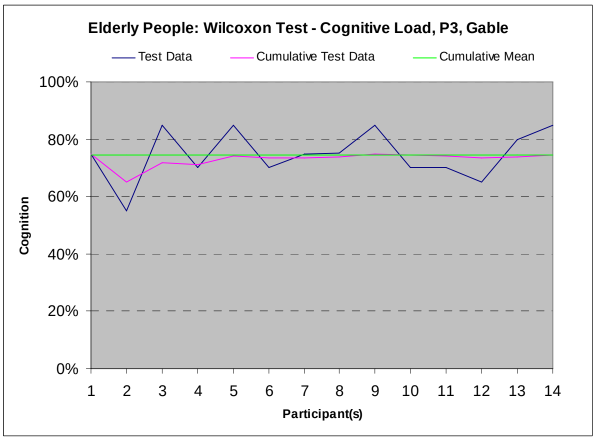
	P1		P2		P3		P4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	80,00%	80,00%	65,00%	65,00%	75,00%	75,00%	85,00%	85,00%
2	72,50%	65,00%	72,50%	80,00%	65,00%	55,00%	75,00%	65,00%
3	75,00%	80,00%	76,67%	85,00%	71,67%	85,00%	81,67%	95,00%
4	73,75%	70,00%	75,00%	70,00%	71,25%	70,00%	78,75%	70,00%
5	74,00%	75,00%	79,00%	95,00%	74,00%	85,00%	81,00%	90,00%
6	75,83%	85,00%	80,83%	90,00%	73,33%	70,00%	80,00%	75,00%
7	75,71%	75,00%	80,71%	80,00%	73,57%	75,00%	78,57%	70,00%
8	76,88%	85,00%	82,50%	95,00%	73,75%	75,00%	78,13%	75,00%
9	77,22%	80,00%	83,33%	90,00%	75,00%	85,00%	78,33%	80,00%
10	77,00%	75,00%	83,50%	85,00%	74,50%	70,00%	78,00%	75,00%
11	74,55%	50,00%	83,18%	80,00%	74,09%	70,00%	79,09%	90,00%
12	73,33%	60,00%	84,58%	100,00%	73,33%	65,00%	80,42%	95,00%
13	71,92%	55,00%	85,00%	90,00%	73,85%	80,00%	81,54%	95,00%
14	71,79%	70,00%	83,21%	60,00%	74,64%	85,00%	81,07%	75,00%



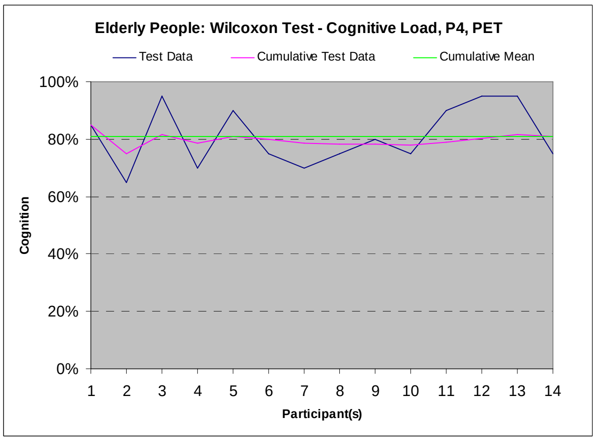
Elderly People: Wilcoxon Test - Cognitive Load, P1, Brik



Elderly People: Wilcoxon Test - Cognitive Load, P2, Prisma



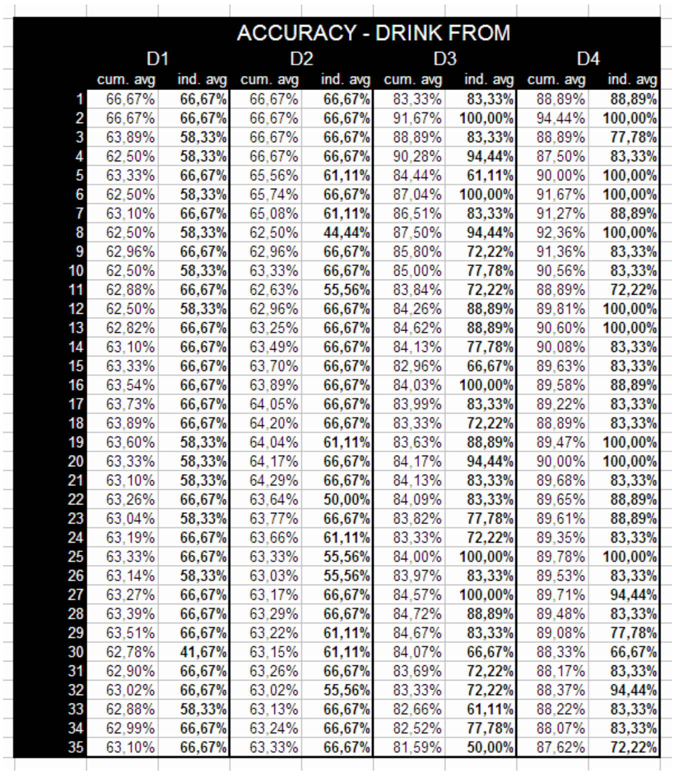
Elderly People: Wilcoxon Test - Cognitive Load, P3, Gable



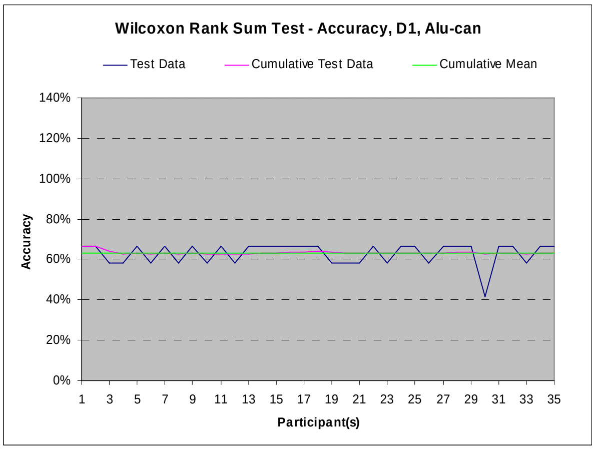
Elderly People: Wilcoxon Test - Cognitive Load, P4, PET

Drinking Test, Students (TP1-TP35)

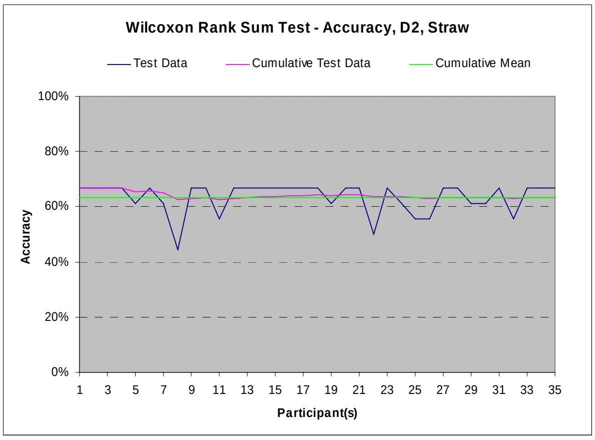
Accuracy – Exponential Reduction



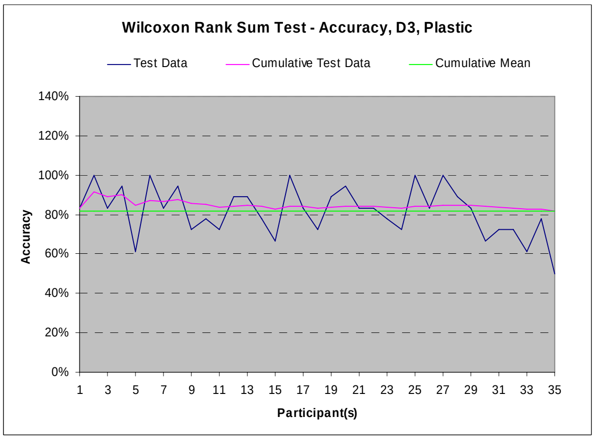
ACCURACY - DRINK FROM								
	D1		D2		D3		D4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	66,67%	66,67%	66,67%	66,67%	83,33%	83,33%	88,89%	88,89%
2	66,67%	66,67%	66,67%	66,67%	91,67%	100,00%	94,44%	100,00%
3	63,89%	58,33%	66,67%	66,67%	88,89%	83,33%	88,89%	77,78%
4	62,50%	58,33%	66,67%	66,67%	90,28%	94,44%	87,50%	83,33%
5	63,33%	66,67%	65,56%	61,11%	84,44%	61,11%	90,00%	100,00%
6	62,50%	58,33%	65,74%	66,67%	87,04%	100,00%	91,67%	100,00%
7	63,10%	66,67%	65,08%	61,11%	86,51%	83,33%	91,27%	88,89%
8	62,50%	58,33%	62,50%	44,44%	87,50%	94,44%	92,36%	100,00%
9	62,96%	66,67%	62,96%	66,67%	85,80%	72,22%	91,36%	83,33%
10	62,50%	58,33%	63,33%	66,67%	85,00%	77,78%	90,56%	83,33%
11	62,88%	66,67%	62,63%	55,56%	83,84%	72,22%	88,89%	72,22%
12	62,50%	58,33%	62,96%	66,67%	84,26%	88,89%	89,81%	100,00%
13	62,82%	66,67%	63,25%	66,67%	84,62%	88,89%	90,60%	100,00%
14	63,10%	66,67%	63,49%	66,67%	84,13%	77,78%	90,08%	83,33%
15	63,33%	66,67%	63,70%	66,67%	82,96%	66,67%	89,63%	83,33%
16	63,54%	66,67%	63,89%	66,67%	84,03%	100,00%	89,58%	88,89%
17	63,73%	66,67%	64,05%	66,67%	83,99%	83,33%	89,22%	83,33%
18	63,89%	66,67%	64,20%	66,67%	83,33%	72,22%	88,89%	83,33%
19	63,60%	58,33%	64,04%	61,11%	83,63%	88,89%	89,47%	100,00%
20	63,33%	58,33%	64,17%	66,67%	84,17%	94,44%	90,00%	100,00%
21	63,10%	58,33%	64,29%	66,67%	84,13%	83,33%	89,68%	83,33%
22	63,26%	66,67%	63,64%	50,00%	84,09%	83,33%	89,65%	88,89%
23	63,04%	58,33%	63,77%	66,67%	83,82%	77,78%	89,61%	88,89%
24	63,19%	66,67%	63,66%	61,11%	83,33%	72,22%	89,35%	83,33%
25	63,33%	66,67%	63,33%	55,56%	84,00%	100,00%	89,78%	100,00%
26	63,14%	58,33%	63,03%	55,56%	83,97%	83,33%	89,53%	83,33%
27	63,27%	66,67%	63,17%	66,67%	84,57%	100,00%	89,71%	94,44%
28	63,39%	66,67%	63,29%	66,67%	84,72%	88,89%	89,48%	83,33%
29	63,51%	66,67%	63,22%	61,11%	84,67%	83,33%	89,08%	77,78%
30	62,78%	41,67%	63,15%	61,11%	84,07%	66,67%	88,33%	66,67%
31	62,90%	66,67%	63,26%	66,67%	83,69%	72,22%	88,17%	83,33%
32	63,02%	66,67%	63,02%	55,56%	83,33%	72,22%	88,37%	94,44%
33	62,88%	58,33%	63,13%	66,67%	82,66%	61,11%	88,22%	83,33%
34	62,99%	66,67%	63,24%	66,67%	82,52%	77,78%	88,07%	83,33%
35	63,10%	66,67%	63,33%	66,67%	81,59%	50,00%	87,62%	72,22%



