MASTER THESIS

How do elderly people learn to work with an online system?

An approach based on usability testing, literature and error classification theory

Deborah Oosting

August 2016, Enschede

Faculty of behavioral sciences (BMS)

University of Twente

First Supervisor: Dr. Martin Schmettow

Second Supervisor: Suzanne Vosslamber, MSc

UNIVERSITEIT TWENTE.

Table of content

How do elderly people learn to work with an online system?

Preface

This study was performed as a master thesis for psychology at the University of Twente. The data used was collected during a project at the National Foundation for the Elderly (NFE) where I had an internship and, afterwards on, had the wonderful opportunity to work fulltime for another year. I would really like to thank a number of people, as they have helped me a lot during the long time it took me to finish this master thesis.

First of all, I would like to thank Martin Schmettow (University of Twente) for being my supervisor throughout the entire thesis. Thank you for your valuable insights and for teaching me the true meaning of the active user paradox (It turned out to be quite nice learning new things in SPSS and R). I would also like to thank Suzanne Vosslamber for finding the time to be my second supervisor.

I would also like to thank Gerard van Loon and Nina van der Vaart (NFE) for their great help during my internship and for giving me the amazing possibility to present some of my results in one of the annual project meetings in Bucharest. I would like to thank the rest of the Care@Home consortium as well, as they provided me with a great opportunity to work in the field while they treated me with respect despite being an intern. I enjoyed working with all of you!

A big thank you to Ruud Zandbergen for his help throughout the thesis as well. Thank you for exchanging ideas and discussing about both of our theses, it was very valuable for me.

I simply can't forget to thank De Koperhorst for their help with the testing location too, as well as all the participants. I had the most wonderful time working with them and (often) listening to the wide range of stories they told as well while having coffee and cookies in between testing rounds. I could not have done this thesis without them.

There are a bunch of other people I would like to thank as well. Thank you to Sanne, Tom, Stephan, Menno, and Michou, for listening, advising, helping, proofreading and even sending motivational chocolate (yes, that exists!). I would like to thank my family and boyfriend's family for all of the above and for simply being there for me all the time throughout this project. Last, I would specially like to thank my wonderful boyfriend Lennert. I could not have done this without your everlasting support.

How do elderly people learn to work with an online system?

Abstract (English)

With a rising elderly population, there is an increasing interest in the development of technology for elderly people which will help them in their everyday lives. However, these technologies are often not adapted to them properly. As the learning processes of this specific user group is still not well-understood, this longitudinal study researched how elderly people learn over time, which types of mistakes they encountered, which of those were the most and least severe and the role of previous experience in learning. Twenty healthy elderly aged 66 to 92 years old participated in a three-trial study in which the Care@Home system was tested by means of 9 tasks. Data was collected by video recording, usage of a concurrent think-aloud protocol and questionnaires. This study wanted to use the results of the Care@Home project for research on elderly learning. It was chosen to perform a longitudinal study to be able to trace learning curves over time as well as the occurrence of different types of mistakes over time. After incident matching, usability problems were listed in order of severity. Severity was rated by a set of rules based on frequency, impact and persistence. To be able to tell something about which types of problems became more and less severe over time, problems were classified into different subtypes, each with another background based on dimensions of conscious control and time in the process. An error classification guideline was constructed to do so, dividing all usability problems into 8 different categories. Previous experience was taken into account as well as a possible influence on learning rate and number of mistakes. It was found that most , but not all elderly people became faster over time working with the care@Home system and thus, were able to learn, albeit in very different rates. Age had an effect on time on task, with older participants being slower than younger ones. Problems found the most often were thought problems and sensorimotor problems. The most severe problems were often based on ambiguous menu's, inconsistencies, design flaws and the use of the remote control. Previous experience in hours was found to have a positive effect on the number of problems in general but especially on knowledge problems, while previous experience in components was found to have a positive effect on time on task. While this study showed that previous experience, thus practice, can improve elderly learning by at least shortening time on task and the number of knowledge problems, this study did not find factors influencing rules and skills problems. Follow-up studies might want to try and find such factors in order to reach a fuller understanding of elderly learning.

How do elderly people learn to work with an online system?

Abstract (Dutch)

Nu de hoeveelheid ouderen in de populatie toeneemt ontstaat er steeds meer interesse in technologie die hun hulp bieden in hun alledaagse leven. Toch is deze technologie vaak niet goed ontworpen voor deze specifieke doelgroep omdat er onvoldoende bekend is over hoe zij leren. In deze studie is onderzocht hoe ouderen leren, welke typen fouten zij tegenkomen en welke van deze fouten het meest en minst ernstig waren. Twintig gezonde ouderen van 66 tot 92 jaar namen deel aan deze studie waarin het Care@Home systeem in 9 taken werd getest over 3 rondes. Data werd verzameld door video-opnamen, *concurrent think-aloud* methode en vragenlijsten. Deze studie wilde de resultaten van het Care@Home project gebruiken voor onderzoek naar hoe ouderen leren. Er is voor een longitudinale studie gekozen omdat dit het mogelijk maakt om leercurven over tijd te bekijken, evenals hoe en wanneer verschillende typen fouten voorkomen. Alle gevonden usability problemen zijn op volgorde van ernst gezet, waarbij de ernst van problemen werd beoordeeld aan de hand van regels voor de score, gebaseerd op frequentie, impact en hardnekkigheid. Om iets te kunnen vertellen over welke typen problemen meer of minder ernstig worden in de loop der tijd werden de problemen geclassificeerd in verschillende subtypes, gebaseerd op dimensies van bewuste uitvoer en de tijd waarop een probleem plaatsvond. Hiervoor is er in deze studie een foutenclassificatie handleiding ontwikkeld waarmee alle problemen in 8 categorieën werden onderverdeeld. Ook voorgaande ervaring is meegenomen als mogelijke invloed op leercurven en aantal fouten. Bijna alle ouderen werden in de loop van tijd sneller in het werken met het Care@Home systeem en waren dus in staat om er mee te leren werken, hoewel de mate waarin wel erg verschilde onderling. Leeftijd had een effect op de tijd die men nodig had om taken af te ronden, waarbij oudere deelnemers langzamer waren dan jongere deelnemers. De gevonden usability problemen waren voornamelijk ‘thought’ en ‘sensorimotor’ problemen. De ernstigste problemen kwamen vaak door ambigue menu’s, inconsistentie, designfouten en door de afstandsbediening. Voorgaande ervaring in uren bleek een positief effect te hebben op het aantal problemen, vooral ‘knowledge’ problemen, terwijl ervaring in onderdelen een positief effect had op de tijdsduur voor een taak. Hoewel ouderen door voorgaande ervaring – dus door oefening, beter kunnen worden in bepaalde aspecten vond deze studie geen factoren die van invloed waren op ‘rules’ en ‘skills’ problemen. Toekomstige studies zouden hier mee verder kunnen gaan om zo de kennis over hoe ouderen leren te vergroten.

How do elderly people learn to work with an online system?

1.0 Introduction

In the last few years , there has been an increase in attention for the elderly population, for a multitude of reasons that initially don't seem to have much to do with the process of learning. The world population is ageing rapidly. According to the World Health Organization (WHO), the proportion of people who will be older than 60 will double from 11% to 22% between 2000 and 2050 (World Health Organization, 2014). In the year 2100, it is expected that a minimum of 30% of all people worldwide will be aged 60 and over (Lutz, Sanderson, & Scherbov, 2008). An ageing population has its consequences, the first being loneliness. The number of (Dutch) people who feel lonely increases steadily with age, from 40.9% of people between the age of 65 and 74 saying that they are lonely to 49.5% of elderly between the age 75 and 84. Elderly people who are divorced or become widow(er)s are even more prone to feelings of severe loneliness (Zantinge, 2014). Other events contributing to increased loneliness are retirement, loss of family members and friends and a decrease in mobility (Centraal Bureau voor de Statistiek, 2012). Loneliness is detrimental to health; It is a risk-factor for developing depression (Pinquart & Sörensen, 2001). People who are lonely are also found to have poorer quality of sleep and an increase in blood pressure, even when their health behaviours and self-reported health does not differ from people who are not lonely (Cacioppo et al., 2002). Research has also shown that for elderly over the age of 60, loneliness is a fair predictor of functional decline in daily activities, daily walking, movements using the upper arm and stair walking. It was a significant predictor of an early death as well (Stijacic Cenzer, 2012).

As a second consequence of an ageing population, predictions show that there will be a shortage of residencies in retirement homes, while at the same time, there will also be a shortage of staff to take care of the rising number of elderly people (Broadbent et al., 2010). In the Netherlands as well, there will be a rise in demand at one side and the shortage of healthcare supplies and services on the other. However, Dutch elderly also have the wish to live independently in their own homes; The Netherlands Institute for Social Research (SPB) interviewed elderly people of whom most indicated that they had the wish to live in their own house for as long as possible (Campen, 2011).

A multitude of projects have already been set up to make it possible for elderly people to live independently for a longer period of time while at the same time targeting loneliness.

One of them is the Active Assisted Living programme (AAL, formerly known as Ambient Assisted Living, also see www.aal-europe.eu), a European funding program aiming to create a better condition of life for the older adults, but also to strengthen the industrial opportunities in Europe through the use of information and communication technology (ICT). Care@Home is one of the projects within the AAL-programme, which aims to improve wellness, social care services and social support for the elderly by means of a personalized communication and service channel on an interactive multimedia Smart TV in their home. One of the Dutch stakeholders in this project is Nationaal Ouderenfonds (National Foundation for the Eldery, or NFE), whose major goal in the project is to represent the needs and wishes of the elderly population in order to create a product that is really suited to the target group. They do so by performing usability tests together with the elderly to improve the product each step along the way of development.

While these interventions and projects are mostly set up with the best intentions for the elderly, two important things have to be taken into account, both regarding learnability. The first is an aspect of motivation; Do elderly people want to work with such systems at all? What is their motivation for accepting or declining? Second, the question of learnability in itself: Can elderly people actually learn how to work with such systems at all? If so, how do they learn this and does this differ from the way younger people learn how to work with online systems? As noted earlier, the process of learning consists of many facets. In order to find out if and how elderly people learn to work with such systems, it is needed to know more about elderly learning in general.