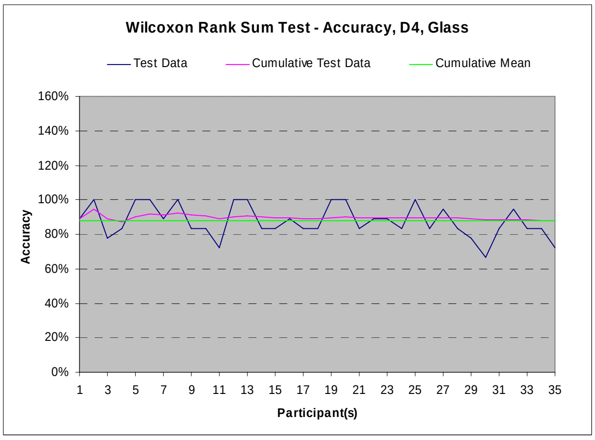
Wilcoxon Rank Sum Test - Accuracy, D1, Alu-can



Wilcoxon Rank Sum Test - Accuracy, D2, Straw

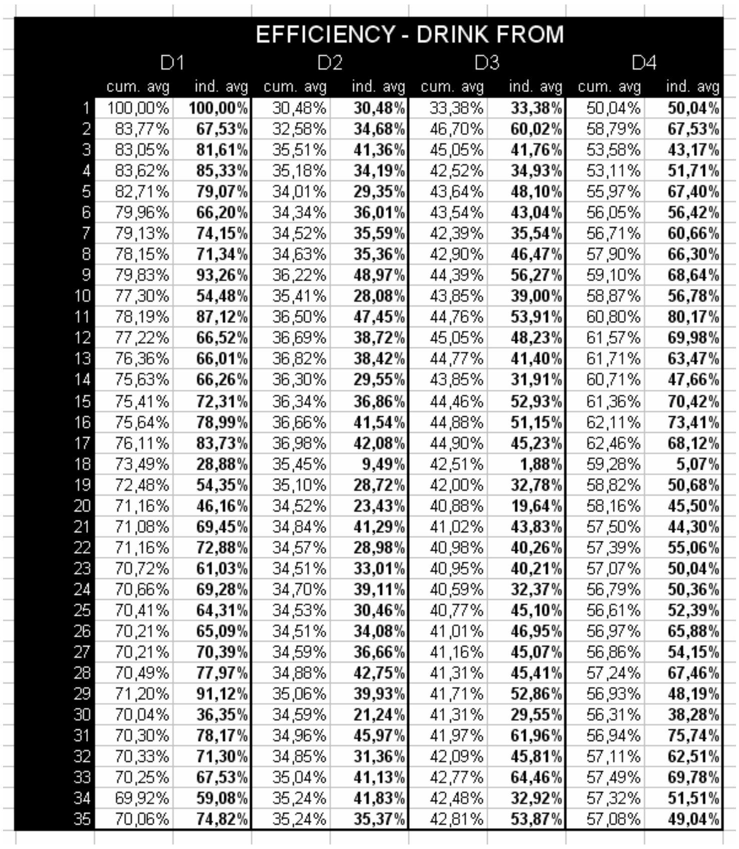


Wilcoxon Rank Sum Test - Accuracy, D3, Plastic

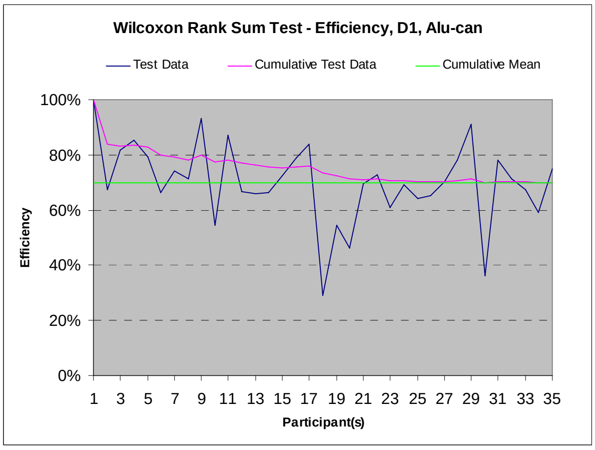


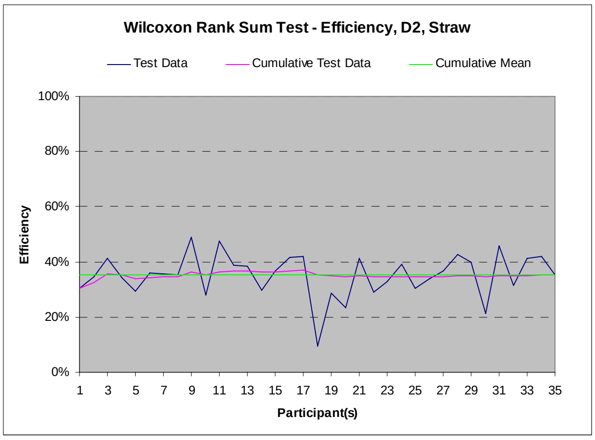
Wilcoxon Rank Sum Test - Accuracy, D4, Glass

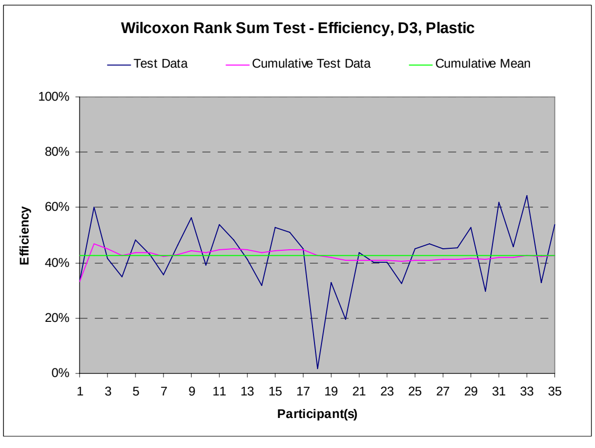
Efficiency – Handling Time

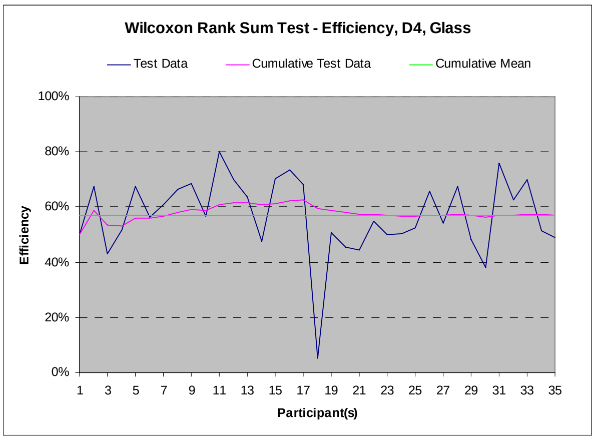


EFFICIENCY - DRINK FROM								
	D1		D2		D3		D4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	100,00%	**100,00%**	30,48%	**30,48%**	33,38%	**33,38%**	50,04%	**50,04%**
2	83,77%	**67,53%**	32,58%	**34,68%**	46,70%	**60,02%**	58,79%	**67,53%**
3	83,05%	**81,61%**	35,51%	**41,36%**	45,05%	**41,76%**	53,58%	**43,17%**
4	83,62%	**85,33%**	35,18%	**34,19%**	42,52%	**34,93%**	53,11%	**51,71%**
5	82,71%	**79,07%**	34,01%	**29,35%**	43,64%	**48,10%**	55,97%	**67,40%**
6	79,96%	**66,20%**	34,34%	**36,01%**	43,54%	**43,04%**	56,05%	**56,42%**
7	79,13%	**74,15%**	34,52%	**35,59%**	42,39%	**35,54%**	56,71%	**60,66%**
8	78,15%	**71,34%**	34,63%	**35,36%**	42,90%	**46,47%**	57,90%	**66,30%**
9	79,83%	**93,26%**	36,22%	**48,97%**	44,39%	**56,27%**	59,10%	**68,64%**
10	77,30%	**54,48%**	35,41%	**28,08%**	43,85%	**39,00%**	58,87%	**56,78%**
11	78,19%	**87,12%**	36,50%	**47,45%**	44,76%	**53,91%**	60,80%	**80,17%**
12	77,22%	**66,52%**	36,69%	**38,72%**	45,05%	**48,23%**	61,57%	**69,98%**
13	76,36%	**66,01%**	36,82%	**38,42%**	44,77%	**41,40%**	61,71%	**63,47%**
14	75,63%	**66,26%**	36,30%	**29,55%**	43,85%	**31,91%**	60,71%	**47,66%**
15	75,41%	**72,31%**	36,34%	**36,86%**	44,46%	**52,93%**	61,36%	**70,42%**
16	75,64%	**78,99%**	36,66%	**41,54%**	44,88%	**51,15%**	62,11%	**73,41%**
17	76,11%	**83,73%**	36,98%	**42,08%**	44,90%	**45,23%**	62,46%	**68,12%**
18	73,49%	**28,88%**	35,45%	**9,49%**	42,51%	**1,88%**	59,28%	**5,07%**
19	72,48%	**54,35%**	35,10%	**28,72%**	42,00%	**32,78%**	58,82%	**50,68%**
20	71,16%	**46,16%**	34,52%	**23,43%**	40,88%	**19,64%**	58,16%	**45,50%**
21	71,08%	**69,45%**	34,84%	**41,29%**	41,02%	**43,83%**	57,50%	**44,30%**
22	71,16%	**72,88%**	34,57%	**28,98%**	40,98%	**40,26%**	57,39%	**55,06%**
23	70,72%	**61,03%**	34,51%	**33,01%**	40,95%	**40,21%**	57,07%	**50,04%**
24	70,66%	**69,28%**	34,70%	**39,11%**	40,59%	**32,37%**	56,79%	**50,36%**
25	70,41%	**64,31%**	34,53%	**30,46%**	40,77%	**45,10%**	56,61%	**52,39%**
26	70,21%	**65,09%**	34,51%	**34,08%**	41,01%	**46,95%**	56,97%	**65,88%**
27	70,21%	**70,39%**	34,59%	**36,66%**	41,16%	**45,07%**	56,86%	**54,15%**
28	70,49%	**77,97%**	34,88%	**42,75%**	41,31%	**45,41%**	57,24%	**67,46%**
29	71,20%	**91,12%**	35,06%	**39,93%**	41,71%	**52,86%**	56,93%	**48,19%**
30	70,04%	**36,35%**	34,59%	**21,24%**	41,31%	**29,55%**	56,31%	**38,28%**
31	70,30%	**78,17%**	34,96%	**45,97%**	41,97%	**61,96%**	56,94%	**75,74%**
32	70,33%	**71,30%**	34,85%	**31,36%**	42,09%	**45,81%**	57,11%	**62,51%**
33	70,25%	**67,53%**	35,04%	**41,13%**	42,77%	**64,46%**	57,49%	**69,78%**
34	69,92%	**59,08%**	35,24%	**41,83%**	42,48%	**32,92%**	57,32%	**51,51%**
35	70,06%	**74,82%**	35,24%	**35,37%**	42,81%	**53,87%**	57,08%	**49,04%**