This paper will first describe a longitudinal usability test of the Care@Home project including the participants, Care@Home system, tests and questionnaires. After that, incident matching, error classification and finally the usability problems found based on the incidents will be described. Following this are the results, consisting of an outline on whether the elderly participants did or did not learn how to work with the Care@Home system, together with a ranking of which type of usability problems are the least and most severe over time. The usability problems will be ranked in a descriptive way as well as being classified according to error classification theory as well to find out whether one type of problem is more or less severe than other types for elderly. Then, an outline will be given on the effect of previous experience on the least and most severe problems and their types based on classification

theory. The most and least severe problems will then be linked to the literature on elderly learning to find out where it coincides and where it does not – and where more information on elderly learning perhaps is needed. But first of all, before describing the usability study that was performed, it is needed to understand the aspects of elderly learning and error classification theory. The next sections 1.2 and 1.3 will give more information about both of these subjects.

1.2 Aspects of learning

So, what is learning? First of all, there are different viewpoints based on whether you are an observer or a learner. According to Washburne (1936), "*learning is an increase, through experience, of ability to gain goals in spite of obstacles*". Here, either the goals become more complex, or the level of effort needed to attain goals will become less over time. Learning entails many factors next to goals and obstacles, such as the level of effort taken to learn and help received.

While this definition gives insight in what learning in itself entails, it does not explain the process of learning. The process of learning is broad. First of all, one factor that enhances learning is motivation to do so, such as positive memories (Maehr & Meyer, 1997; Washburne, 1936). If people are not motivated to learn something, this may be inhibiting their progress (Caplin, 1969). Attention is another important factor, as it creates a focus on stimuli that are important for what needs to be learned, while at the same time channelling out irrelevant stimuli (Chun, Golomb, & Turk-Browne, 2011). Information is then encoded and stored in memory, where it is processed, categorized and clustered. It is linked to things people already know and understand. For learning to happen, diverse facts, ideas and concepts must also be connected to other facts, ideas and concepts in order to create a bigger picture of how everything related to one another. This is needed in order to be able to retrieve the information later, when it is needed in another situation, or when it needs to be linked to new information again (Austin, Orcutt, & Rosso, 2001). So, how do these processes of learning work exactly?

1.2.1 Motivation to learn

Regardless of a learner's age, motivation is a key aspect of learning. Motivation helps learning in different ways. It activates someone by directing behavior towards goals (Maehr & Meyer, 1997). It also helps initiation; Someone who is more motivated to learn something

will start more eager than someone who is less motivated (Larson, 2000). After starting a learning process, motivation also determines persistence, as people who are highly motivated to learn something will maintain or update their goals over time, and spend more time trying to master what they want to learn, for example by taking more courses to master their goal (Maehr & Meyer, 1997). A third aspect of motivation is intensity. Intensity can be seen through the vigor that goes into pursuing a goal, for example by regularly taking opportunities to practice the learned material outside of classes (e.g., someone who tries to learn Russian chooses to watch Russian TV programs in their spare time). This, in turn, influences performance, working as an enhancement for learning and mastering of the desired skill (Pugh & Bergin, 2006).

Motivation can be divided in two different kinds, intrinsic and extrinsic motivation. The first kind of motivation is the one where the learner does something because it is truly interesting and enjoyable to him or her (e.g. reading French literature in your spare time because you really want to learn it). The motivation comes from within the learner, and not from an external reward. The latter is the kind of motivation someone has when it leads to a specific outcome (e.g. learning French grammar rules for an exam in high school because you want to get a good grade) and is mostly associated with some kind of a reward (Ormrod, 2013; Ryan & Deci, 2000). Motivation seems to be the most beneficial to learning when it is intrinsic, though research has shown that learners are motivated by both intrinsic and extrinsic motivation simultaneously (Eisenberger, Rhoades, & Cameron, 1999; Lin, McKeachie, & Kim, 2001; Ormrod, 2013).

1.2.2 Attention and learning

Next to motivation, attention is important for learning. As learning is the process of memorization, integration and application of new information, attention gives the learner an initial focus on whatever has to be learned (Gottlieb, 2012). This is needed because the environment in general presents more information than someone can process at once (Chun et al., 2011). There are two types of attention, external and internal. The first type related to selecting information that comes in through the senses. This is also called selective attention, and it helps the learner not to get distracted by irrelevant sensory stimuli (Chun et al., 2011; Kruschke, 2011; Sauce, Wass, Smith, Kwan, & Matzel, 2014). The second, internal attention, related to the selection of internally related material, such as what is in the working memory,

or even long-term memory or response selection (Chun et al., 2011). This form of attention protects the learner thus from distractions caused by emotional impulses, irrelevant memories and automatic responses (Sauce et al., 2014). For learning, external attention selects the sensory information that will continue for processing in the working memory (needed because working memory has limited capacity), while internal attention includes the cognitive control or executive mechanisms that give priority to which sensory information continues to working memory for encoding. At the same time it suppresses possible distractions as well (Chun et al., 2011). When the flow of information continues, learning takes place – involving memory.