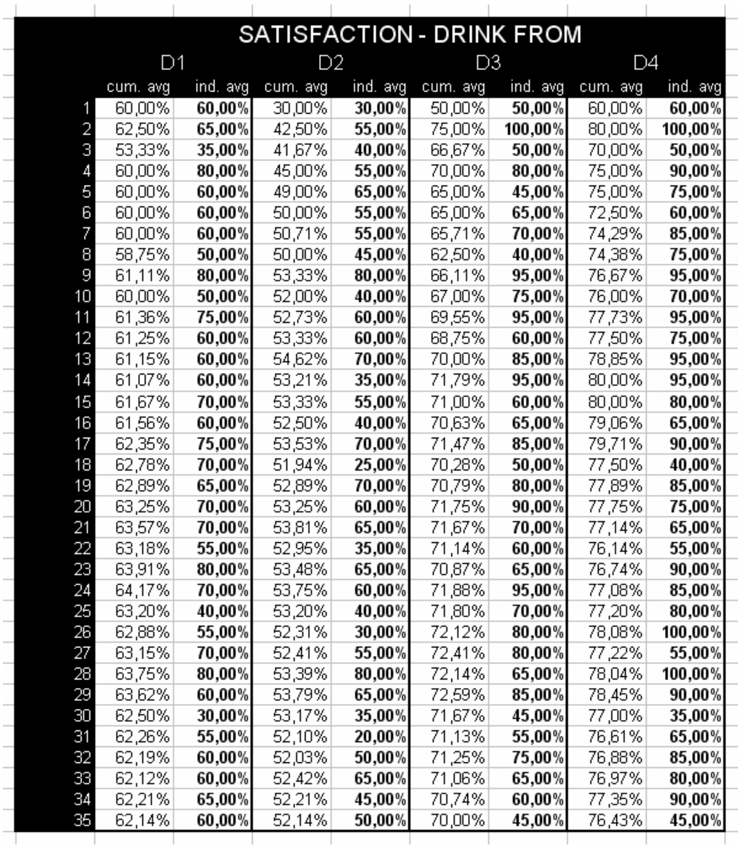

Wilcoxon Rank Sum Test - Efficiency, D1, Alu-can


Wilcoxon Rank Sum Test - Efficiency, D2, Straw

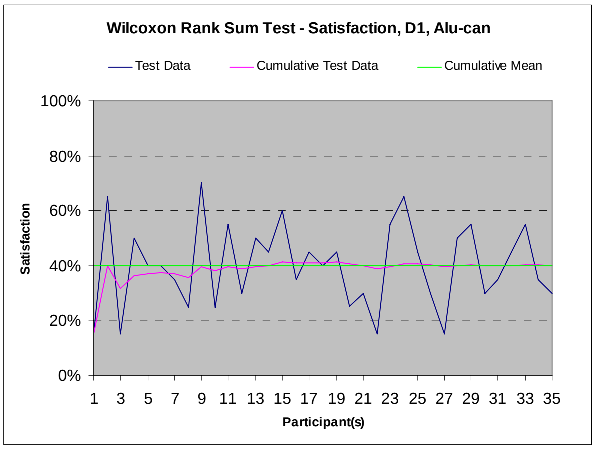

Wilcoxon Rank Sum Test - Efficiency, D3, Plastic

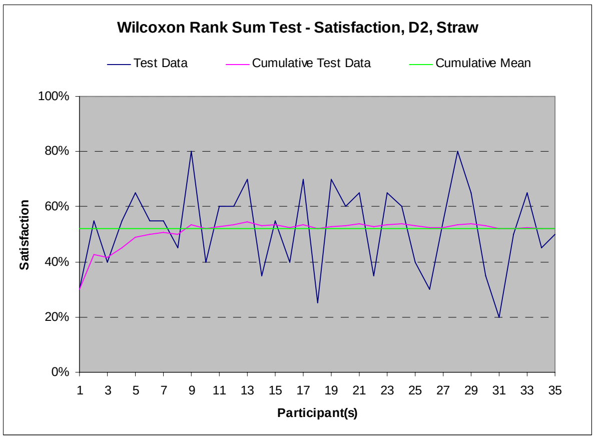

Wilcoxon Rank Sum Test - Efficiency, D4, Glass

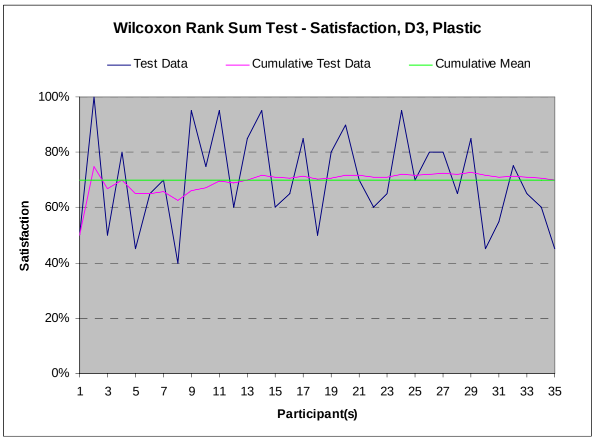
Satisfaction

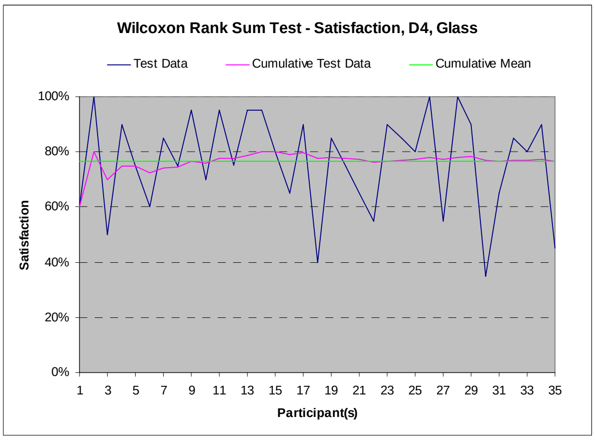


SATISFACTION - DRINK FROM								
	D1		D2		D3		D4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	60,00%	**60,00%**	30,00%	**30,00%**	50,00%	**50,00%**	60,00%	**60,00%**
2	62,50%	**65,00%**	42,50%	**55,00%**	75,00%	**100,00%**	80,00%	**100,00%**
3	53,33%	**35,00%**	41,67%	**40,00%**	66,67%	**50,00%**	70,00%	**50,00%**
4	60,00%	**80,00%**	45,00%	**55,00%**	70,00%	**80,00%**	75,00%	**90,00%**
5	60,00%	**60,00%**	49,00%	**65,00%**	65,00%	**45,00%**	75,00%	**75,00%**
6	60,00%	**60,00%**	50,00%	**55,00%**	65,00%	**65,00%**	72,50%	**60,00%**
7	60,00%	**60,00%**	50,71%	**55,00%**	65,71%	**70,00%**	74,29%	**85,00%**
8	58,75%	**50,00%**	50,00%	**45,00%**	62,50%	**40,00%**	74,38%	**75,00%**
9	61,11%	**80,00%**	53,33%	**80,00%**	66,11%	**95,00%**	76,67%	**95,00%**
10	60,00%	**50,00%**	52,00%	**40,00%**	67,00%	**75,00%**	76,00%	**70,00%**
11	61,36%	**75,00%**	52,73%	**60,00%**	69,55%	**95,00%**	77,73%	**95,00%**
12	61,25%	**60,00%**	53,33%	**60,00%**	68,75%	**60,00%**	77,50%	**75,00%**
13	61,15%	**60,00%**	54,62%	**70,00%**	70,00%	**85,00%**	78,85%	**95,00%**
14	61,07%	**60,00%**	53,21%	**35,00%**	71,79%	**95,00%**	80,00%	**95,00%**
15	61,67%	**70,00%**	53,33%	**55,00%**	71,00%	**60,00%**	80,00%	**80,00%**
16	61,56%	**60,00%**	52,50%	**40,00%**	70,63%	**65,00%**	79,06%	**65,00%**
17	62,35%	**75,00%**	53,53%	**70,00%**	71,47%	**85,00%**	79,71%	**90,00%**
18	62,78%	**70,00%**	51,94%	**25,00%**	70,28%	**50,00%**	77,50%	**40,00%**
19	62,89%	**65,00%**	52,89%	**70,00%**	70,79%	**80,00%**	77,89%	**85,00%**
20	63,25%	**70,00%**	53,25%	**60,00%**	71,75%	**90,00%**	77,75%	**75,00%**
21	63,57%	**70,00%**	53,81%	**65,00%**	71,67%	**70,00%**	77,14%	**65,00%**
22	63,18%	**55,00%**	52,95%	**35,00%**	71,14%	**60,00%**	76,14%	**55,00%**
23	63,91%	**80,00%**	53,48%	**65,00%**	70,87%	**65,00%**	76,74%	**90,00%**
24	64,17%	**70,00%**	53,75%	**60,00%**	71,88%	**95,00%**	77,08%	**85,00%**
25	63,20%	**40,00%**	53,20%	**40,00%**	71,80%	**70,00%**	77,20%	**80,00%**
26	62,88%	**55,00%**	52,31%	**30,00%**	72,12%	**80,00%**	78,08%	**100,00%**
27	63,15%	**70,00%**	52,41%	**55,00%**	72,41%	**80,00%**	77,22%	**55,00%**
28	63,75%	**80,00%**	53,39%	**80,00%**	72,14%	**65,00%**	78,04%	**100,00%**
29	63,62%	**60,00%**	53,79%	**65,00%**	72,59%	**85,00%**	78,45%	**90,00%**
30	62,50%	**30,00%**	53,17%	**35,00%**	71,67%	**45,00%**	77,00%	**35,00%**
31	62,26%	**55,00%**	52,10%	**20,00%**	71,13%	**55,00%**	76,61%	**65,00%**
32	62,19%	**60,00%**	52,03%	**50,00%**	71,25%	**75,00%**	76,88%	**85,00%**
33	62,12%	**60,00%**	52,42%	**65,00%**	71,06%	**65,00%**	76,97%	**80,00%**
34	62,21%	**65,00%**	52,21%	**45,00%**	70,74%	**60,00%**	77,35%	**90,00%**
35	62,14%	**60,00%**	52,14%	**50,00%**	70,00%	**45,00%**	76,43%	**45,00%**

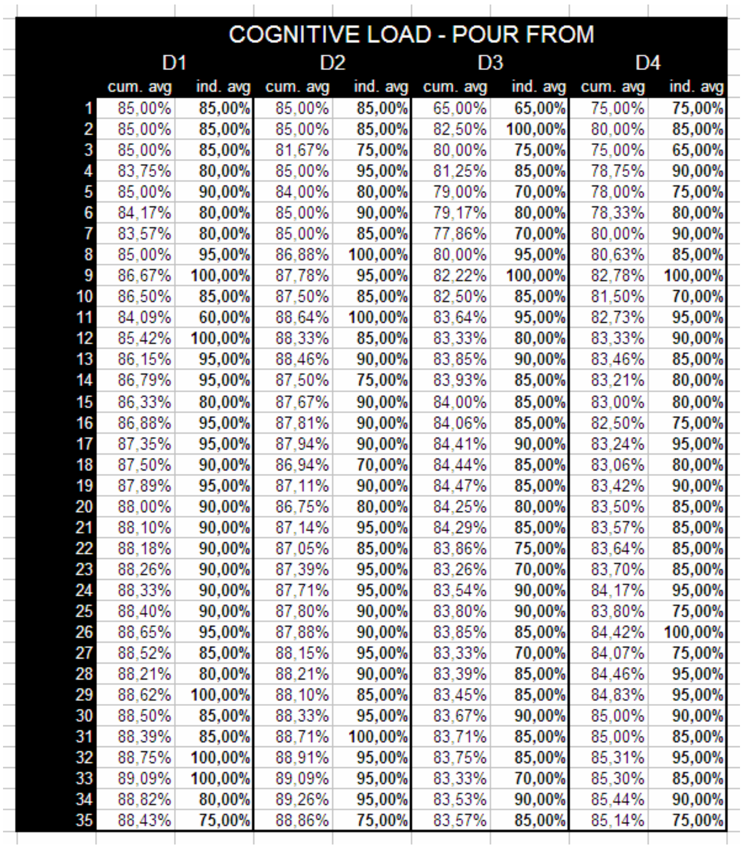




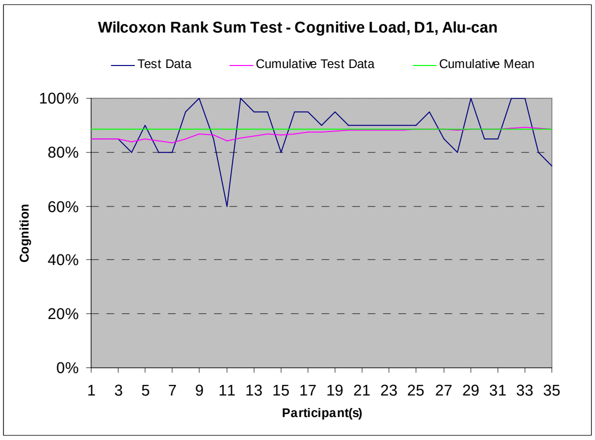




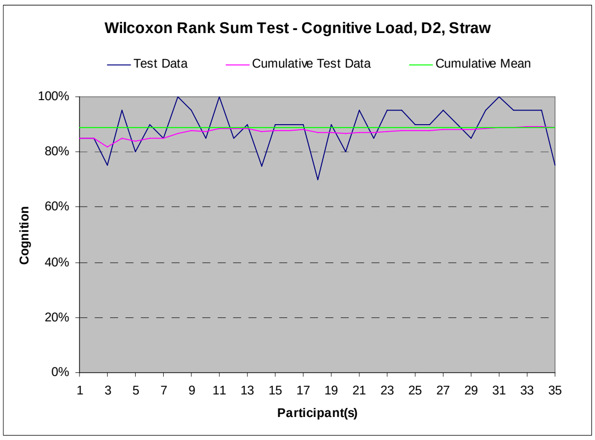
Cognitive Load



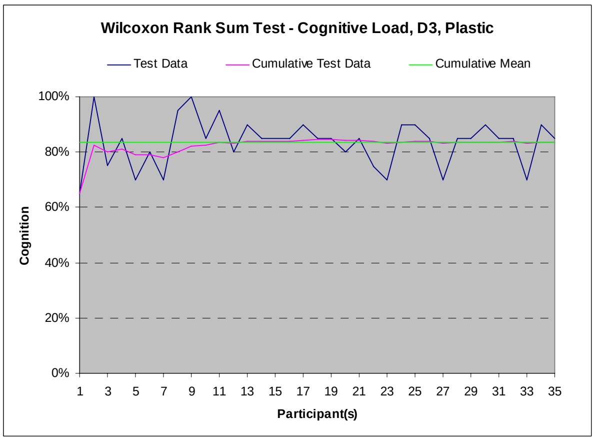
COGNITIVE LOAD - POUR FROM								
	D1		D2		D3		D4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	85,00%	85,00%	85,00%	85,00%	65,00%	65,00%	75,00%	75,00%
2	85,00%	85,00%	85,00%	85,00%	82,50%	100,00%	80,00%	85,00%
3	85,00%	85,00%	81,67%	75,00%	80,00%	75,00%	75,00%	65,00%
4	83,75%	80,00%	85,00%	95,00%	81,25%	85,00%	78,75%	90,00%
5	85,00%	90,00%	84,00%	80,00%	79,00%	70,00%	78,00%	75,00%
6	84,17%	80,00%	85,00%	90,00%	79,17%	80,00%	78,33%	80,00%
7	83,57%	80,00%	85,00%	85,00%	77,86%	70,00%	80,00%	90,00%
8	85,00%	95,00%	86,88%	100,00%	80,00%	95,00%	80,63%	85,00%
9	86,67%	100,00%	87,78%	95,00%	82,22%	100,00%	82,78%	100,00%
10	86,50%	85,00%	87,50%	85,00%	82,50%	85,00%	81,50%	70,00%
11	84,09%	60,00%	88,64%	100,00%	83,64%	95,00%	82,73%	95,00%
12	85,42%	100,00%	88,33%	85,00%	83,33%	80,00%	83,33%	90,00%
13	86,15%	95,00%	88,46%	90,00%	83,85%	90,00%	83,46%	85,00%
14	86,79%	95,00%	87,50%	75,00%	83,93%	85,00%	83,21%	80,00%
15	86,33%	80,00%	87,67%	90,00%	84,00%	85,00%	83,00%	80,00%
16	86,88%	95,00%	87,81%	90,00%	84,06%	85,00%	82,50%	75,00%
17	87,35%	95,00%	87,94%	90,00%	84,41%	90,00%	83,24%	95,00%
18	87,50%	90,00%	86,94%	70,00%	84,44%	85,00%	83,06%	80,00%
19	87,89%	95,00%	87,11%	90,00%	84,47%	85,00%	83,42%	90,00%
20	88,00%	90,00%	86,75%	80,00%	84,25%	80,00%	83,50%	85,00%
21	88,10%	90,00%	87,14%	95,00%	84,29%	85,00%	83,57%	85,00%
22	88,18%	90,00%	87,05%	85,00%	83,86%	75,00%	83,64%	85,00%
23	88,26%	90,00%	87,39%	95,00%	83,26%	70,00%	83,70%	85,00%
24	88,33%	90,00%	87,71%	95,00%	83,54%	90,00%	84,17%	95,00%
25	88,40%	90,00%	87,80%	90,00%	83,80%	90,00%	83,80%	75,00%
26	88,65%	95,00%	87,88%	90,00%	83,85%	85,00%	84,42%	100,00%
27	88,52%	85,00%	88,15%	95,00%	83,33%	70,00%	84,07%	75,00%
28	88,21%	80,00%	88,21%	90,00%	83,39%	85,00%	84,46%	95,00%
29	88,62%	100,00%	88,10%	85,00%	83,45%	85,00%	84,83%	95,00%
30	88,50%	85,00%	88,33%	95,00%	83,67%	90,00%	85,00%	90,00%
31	88,39%	85,00%	88,71%	100,00%	83,71%	85,00%	85,00%	85,00%
32	88,75%	100,00%	88,91%	95,00%	83,75%	85,00%	85,31%	95,00%
33	89,09%	100,00%	89,09%	95,00%	83,33%	70,00%	85,30%	85,00%
34	88,82%	80,00%	89,26%	95,00%	83,53%	90,00%	85,44%	90,00%
35	88,43%	75,00%	88,86%	75,00%	83,57%	85,00%	85,14%	75,00%



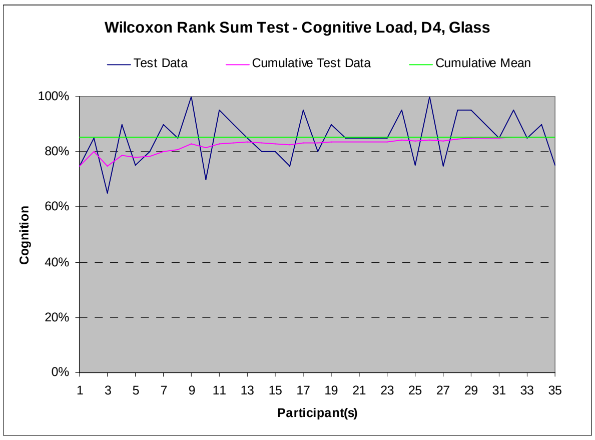
Wilcoxon Rank Sum Test - Cognitive Load, D1, Alu-can



Wilcoxon Rank Sum Test - Cognitive Load, D2, Straw



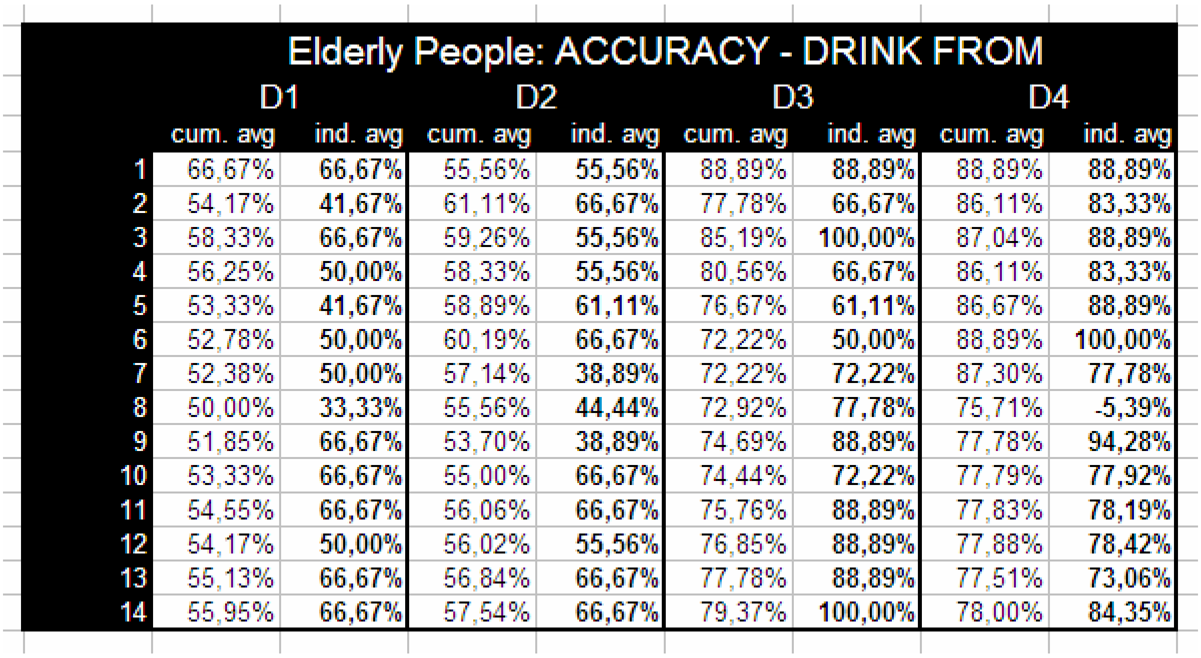
Wilcoxon Rank Sum Test - Cognitive Load, D3, Plastic