1.2.3 The role of memory in learning

Memory and learning are related, but different. While learning means the acquisition of a skill or knowledge itself, memory is the expression of what you have learned. The two also differ based on speed. Acquiring a new skill or knowledge mostly costs quite some work and goes relatively slow, this is learning. If acquisition goes instantly, this is making a memory (Kazdin, 2000). So, how does memory enhance learning? First of all, there are different types of memory. One type is short-term memory, the place where information enters the brain in order to be processed. Here, it is stored for a few seconds, after which the information enters the working memory where it is processed further (Ricker, 2012). This is also where attention plays a big role, as attention makes it possible for information to be processed further (Chun et al., 2011; Gottlieb, 2012). Information that enters working memory is scanned. If it is deemed important enough, it is actively processed so it won't be forgotten. Active processing can, for example, consist of rehearsing. Only when information is actively used in the working memory, it can be transferred to long-term memory (Baddeley & Hitch, 1974, 2010). In the long-term memory, information is classified, organized, connected to other, already known concepts and information and finally, stored permanently (Ricker, 2012). From here, information can then be retrieved in order to be used again. The process of transferring new information in the long-term memory takes time and costs effort. When this transfer is complete and someone is able to retrieve the information as well, learning took place (Kazdin, 2000).

Learning can be seen in the brain as well, a research on sea slugs (Aplysia) showed earlier. Basically, the principle of learning and memory is that it involves changes in the strength of synaptic connections between neurons. There is no reorganization of the nervous

system, or in the growth of new neurons – what changes is the strength of a connection that was already there before. Involved in short-term memory are underlying biochemical changes that are transient – therefore, the memories won't last. Long-term memories involve less transient biochemical changes and often, changes in the structure of neurons as well. This can include growth of new processes, synapses and branches (Byrne, 2015). Summarizing, the brain changes the strength of existing connections between neurons using non-transient mechanisms in order to transfer information into long-term memory. This, learning, makes it possible to retrieve information again at another time in another situation.

1.2.4 Motivation and attention in elderly learning

As explained above, motivation, attention and memory are important aspects of learning. So, how are these aspects related specifically to elderly learning then? Related to motivation, research showed that elderly really *want* to learn, especially about how to work with a computer (Purdie & Boulton-Lewis, 2003). Moreover, their motivation seems to be mainly of an intrinsic kind: In general, online elderly see the online world as a source of information, for personal, day-to-day contact with family and friends. They are inclined to use the internet to stay up-to-date, however, not to meet new people; rather they just want to stay in touch with people they already know (Martinez, Cabecinhas, Loscertales, Loscertales Abril, & Martínez Pecino, 2011). However, while a multitude of elderly people are online, a lot of technologies – even those currently aimed at this particular user group – are not properly designed or even downright user-unfriendly for the elderly user. (Wisniewski & Polak-sopinska, 2009). While the elderly are willing to use new, online technologies, the designs are not yet adapted to their needs. This may be detrimental for motivation, as elderly people might think it is too difficult for them to work with such technologies.

Research about attention in elderly people has shown that overall, there seems to a decline in attention span with age (Hawthorn, 2000; Purdie & Boulton-Lewis, 2003). Compared to younger people, elderly have severe problems when it comes to inhibiting irrelevant information, while at the same time maintaining relevant information (Jones & Bayen, 1998). In a test based on learning rules, simultaneously testing attention span, it was found that elderly had a lower attention span, leading to lower accuracy rates in performance (Bauer, Toepper, Gebhardt, Gallhofer, & Sammer, 2015). This difficulty in inhibition was also shown in an experiment by Rodrigues and Pandeirada (2015), who let elderly perform

attention tasks in either an environment without any distractions, or in an environment with distractions as they occur in everyday life. It was found that in the second condition, elderly were not able to filter out the irrelevant stimuli and were consequentially less accurate in their performance. Overall, sustained attention (focus on one thing) seems to decline. For divided attention (focus on multiple things at the same time), conclusions are not solely based on a decline, but on another factor as well: As an elderly person needs to pay attention to more things at the same time, task complexity increases, needing more cognitive resources. This in turn may account for a slower performance and less accuracy (Hawthorn, 2000). Even though elderly experience a decrease in attention span, information definitely comes through – even if it takes longer (Purdie & Boulton-Lewis, 2003; White et al., 1999).

1.2.5 Memory in elderly learning

When information flow continues, memory has a great part in learning for elderly as well. A lot of research has been performed about the different parts of memory and how they affect elderly learning, especially compared to younger people's learning abilities. Overall, it seems older age is associated with a worse performance related to time and accuracy in general, possibly caused by a decrease in overall myelin and number of active synapses (Salthouse, 2000). Possibly, retrieval or building of memories also suffers from this, making learning harder. This already happens in the storage capacity and the manipulation of information in working memory, which declines in older age. (Balota, Dolan, & Duchek, 2000). Even so, this is not the case for all types of memory: According to multiple studies, the effect of priming (or implicit learning) does not diminish over time, and if it does it's mostly linked to neurological illness such as Alzheimer's disease (Fleischman, 2007; Schugens, Daum, & Spindler, 1997; Spaan & Raaijmakers, 2011). Results on skill learning (also implicit) are not uniform. One study found a non-significant difference related to a word-stem completion task where elderly performed worse than younger people (Schugens et al., 1997). Another study claimed that implicit memory, and thus implicit learning, was not measured correctly before, influencing findings. Their study found a decline of implicit learning with older age (Ward, Berry, & Shanks, 2013).

Measures of explicit learning (conscious long-term memory and learning) show that episodic memory, which is the memory about specific personal events, does decline over time eventually (Baddeley, 2001). In normal aging this does not happen before the age of 60 and

semantic memory (knowledge about the world, facts and concepts), even keeps on improving until approximately the same age (Aine et al., 2011; Rönnlund, Nyberg, & Bäckman, 2005). Related to learning, this would mean that elderly are still able to grasp new concepts and link these to already available knowledge, declining to some degree after the age of 60. Another study showed this to be true, finding that while semantic encoding was relatively the same for elder (60 years and older) and younger participants, younger participants outperformed the elderly regarding episodic encoding. Their conclusion was that encoding processes may suffer more over time than retrieval processes (Friedman, Nessler, & Johnson, 2007; McDonough, Cervantes, Gray, & Gallo, 2014).

While overall, memory does seem to decline over time, some types are more robust than others. For example in encoding information, elderly seem to encode information less specific than younger people, making retrieval later on more difficult (Balota et al., 2000). This was shown in an fMRI study in which young and older participants had to recollect details of pictures showing complex scenes. Not only were elderly less accurate in their recollection, they also remembered less details than the younger participants (McDonough et al., 2014). Next to this, older age comes with a shortening of mental resources. When the number of mental operations rises, or a task becomes more complex, elderly performance lowers. In an experiment where the backwards digit span test was done, elderly performed not worse than younger participants on merely one task (Meadmore, Dror, & Bucks, 2009). However, when a second task was added, younger participants outperformed the older ones by far (Craik, 1994; Van der Linden, Bredart, & Beerten, 1994). Another experiment confirmed this finding, showing that this effect did not depend as much on the type of information processed but more on how much resources it takes to perform a task (Cansino et al., 2013). As elderly people have less processing resources available, learning for them works better on a base of recognition than on recall (Jones & Bayen, 1998). A learning experiment was performed based on these findings as well. In this experiment, younger and older people were exposed to letter strings generated by artificial grammar. At the same time, they had to perform another task (so, they were not aware of learning grammar implicitly). Afterwards, they had to retype every string after they had seen it for 5 seconds (explicit learning). Then, they had to classify novel strings as either grammar or non-grammar (implicit learning), based on their intuition (or, implicit learning). Elderly people were worse than younger people on retyping the strings, but no difference in classifying strings were found

between elderly and younger participants (Kürten, De Vries, Kowal, Zwitserlood, & Flöel, 2012).

Another form of learning where the differences between older and younger participants is quite visible, is spatial learning. This type of learning is important for learning routes or landmarks. Over time, route learning itself seems to have a minor decline – elderly making more mistakes and showing a decline in route efficiency compared to younger people. However, younger people are better at map learning and place learning (Klencklen, Després, Dufour, & Despres, 2012). Holzinger, Searle and Nischelwitzer (2007) found that elderly need more time than younger people to learn navigation processes, also on a mobile phone. They also had a higher error rate and less knowledge retention of the interface over time. Therefore consistency in a system and its interface is very important to enhance elderly learning and inconsistencies – making problems worse – should be avoided (Holzinger et al., 2007; Van Veldhoven et al., 2008).

Associative learning as well is affected by age. In a test, participants had to learn rules based on feedback and the reversion of this feedback. If the rules changed, participants had to change their focus on other information. It was found that elderly showed lower accuracy rates in this task than younger people did while (re)learning the rules. However, it was thought that associative learning in this context also included attention, as participants had to shift their attention to something else when the rules changed. As attention was also found to be lower in elderly participants than in younger, this offers a partial explanation as well (Bauer et al., 2015). Associative learning also has great part in keystroke actions and (mouse) button clicking, as people have to learn the association between keys or buttons and what they do when stroking or clicking it (Chou, Lai, & Liu, 2013). Basically, this is paired-association learning. An experiment in which participants, either novices or advanced computer users, learned how to use Word, it was found that for elderly novices, learning how to work with the mouse and learning how to open menu's using key commands was more troublesome. They were slower than younger and middle-aged novice participants to complete the learning process (it was self-paced) and even after completing the learning tasks, elderly novices performed worse than the younger and middle-aged. For advanced computer users, effects of age were also found on learning and performance, but the correlation was far lower than for the novices. For the final measures, an age effect was found for time but again, with a lower

correlation than in the novices group. For accuracy, no age effect was found in the final measure. These results show that associative learning based on keystrokes, mouse clicks and button clicks is overall worse for elderly than for younger people. Still, it also shows that elderly people can overcome some decline in learning merely by experience (Charness, Kelley, Bosman, & Mottram, 2001).

Last, motor learning experiences some influences of age as well. Overall, studies agree that learning motor skills is harder for elderly people than for younger people (Salat et al., 2004; Voelcker-Rehage, 2008). For learning fine motor skills, results seem to be task specific. Learning small movements (such as moving a lever to a target place) costs elderly people more effort and time than younger people. Learning differences between elderly and younger become increasingly visible over the course of practicing, not right from the start. Elderly are well capable of learning new motor sequences, but they are slower in learning than younger people, especially when task complexity increases (Voelcker-Rehage, 2008). They need more practice than younger people to get to the same level of skill (Cai, Chan, Yan, & Peng, 2014; Ketcham & Stelmach, 2004). One explanation for slower responses – playing a role in pace of learning, is the lessening of dopamine over the years (Seidler et al., 2010; van Dyck et al., 2008). These results in motor learning show that it is possible that elderly might experience difficulties in working with technical devices that also include motor learning of some kind.