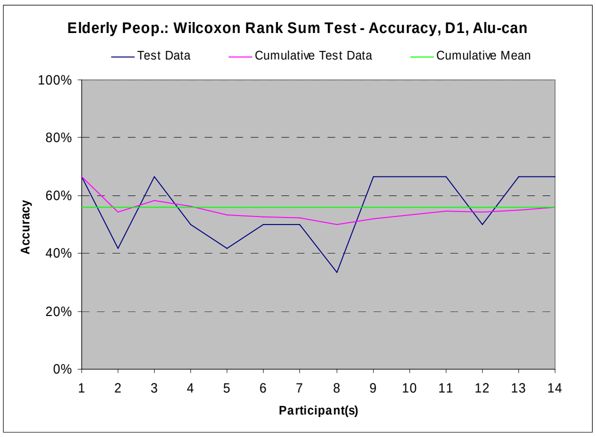
Wilcoxon Rank Sum Test - Cognitive Load, D4, Glass

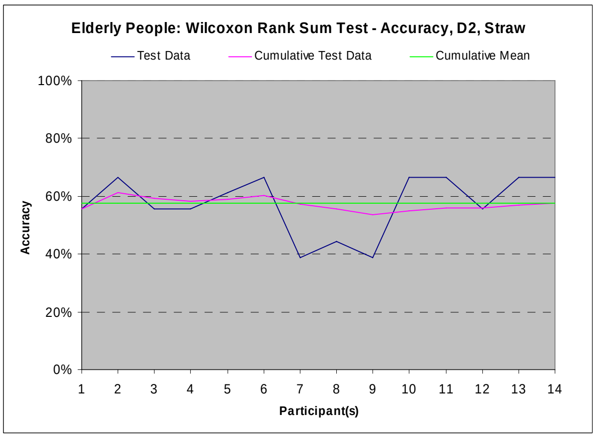
Drinking Test, Elderly People (TP36-TP49)

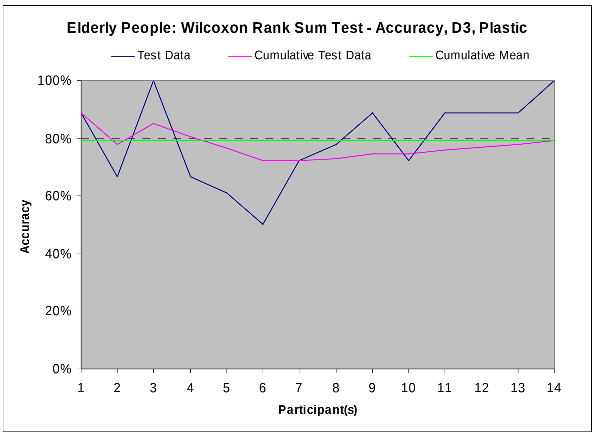
Accuracy – Exponential Reduction

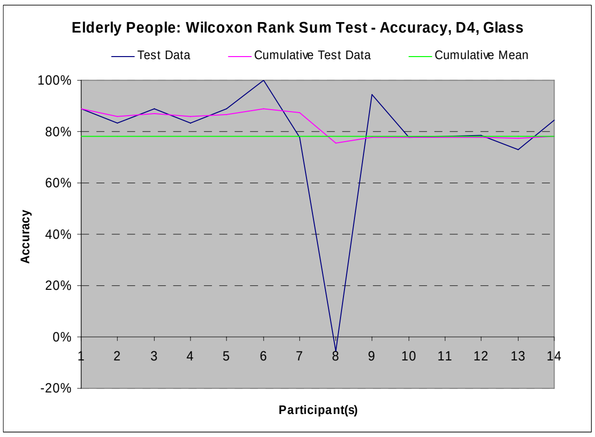


Elderly People: ACCURACY - DRINK FROM	D1		D2		D3		D4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	66,67%	66,67%	55,56%	55,56%	88,89%	88,89%	88,89%	88,89%
2	54,17%	41,67%	61,11%	66,67%	77,78%	66,67%	86,11%	83,33%
3	58,33%	66,67%	59,26%	55,56%	85,19%	100,00%	87,04%	88,89%
4	56,25%	50,00%	58,33%	55,56%	80,56%	66,67%	86,11%	83,33%
5	53,33%	41,67%	58,89%	61,11%	76,67%	61,11%	86,67%	88,89%
6	52,78%	50,00%	60,19%	66,67%	72,22%	50,00%	88,89%	100,00%
7	52,38%	50,00%	57,14%	38,89%	72,22%	72,22%	87,30%	77,78%
8	50,00%	33,33%	55,56%	44,44%	72,92%	77,78%	75,71%	-5,39%
9	51,85%	66,67%	53,70%	38,89%	74,69%	88,89%	77,78%	94,28%
10	53,33%	66,67%	55,00%	66,67%	74,44%	72,22%	77,79%	77,92%
11	54,55%	66,67%	56,06%	66,67%	75,76%	88,89%	77,83%	78,19%
12	54,17%	50,00%	56,02%	55,56%	76,85%	88,89%	77,88%	78,42%
13	55,13%	66,67%	56,84%	66,67%	77,78%	88,89%	77,51%	73,06%
14	55,95%	66,67%	57,54%	66,67%	79,37%	100,00%	78,00%	84,35%