1.2.6 Other aspects important to elderly learning

Next to all these changes in learning over time, there are also some other aspects to take into account considering elderly learning that do not relate to specific forms of learning. First of all, elderly people seem to learn more from positive feedback than from negative compared to younger people. In a learning experiment, younger people outperformed the elderly except in a positive learning condition (Eppinger, Herbert, & Kray, 2010). Implicating that older people learn more from positive than from negative feedback, this is an important note to take into account when building a system.

Second, level of previous experience has an impact on learning. As was seen in the experiment by Charness, Kelley, Bosman and Mottram (2001), differences in learning were visible when comparing advanced computer users to novices, giving advanced users certain advantages. Even though their experiment merely found effects of experience on accuracy and not on time, other experiments did found such effects, also in learning how to work with

online systems (Czaja & Sharit, 1993; Hurtienne, Horn, Langdon, & Clarkson, 2013; Laberge & Scialfa, 2005). Advanced users may also make different types of mistakes than novice users while learning. Advanced users, or experts make more mistakes based on habits; they know how something needs to be done, but forget one aspect. Or in a likewise system, something works just slightly different than they are used to. Novice users, on the other hand, make more diverse mistakes because of their lack of knowledge and experience (Carroll & Rosson, 1987). Learning is always accompanied by making mistakes. These mistakes, or errors, can be divided into different types, each representing other underlying causes. The next paragraph will handle errors and their classification.

1.3 Error classification theory

As explained above, learning for elderly people is different than for younger people. Inevitably included in learning is making mistakes or encountering errors. If learning is different for elderly, it might be that the mistakes they encounter are different as well. In order to find out, this study uses error classification theory. Error classification theory can be used to classify user mistakes, or usability problems, to certain types of errors that each have a different background as to why they occur. Comparing which type of errors occur the most and least gives insight in elderly learning, especially when the number of errors of different types are compared over the course of multiple trials, such as was done in this study. As there are multiple ways of error classification, the theory behind it and some examples of classification models will be described below, after which the error classification model used in this study will be described in full detail.

1.3.1 Skills, rules, knowledge

Some studies regarding usability problem classification are based on the skills, rules, knowledge model as proposed by Rasmussen (1983). This model, created to predict human performance, divides behavior into three levels based on the amount of conscious control. First, there is the skill-based behavior, actions for which no conscious control is needed as they rely on automaticity. The rule-based level is based on more conscious control, as a person will use previous experience and certain rules to do something. Last, the knowledge-based level needs conscious control and a lot of attention; Hardly any previous experience is used here. This level is mostly used when people are unfamiliar with a task or action. As learning takes place, people will first approach new situations on the knowledge level. When

a situation becomes more familiar, people will be more on the rule-based level, and when situations are fully known, people will process it on the skill-based level. It is therefore expected that when people learn to work with a new system, mistakes will at first mostly be on the knowledge level and shift to more mistakes being made on the rule-based and skill-based level after more practice. It is shown that when people become more familiar with doing something, the number of mistakes they make overall lowers (Kjeldskov, Skov, & Stage, 2005). Taking this into account, it might be that especially knowledge- and rule-based problems lessen to some extent over time while the number of skill-based problems rises. If so, a problem classification based on the skills, rules, knowledge model might give more insight in which problems will stay and which will disappear over time.

1.3.2 Slips and lapses

Adding another division of usability problems to the skills, rules, knowledge model, Reason (1990) argued that problems can be divided into slips and lapses, and mistakes. Slips and lapses are execution failures based on automatic processes (comparable to knowledge level from Rasmussen (1983)), with the difference that a slip concerns a situation in which the execution was incorrect, while a lapse concerns a situation with no execution at all. The other level, mistakes, are planning failures, coming from higher-order cognitive processes, which can be divided into rule-based- and knowledge based mistakes, comparable to the rules and knowledge levels as described by Rasmussen (1983)

1.3.3 Goal-orientation and level of action

Zapf, Brodbeck, Frese, peters and Prümper (1992) observed usability problems that were found for office workers using computers to complete certain tasks. Hereafter, they created different classes in order to classify these problems, in multiple dimensions as seen below in figure 1. The first dimension is based on the assumption that actions are goal-oriented. Plans have to be developed (goals/planning), monitored/executed (monitoring) and then evaluated (feedback). The second dimension shows three levels of action which can be compared to Rasmussen (1983). The intellectual level of regulation is similar to the knowledge-based level, the level of flexible action patterns to Rasmussen's rule-based level and the sensorimotor level of regulation, or automaticity is similar to the skill-based level by Rasmussen. Zapf also included a knowledge base which can't be compared to Rasmussen's model nor the one by Reason (1990), but is used for developing goals and plans (Zapf et al.,

1992). So, in some way, this model shares a resemblance with the work of Reason (1990) when compared to the first dimension (Zapf et al added feedback as a third goal-oriented stage) and also with the model of Rasmussen (1983) if compared to the second dimension. Classification by the Zapf et al. model (1992) gives more insight in when usability problems occur in the action process. Moreover, it also shows whether these problems occur by use of (full or partial) conscious control or by automaticity.

Knowledge base for regulation	Knowledge problems		
Level of action regulation	**Steps in the action process**		
	Goals/planning	**Monitoring**	**Feedback**
Intellectual level of regulation	Thought problems	Memory problems	Judgment problems
Level of flexible action patterns	Habit problems	Omission problems	Recognition problems
Sensorimotor level of regulation	Sensorimotor problems		

Figure 1: *error classification*

1.3.3.1 Types of errors

Knowledge error: At the base of being able to do certain tasks at all is knowledge of the device one has to perform the task on. So, for example, someone who has no idea how to work with a computer will most certainly make mistakes based on commands that someone who has experience with working on a computer will probably not make (Zapf et al., 1992). So, the basics of the device, system, program or platform the user will work with have to be explained to the user beforehand. If a mistake was made based on a lack of understanding beyond the parts that needed to be tested (so, the basics), it is probably a knowledge error (Barendregt, Bekker, Bouwhuis, & Baauw, 2006). As the literature showed that elderly are slower learning things, mistakes on this level are to be expected (Balota et al., 2000; Salthouse, 2000)

Thought error: This type of error occurs when a user develops an inadequate goal or plan to do a task, even though the user knows all the features of the system (Barendregt et al., 2006;

Zapf et al., 1992). So, for example, when a user wants to go to another page but the navigation buttons all look the same, so s/he clicks the wrong one and ends up on another page than intended. Mistakes of this kind can occur in all kind of ways, related to either insufficient attention as well as not being able to tell which button does what (Bauer et al., 2015; Chou et al., 2013)

Memory errors: Occurring during the monitoring phase, this type of error occurs when the user had the right plan or goal, but a part of the plan was forgotten during execution (Barendregt et al., 2006; Zapf et al., 1992). Just like above, less detailed encoding might prove some difficulties here. At the same time, when a task is complex, the elderly user might forget some things due to limited resources available (Cansino et al., 2013; Friedman et al., 2007; McDonough et al., 2014)

Judgment errors: Happens when the user cannot interpret the feedback gotten from the system after an action was performed (Barendregt et al., 2006; Zapf et al., 1992). Barendregt et al. (2006) use an example in which a child has played a game correctly, but does not understand the feedback afterwards, and thus, does not know whether s/he performed right or wrong. This might interact with a multitude of resources being drawn at the same time, but also by less attention, or even feedback with a too negative connotation (Cansino et al., 2013; Eppinger et al., 2010; Hawthorn, 2000; McDonough et al., 2014). If a lot is going on at the same time on the screen, it might even be the case that the user is not able to inhibit the irrelevant details and fails to notice the feedback properly altogether (Rodrigues & Pandeirada, 2015).

Habit errors: This type of error takes place during the goal and planning phase. They occur when a user performs a correct action in a wrong setting. Most of the time, this is an action that worked in an earlier (different) situation, which makes the user thinks it will work in this setting as well (Barendregt et al., 2006; Zapf et al., 1992). Results related to skill learning were ambiguous, but mostly in the direction of not too much of a decline over the years (Schugens et al., 1997; Ward et al., 2013). Because of this, some, but not much of this type of errors are to be expected. It is found though, that more experienced users might make mistakes based on habits more often (Carroll & Rosson, 1987).

Omission errors: An omission error occurs when a person does not complete a (sub) plan that s/he usually knows how to do (Barendregt et al., 2006; Zapf et al., 1992). This might happen because the user has too much focus on the next step (Barendregt et al., 2006). Zapf et al. (1992) give the example of a person who usually always saves his or her files but for once, forgets to do so. This might occur because a task is complex and draws on many memory related resources at once, but also because less details are remembered and thus, parts are easily forgotten for once (Cansino et al., 2013; Meadmore et al., 2009; Van der Linden et al., 1994).

Recognition errors: Occurring during the feedback phase, this type of error happens when a person fails to notice or recognize a feedback message that would normally be well-understood, leaving the participant confused (Barendregt et al., 2006; Zapf et al., 1992). It is important to note that the difference between recognition errors and feedback errors is that with the feedback errors, the user received new feedback while with recognition errors, the user receives feedback that was given (and possibly understood) before. As a lot of things going on simultaneously might draw on a lot of resources, it might be that feedback is misinterpreted (Friedman et al., 2007; Laberge & Scialfa, 2005; McDonough et al., 2014; Rodrigues & Pandeirada, 2015).

Sensorimotor errors: As it is empirically very difficult to differentiate between the three levels of action for this category, there is only one type of error here. Errors here are related to motor-skills, such as touching the wrong button because the buttons are placed too close to each other, or because they are too small (Barendregt et al., 2006; Zapf et al., 1992). On the other hand, errors can also occur because elderly users experience difficulties in learning the connection between a keystroke or mouse click and what happens (Charness et al., 2001; Chou et al., 2013). Together with slower motor execution altogether, these mistakes are to be expected when working with a system that uses a mouse pad, remote control or in general small buttons (Seidler et al., 2010; van Dyck et al., 2008).

1.4 Study goal and research questions

As described in detail above, it is possible to describe the process of elderly learning in multiple ways. One way shows how their learning differs from how younger people learn and how this related to the (memory) processes that entail learning, another is by means of classification theory of the errors they encounter while learning to work with a new system.

How do elderly people learn to work with an online system?