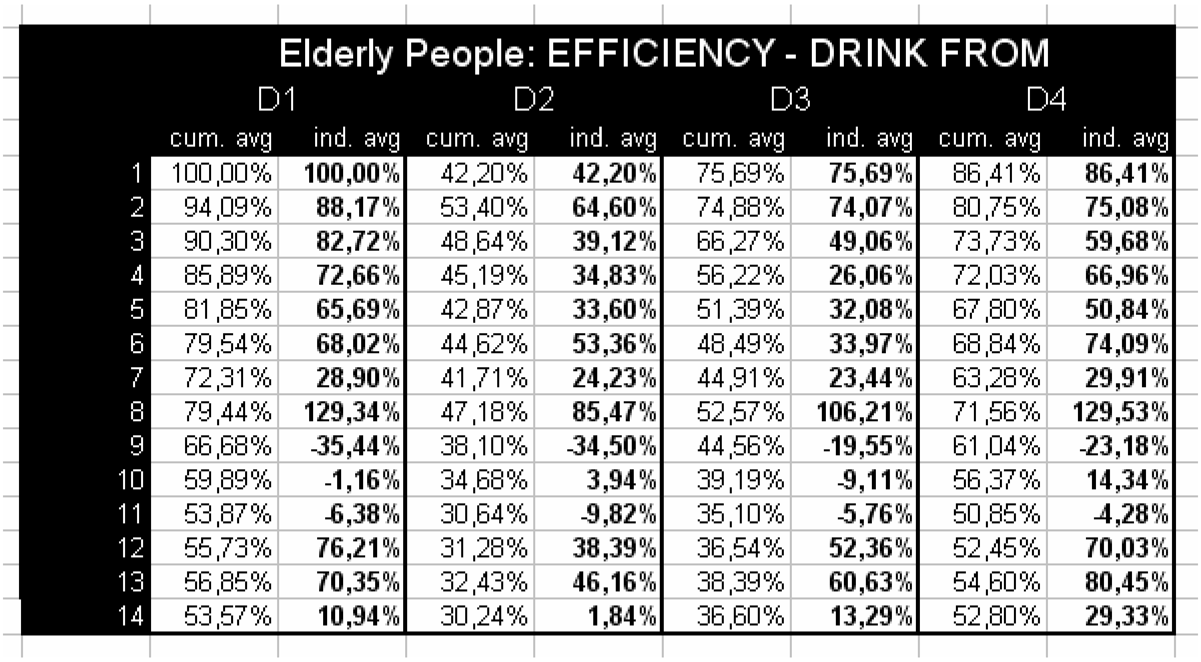






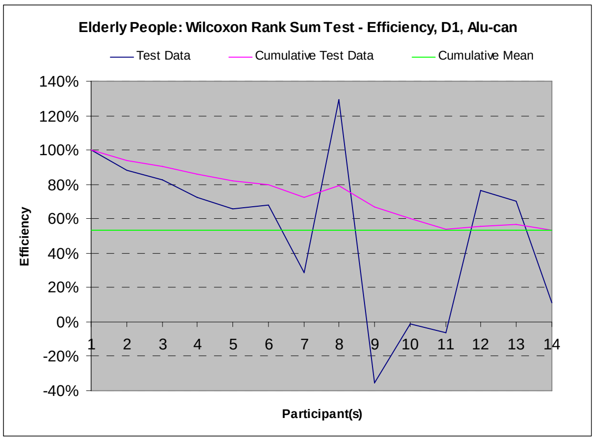


Efficiency – Handling Time

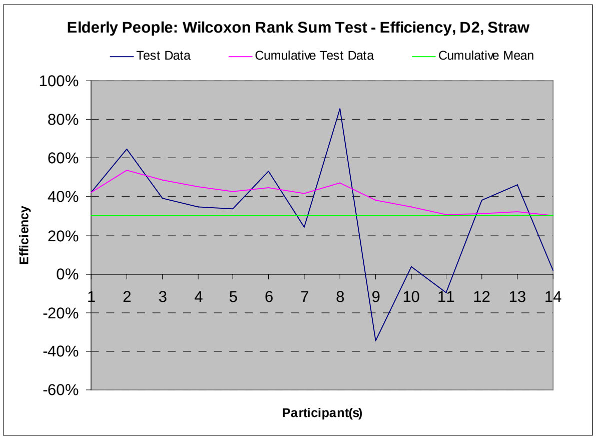


Elderly People: EFFICIENCY - DRINK FROM

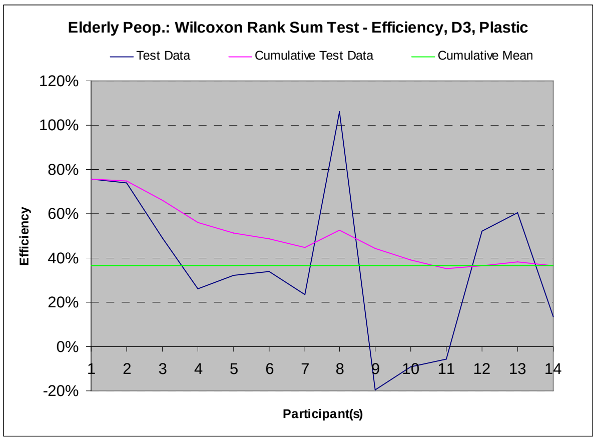
	D1		D2		D3		D4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	100,00%	**100,00%**	42,20%	**42,20%**	75,69%	**75,69%**	86,41%	**86,41%**
2	94,09%	**88,17%**	53,40%	**64,60%**	74,88%	**74,07%**	80,75%	**75,08%**
3	90,30%	**82,72%**	48,64%	**39,12%**	66,27%	**49,06%**	73,73%	**59,68%**
4	85,89%	**72,66%**	45,19%	**34,83%**	56,22%	**26,06%**	72,03%	**66,96%**
5	81,85%	**65,69%**	42,87%	**33,60%**	51,39%	**32,08%**	67,80%	**50,84%**
6	79,54%	**68,02%**	44,62%	**53,36%**	48,49%	**33,97%**	68,84%	**74,09%**
7	72,31%	**28,90%**	41,71%	**24,23%**	44,91%	**23,44%**	63,28%	**29,91%**
8	79,44%	**129,34%**	47,18%	**85,47%**	52,57%	**106,21%**	71,56%	**129,53%**
9	66,68%	**-35,44%**	38,10%	**-34,50%**	44,56%	**-19,55%**	61,04%	**-23,18%**
10	59,89%	**-1,16%**	34,68%	**3,94%**	39,19%	**-9,11%**	56,37%	**14,34%**
11	53,87%	**-6,38%**	30,64%	**-9,82%**	35,10%	**-5,76%**	50,85%	**-4,28%**
12	55,73%	**76,21%**	31,28%	**38,39%**	36,54%	**52,36%**	52,45%	**70,03%**
13	56,85%	**70,35%**	32,43%	**46,16%**	38,39%	**60,63%**	54,60%	**80,45%**
14	53,57%	**10,94%**	30,24%	**1,84%**	36,60%	**13,29%**	52,80%	**29,33%**



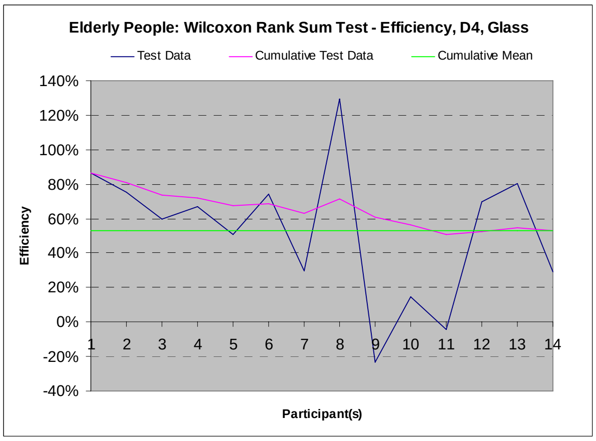
Elderly People: Wilcoxon Rank Sum Test - Efficiency, D1, Alu-can



Elderly People: Wilcoxon Rank Sum Test - Efficiency, D2, Straw

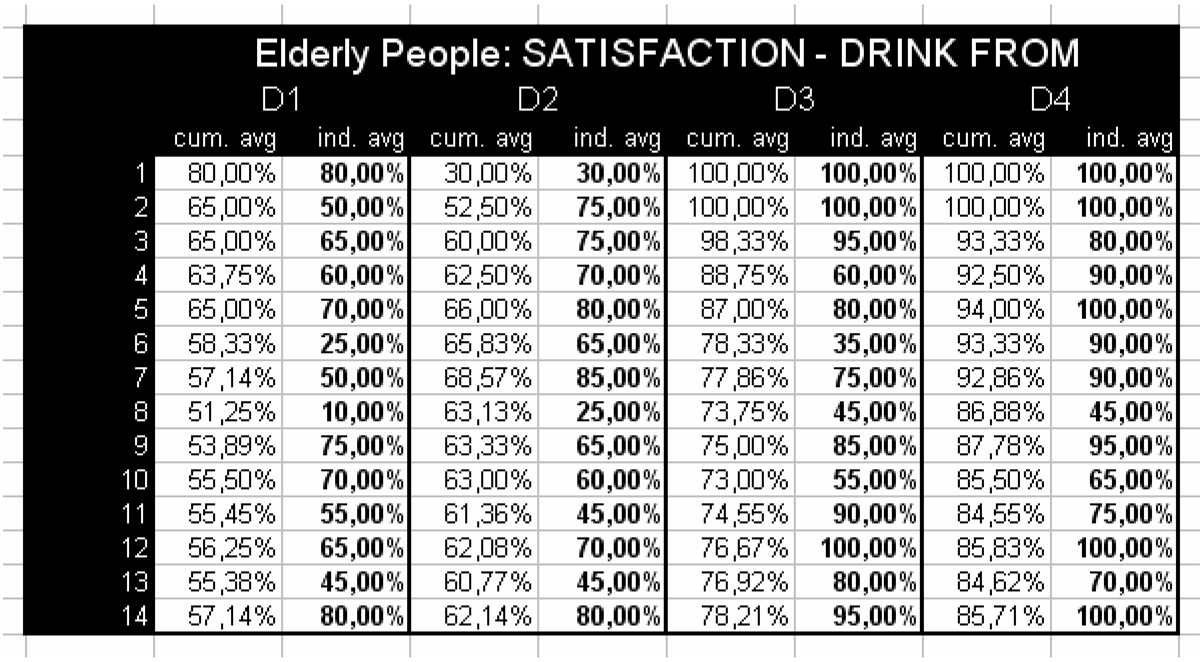


Elderly Peop.: Wilcoxon Rank Sum Test - Efficiency, D3, Plastic

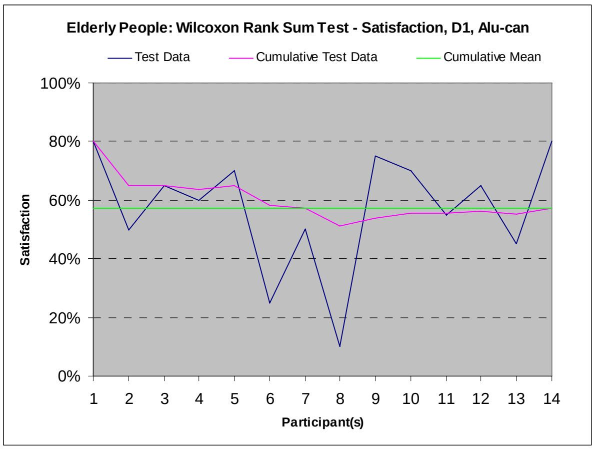


Elderly People: Wilcoxon Rank Sum Test - Efficiency, D4, Glass

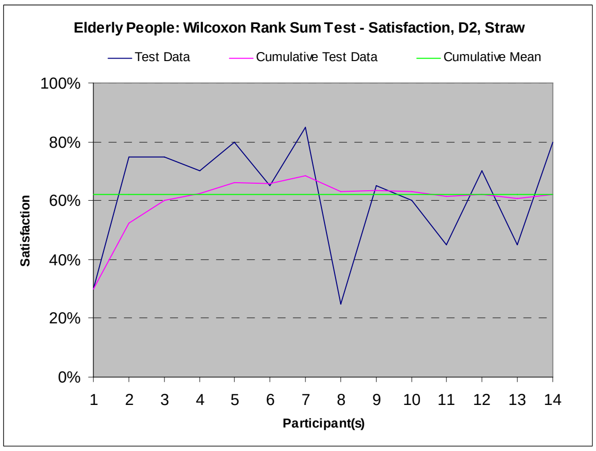
Satisfaction



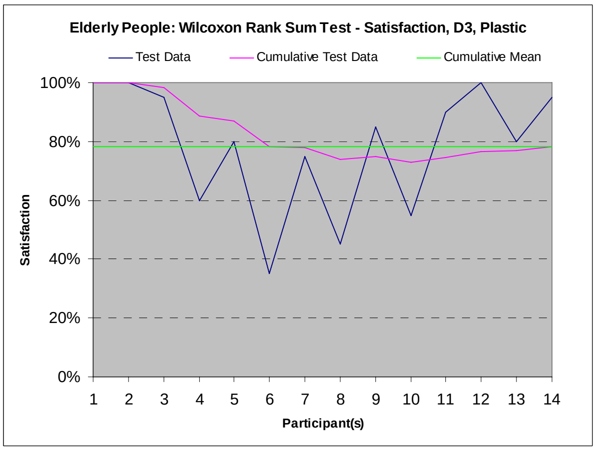
Elderly People: SATISFACTION - DRINK FROM								
	D1		D2		D3		D4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	80,00%	**80,00%**	30,00%	**30,00%**	100,00%	**100,00%**	100,00%	**100,00%**
2	65,00%	**50,00%**	52,50%	**75,00%**	100,00%	**100,00%**	100,00%	**100,00%**
3	65,00%	**65,00%**	60,00%	**75,00%**	98,33%	**95,00%**	93,33%	**80,00%**
4	63,75%	**60,00%**	62,50%	**70,00%**	88,75%	**60,00%**	92,50%	**90,00%**
5	65,00%	**70,00%**	66,00%	**80,00%**	87,00%	**80,00%**	94,00%	**100,00%**
6	58,33%	**25,00%**	65,83%	**65,00%**	78,33%	**35,00%**	93,33%	**90,00%**
7	57,14%	**50,00%**	68,57%	**85,00%**	77,86%	**75,00%**	92,86%	**90,00%**
8	51,25%	**10,00%**	63,13%	**25,00%**	73,75%	**45,00%**	86,88%	**45,00%**
9	53,89%	**75,00%**	63,33%	**65,00%**	75,00%	**85,00%**	87,78%	**95,00%**
10	55,50%	**70,00%**	63,00%	**60,00%**	73,00%	**55,00%**	85,50%	**65,00%**
11	55,45%	**55,00%**	61,36%	**45,00%**	74,55%	**90,00%**	84,55%	**75,00%**
12	56,25%	**65,00%**	62,08%	**70,00%**	76,67%	**100,00%**	85,83%	**100,00%**
13	55,38%	**45,00%**	60,77%	**45,00%**	76,92%	**80,00%**	84,62%	**70,00%**
14	57,14%	**80,00%**	62,14%	**80,00%**	78,21%	**95,00%**	85,71%	**100,00%**



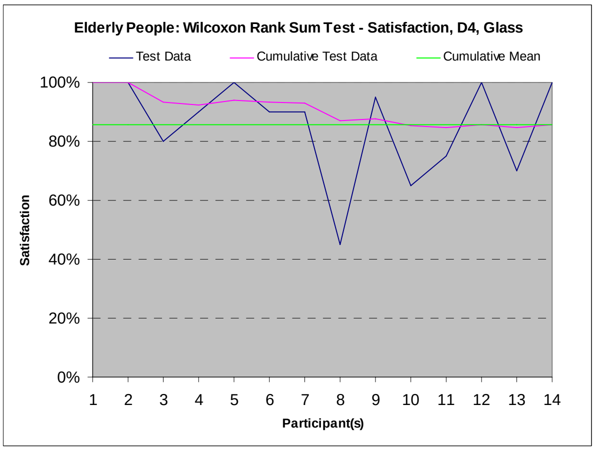
Elderly People: Wilcoxon Rank Sum Test - Satisfaction, D1, Alu-can



Elderly People: Wilcoxon Rank Sum Test - Satisfaction, D2, Straw

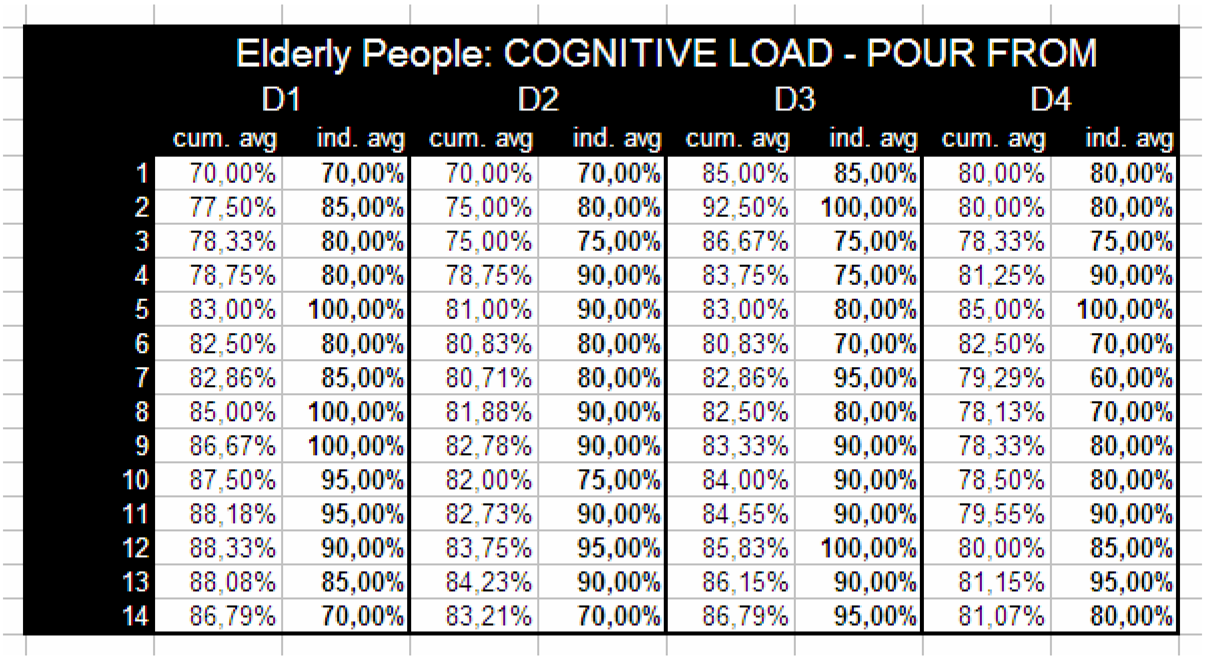


Elderly People: Wilcoxon Rank Sum Test - Satisfaction, D3, Plastic

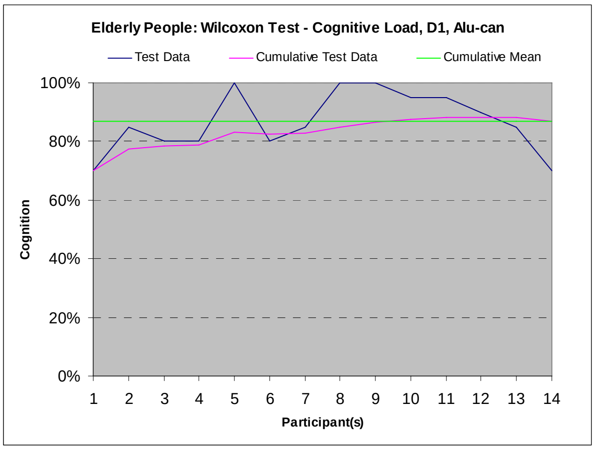


Elderly People: Wilcoxon Rank Sum Test - Satisfaction, D4, Glass

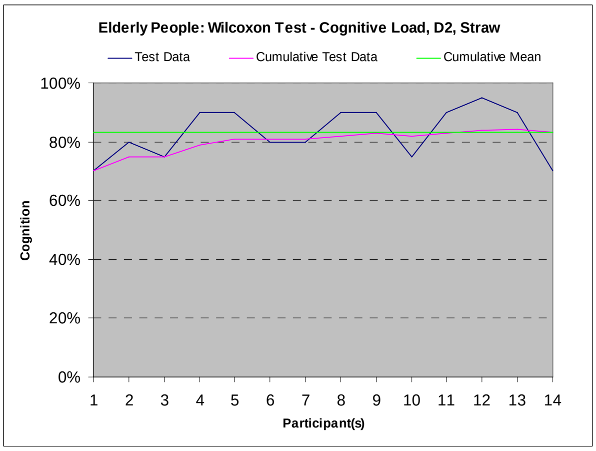
Cognitive Load



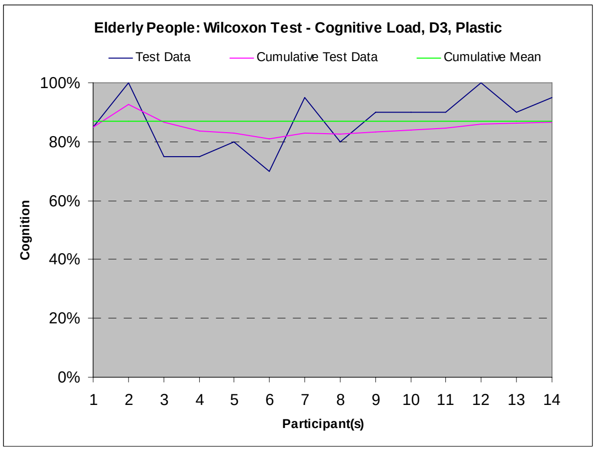
Elderly People: COGNITIVE LOAD - POUR FROM	D1		D2		D3		D4	
	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg	cum. avg	ind. avg
1	70,00%	70,00%	70,00%	70,00%	85,00%	85,00%	80,00%	80,00%
2	77,50%	85,00%	75,00%	80,00%	92,50%	100,00%	80,00%	80,00%
3	78,33%	80,00%	75,00%	75,00%	86,67%	75,00%	78,33%	75,00%
4	78,75%	80,00%	78,75%	90,00%	83,75%	75,00%	81,25%	90,00%
5	83,00%	100,00%	81,00%	90,00%	83,00%	80,00%	85,00%	100,00%
6	82,50%	80,00%	80,83%	80,00%	80,83%	70,00%	82,50%	70,00%
7	82,86%	85,00%	80,71%	80,00%	82,86%	95,00%	79,29%	60,00%
8	85,00%	100,00%	81,88%	90,00%	82,50%	80,00%	78,13%	70,00%
9	86,67%	100,00%	82,78%	90,00%	83,33%	90,00%	78,33%	80,00%
10	87,50%	95,00%	82,00%	75,00%	84,00%	90,00%	78,50%	80,00%
11	88,18%	95,00%	82,73%	90,00%	84,55%	90,00%	79,55%	90,00%
12	88,33%	90,00%	83,75%	95,00%	85,83%	100,00%	80,00%	85,00%
13	88,08%	85,00%	84,23%	90,00%	86,15%	90,00%	81,15%	95,00%
14	86,79%	70,00%	83,21%	70,00%	86,79%	95,00%	81,07%	80,00%



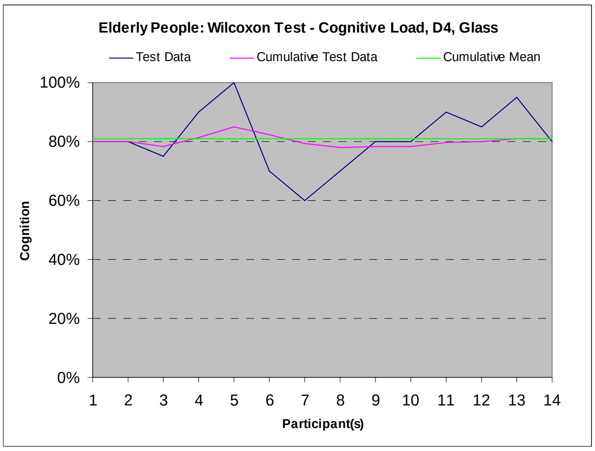
Elderly People: Wilcoxon Test - Cognitive Load, D1, Alu-can



Elderly People: Wilcoxon Test - Cognitive Load, D2, Straw



Elderly People: Wilcoxon Test - Cognitive Load, D3, Plastic



Elderly People: Wilcoxon Test - Cognitive Load, D4, Glass

Appendix K – Usability Measures for Different User Groups

All (Student) Test Participants

Test participants: *TP1-TP35*
Number of male participants: *20*
Number of female participants: *15*
Average age: *24.6 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	65,1%	50,2%	64,8%	73,7%
Accuracy	63,1%	63,3%	81,6%	87,6%
Efficiency	70,1%	35,2%	42,8%	57,1%
Satisfaction	62,1%	52,1%	70,0%	76,4%
Cognitive Load	88,4%	88,9%	83,6%	85,1%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	40,6%	67,4%	68,9%	62,9%
Accuracy	46,2%	89,0%	76,4%	69,8%
Efficiency	35,6%	40,3%	61,4%	45,8%
Satisfaction	40,0%	72,7%	68,9%	73,0%
Cognitive Load	73,7%	80,7%	81,0%	82,3%

Pilot Test

Test participants: *TP1-TP5*
Number of male participants: *3*
Number of female participants: *2*
Average age: *24.8 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	66,3%	48,5%	63,1%	72,0%
Accuracy	63,3%	65,6%	84,4%	90,0%
Efficiency	75,4%	31,0%	39,7%	51,0%
Satisfaction	60,0%	49,0%	65,0%	75,0%
Cognitive Load	85,0%	84,0%	79,0%	78,0%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	37,6%	72,5%	63,4%	69,3%
Accuracy	46,7%	93,3%	78,3%	91,7%
Efficiency	29,2%	37,3%	48,0%	42,4%
Satisfaction	37,0%	87,0%	64,0%	74,0%
Cognitive Load	80,0%	82,0%	86,0%	89,0%

Test 1

Test participants: *TP6-TP20*
Number of male participants: *11*
Number of female participants: *4*
Average age: *25.2 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	62,8%	47,6%	63,7%	73,3%
Accuracy	61,7%	61,9%	83,3%	88,1%
Efficiency	67,1%	34,5%	39,9%	58,2%
Satisfaction	59,7%	46,3%	68,0%	73,7%
Cognitive Load	90,0%	88,0%	85,0%	86,3%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	39,7%	64,0%	67,7%	64,9%
Accuracy	45,6%	88,9%	76,1%	76,0%
Efficiency	38,0%	41,3%	61,4%	47,3%
Satisfaction	35,7%	62,0%	65,7%	71,3%
Cognitive Load	70,7%	81,0%	79,7%	84,0%

Test 2

Test participants: *TP21-TP35*
Number of male participants: *6*
Number of female participants: *9*
Average age: *24.0 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	67,0%	53,5%	66,4%	74,7%
Accuracy	64,4%	64,1%	78,9%	86,3%
Efficiency	71,2%	37,4%	46,8%	58,0%
Satisfaction	65,3%	59,0%	73,7%	79,7%
Cognitive Load	88,0%	91,3%	83,7%	86,3%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	42,4%	69,0%	71,9%	61,7%
Accuracy	46,7%	87,8%	76,1%	65,3%
Efficiency	35,3%	40,5%	65,9%	45,6%
Satisfaction	45,3%	78,7%	73,7%	74,3%
Cognitive Load	74,7%	80,0%	80,7%	78,3%

Erik as Test Monitor

Test participants: *TP1, TP3, TP5-TP20*
Number of male participants: *13*
Number of female participants: *5*
Average age: *25.1 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	63,0%	47,4%	62,0%	72,1%
Accuracy	62,0%	62,3%	82,1%	88,3%
Efficiency	68,8%	33,7%	39,3%	56,4%
Satisfaction	58,3%	46,1%	64,7%	71,7%
Cognitive Load	89,4%	86,7%	82,5%	83,9%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	38,7%	64,9%	66,5%	64,8%
Accuracy	46,3%	89,4%	75,9%	77,1%
Efficiency	36,3%	40,8%	59,0%	46,1%
Satisfaction	33,6%	64,7%	64,4%	71,1%
Cognitive Load	73,1%	80,0%	80,6%	85,6%

Tomasz as Test Monitor

Test participants: *TP2, TP4, TP21-TP35*
Number of male participants: *7*
Number of female participants: *10*
Average age: *24.2 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	67,3%	53,3%	67,7%	75,4%
Accuracy	64,2%	64,4%	81,0%	86,9%
Efficiency	71,4%	36,9%	46,5%	57,8%
Satisfaction	66,2%	58,5%	75,6%	81,5%
Cognitive Load	87,4%	91,2%	84,7%	86,5%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	42,6%	69,9%	71,5%	62,7%
Accuracy	46,1%	88,7%	77,0%	67,4%
Efficiency	34,8%	39,9%	64,0%	45,6%
Satisfaction	46,8%	81,2%	73,5%	75,0%
Cognitive Load	74,4%	81,5%	81,5%	78,8%

All Male (Student) Participants

Test participants: *TP1, TP4-TP5, TP7-TP17, TP21, TP25, TP28, TP31-TP33*
Number of male participants: *20*
Number of female participants: *0*
Average age: *24.7 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	66,8%	49,1%	64,6%	73,8%
Accuracy	63,8%	63,1%	83,1%	88,1%
Efficiency	73,5%	35,4%	42,2%	58,4%
Satisfaction	63,3%	48,8%	68,5%	75,0%
Cognitive Load	89,0%	87,8%	81,5%	84,5%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	40,8%	66,3%	68,3%	71,2%
Accuracy	49,2%	89,6%	75,8%	90,8%
Efficiency	37,2%	40,8%	61,7%	47,9%
Satisfaction	36,0%	68,5%	67,5%	75,0%
Cognitive Load	73,0%	77,5%	80,0%	82,5%

All Female (Student) Participants

Test participants: *TP2-TP3, TP6, TP18-TP20, TP22-TP24, TP26-TP27, TP29-TP30, TP34-TP35*
Number of male participants: *0*
Number of female participants: *15*
Average age: *24.5 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	62,8%	51,8%	65,1%	73,5%
Accuracy	62,2%	63,7%	79,6%	87,0%
Efficiency	65,5%	35,0%	43,6%	55,3%
Satisfaction	60,7%	56,7%	72,0%	78,3%
Cognitive Load	87,7%	90,3%	86,3%	86,0%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	40,3%	68,8%	69,7%	57,2%
Accuracy	42,2%	88,3%	77,2%	58,3%
Efficiency	33,4%	39,7%	61,1%	43,1%
Satisfaction	45,3%	78,3%	70,7%	70,3%
Cognitive Load	74,7%	85,0%	82,3%	82,0%

Erik's Male Participants

Test participants: *TP1, TP5, TP7-TP17*
Number of male participants: *13*
Number of female participants: *0*
Average age: *24.6 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	65,7%	47,5%	63,4%	74,6%
Accuracy	63,5%	62,0%	82,9%	89,7%
Efficiency	72,2%	34,1%	40,3%	59,5%
Satisfaction	61,5%	46,5%	66,9%	74,6%
Cognitive Load	90,8%	85,4%	81,9%	83,8%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	40,5%	64,6%	67,3%	69,7%
Accuracy	50,0%	89,7%	75,6%	89,1%
Efficiency	37,8%	41,4%	60,7%	47,4%
Satisfaction	33,8%	62,7%	65,4%	72,7%
Cognitive Load	73,5%	76,9%	80,0%	84,2%

Erik's Female Participants

Test participants: *TP3, TP6, TP18-TP20*
Number of male participants: *0*
Number of female participants: *5*
Average age: *26.2 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	56,0%	47,0%	58,6%	65,6%
Accuracy	58,3%	63,3%	80,0%	84,4%
Efficiency	59,7%	32,7%	36,7%	48,2%
Satisfaction	50,0%	45,0%	59,0%	64,0%
Cognitive Load	86,0%	90,0%	84,0%	84,0%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	34,0%	65,8%	64,4%	58,5%
Accuracy	36,7%	88,3%	76,7%	66,0%
Efficiency	32,4%	38,9%	54,4%	42,6%
Satisfaction	33,0%	70,0%	62,0%	67,0%
Cognitive Load	72,0%	88,0%	82,0%	89,0%

Tomasz’ Male Participants

Test participants: *TP4, TP21, TP25, TP28, TP31-TP33*
Number of male participants: *7*
Number of female participants: *0*
Average age: *24.9 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	68,8%	51,9%	66,9%	72,4%
Accuracy	64,3%	65,1%	83,3%	84,9%
Efficiency	75,8%	37,9%	45,9%	56,4%
Satisfaction	66,4%	52,9%	71,4%	75,7%
Cognitive Load	85,7%	92,1%	80,7%	85,7%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	41,2%	69,4%	70,3%	74,0%
Accuracy	47,6%	89,3%	76,2%	94,0%
Efficiency	36,0%	39,7%	63,4%	48,8%
Satisfaction	40,0%	79,3%	71,4%	79,3%
Cognitive Load	72,1%	78,6%	80,0%	79,3%

Tomasz’ Female Participants

Test participants: *TP2, TP22-TP24, TP26-TP27, TP29-TP30, TP34-TP35*
Number of male participants: *0*
Number of female participants: *10*
Average age: *23.7 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	66,2%	54,2%	68,3%	77,5%
Accuracy	64,2%	63,9%	79,4%	88,3%
Efficiency	68,4%	36,1%	47,0%	58,8%
Satisfaction	66,0%	62,5%	78,5%	85,5%
Cognitive Load	88,5%	90,5%	87,5%	87,0%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	43,5%	70,3%	72,3%	57,9%
Accuracy	45,0%	88,3%	77,5%	58,4%
Efficiency	34,0%	40,1%	64,5%	43,4%
Satisfaction	51,5%	82,5%	75,0%	72,0%
Cognitive Load	76,0%	83,5%	82,5%	78,5%

Five Males and Five Females from Erik's Participants

Test participants: *TP3, TP6, TP13-TP20*
Number of male participants: *5*
Number of female participants: *5*
Average age: *25.5 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	62,1%	48,7%	63,3%	71,9%
Accuracy	60,8%	61,7%	82,2%	86,1%
Efficiency	67,5%	35,3%	39,3%	56,0%
Satisfaction	58,0%	49,0%	68,5%	73,5%
Cognitive Load	87,5%	89,0%	83,5%	88,0%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	37,8%	65,1%	67,1%	64,4%
Accuracy	43,3%	90,8%	76,7%	75,0%
Efficiency	35,0%	40,0%	59,7%	44,8%
Satisfaction	35,0%	64,5%	65,0%	73,5%
Cognitive Load	73,5%	87,0%	80,5%	87,0%

Five Males and Five Females from Tomasz' Participants

Test participants: *TP25, TP27-TP35*
Number of male participants: *5*
Number of female participants: *5*
Average age: *24.5 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	65,3%	52,0%	64,9%	72,0%
Accuracy	63,3%	64,4%	78,3%	86,7%
Efficiency	69,5%	36,6%	48,8%	55,7%
Satisfaction	63,0%	55,0%	67,5%	73,5%
Cognitive Load	89,0%	91,0%	81,0%	83,0%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	42,9%	65,6%	69,4%	69,8%
Accuracy	50,0%	85,8%	75,8%	90,8%
Efficiency	36,3%	39,6%	65,9%	45,5%
Satisfaction	42,5%	71,5%	66,5%	73,0%
Cognitive Load	78,0%	77,0%	87,0%	80,0%

All Elderly Test Participants

Test participants: *TP36-TP49*
Number of male participants: *5*
Number of female participants: *9*
Average age: *67.1 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	56,1%	50,3%	65,0%	72,6%
Accuracy	56,0%	57,5%	79,4%	78,0%
Efficiency	55,3%	31,1%	37,6%	54,2%
Satisfaction	57,1%	62,1%	78,2%	85,7%
Cognitive Load	86,8%	83,2%	86,8%	81,1%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	40,0%	57,7%	72,1%	45,4%
Accuracy	45,2%	57,8%	75,0%	32,9%
Efficiency	30,4%	33,9%	58,7%	37,7%
Satisfaction	44,3%	81,4%	82,5%	65,7%
Cognitive Load	71,8%	83,2%	74,6%	81,1%

All Elderly Male Participants

Test participants: *TP36, TP38-TP39, TP47-TP48*
Number of male participants: *5*
Number of female participants: *0*
Average age: *68.0 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	61,2%	49,0%	71,5%	76,1%
Accuracy	60,0%	57,8%	86,7%	84,4%
Efficiency	60,6%	31,1%	40,9%	55,9%
Satisfaction	63,0%	58,0%	87,0%	88,0%
Cognitive Load	81,0%	84,0%	85,0%	85,0%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	37,7%	71,3%	69,5%	51,7%
Accuracy	40,0%	86,7%	78,3%	43,6%
Efficiency	27,2%	36,1%	51,2%	39,6%
Satisfaction	46,0%	91,0%	79,0%	72,0%
Cognitive Load	69,0%	82,0%	75,0%	88,0%

All Elderly Female Participants

Test participants: *TP37, TP40-TP46, TP49*
Number of male participants: *0*
Number of female participants: *9*
Average age: *66.7 years*

DRINK	D1, Alu-Can	D2, Straw	D3, Plastic	D4, Glass
Overall benchm.	53,3%	51,0%	61,4%	71,6%
Accuracy	53,7%	57,4%	75,3%	77,2%
Efficiency	52,4%	31,1%	35,7%	53,3%
Satisfaction	53,9%	64,4%	73,3%	84,4%
Cognitive Load	90,0%	82,8%	87,8%	78,9%

POUR	P1, Brik	P2, Prisma	P3, Gable	P4, PET
Overall benchm.	41,2%	53,0%	73,5%	41,9%
Accuracy	48,1%	50,0%	73,1%	26,9%
Efficiency	32,2%	32,7%	62,9%	36,6%
Satisfaction	43,3%	76,1%	84,4%	62,2%
Cognitive Load	73,3%	83,9%	74,4%	77,2%