Based on these two ways, this paper will describe whether elderly people are able to learn how to work with an online system, the types of mistakes they experience while doing so, and which mistakes they are and are not able to recover from easily. Next to that, this paper will compare whether the error classification findings coincide with elderly learning processes as described in the literature and if not – where these differences are in. This study will be of a longitudinal kind, which provides great possibility to see which mistakes will and will not disappear over time, how they are related to both theory of error classification and elderly learning processes as described in the literature. Therefore, research questions will be the following:

R1: Can elderly people learn how to work with an online system?

R1a: What do the learning curves of elderly people, learning how to work with an online system, look like?

R1b: Are there elderly people who do not learn how to work with an online system?

R2: Which mistakes do elderly people make while learning to work with an online system?

R2a: What are the most and least severe mistakes elderly make while learning to work with an online system and how are these problems classified according to error classification theory?

R2b: Do the most and least severe mistakes of elderly people differ based on previous online experience, and if so – are these differences also found in problem types after error classification?

R2b: Do these worst and least severe types of mistakes coincide with the literature on elderly learning processes?

How do elderly people learn to work with an online system?

2. Methods

2.1 Sample

Together with De Koperhorst, a home for the elderly based in Amersfoort (NL), NFE recruited 20 participants to participate in this study on a voluntary base. Eight of these participants (40%) lived in a retirement home while the 12 others (60%) lived on their own. 13 (65%) were female and 7 (35%) were male. Their mean age was 79,5 years old (*SD* = 7,506) ranging from the youngest participants being 66 years old to the oldest who was 92. All 20 participants had the Dutch nationality.

To control for previous experience with technology and level of expertise, participants were asked to fill in questionnaires regarding previous experience with NFE and their previous experience with technical (online/smart) systems. Of all participants, 4 (20%) had been involved in previous projects from NFE. When asked, two of these projects were very different from Care@Home. Two of the participants worked on a project that lead to Care@Home and were thus familiar with the concept of a smart TV. However, working on that project mostly consisted of doing usability interviews, so they could be included in the Care@Home trial nonetheless.

2.2 Materials

2.2.1Care@Home platform

A mock-up version of the Care@Home platform was used in this study. This platform was created by Nationaal Ouderenfonds together with the other partners in this AAL project. The Care@Home platform was meant as a support system for elderly who want to live independently for a longer amount of time. It was intended to do so by means of personalized communication and offering of a service channel in their own home. Using the platform, elderly people would be able to easily contact friends, family and (in)formal caregivers by video, leave reminders for themselves or their spouse, order groceries online, find someone living nearby to help with a chore. The system would also show elderly people mild exercises they could do to stay fit, or show them local news.

As this study only used a mock-up version, not all functionalities were fully included in the platform yet at the point of testing. Therefore, tasks were adapted for this mock-up version of Care@Home to include the parts that were fully functional. These were the main

page, the address book, agenda, contact page (e-mail function) and the 'neighbourhood' page which contained two videos, one about local news and the other had exercises elderly people could do at home.

2.2.2 Questionnaires

The first questionnaire that was given to participants consisted of demographic information. The second questionnaire measured previous experience and self-rated expertise with technical devices and consisted of ten items. Since Care@Home was an online system running on a smart-TV, participants were not only asked whether they had online experience using a computer or a laptop, but also whether they had worked with a smart-TV, tablet or smartphone before. Last, system satisfaction was measured with the After Scenario Questionnaire (ASQ), a three-item seven-points Likerts scale measuring satisfaction with the system for each task, based on the ease of the task, self-rated task-completion task and support that the system provides during the task (J. R. Lewis, 1995). This last questionnaire was used to collect information for further development of the Care@Home platform for Nationaal Ouderenfonds.

2.2.3 Tasks

Participants were presented with the list of the tasks (see figure 2 below). Ten tasks in five parts of the system had to be completed, but after the first two participants worked on these tasks, it seemed ten tasks were just too much so it was chosen to delete one of the tasks that shared the most similarities with one of the other tasks from the list, so participants had to complete nine tasks in total. Tasks were given in a story wise way, to make it easier for participants to grasp the concept of what Care@Home could do in a daily life setting.

Task	Description
1	Check whether there will be events in the upcoming week. Are there any?
2	You'll need to call your GP tomorrow. Create a reminder so you won't forget this, then put it in your agenda for tomorrow at 12:00 'o clock.
3	Add Gert Dijkstra to your contacts/address book.
4	The phone numer of Miep Jansen is incorrect. It is supposed to be 034 – 669555. Please correct it.
5	Oops! Apparently, more people are named Gert Dijkstra, and it happened that you accidentally added the wrong one to your contacts. Please remove him from your contacts.
6	Check your e-mail. Did you receive any new e-mails? Please remove the e-mail titled 'tea'.

7	You have received an important e-mail about C1000 (local supermarket). Please send it forward to one of your contacts.
8	The day after tomorrow is your birthday, congratulations! Add this festive event to your agenda.
9	Since it will be your birthday, it would be nice to have some visitors over. Please write an e-mail to invite one of your contacts to come over for a cup of coffee for your birthday. Check whether you've sent the e-mail afterwards.
10	It is a beautiful day and you feel like working out! Luckily for you, Care@Home has a built-in option to help you with this. Please find this option and open it.

Figure 2*: Task list*

2.2.4 Think Aloud Protocol

It was chosen to use a think-aloud protocol during the studies. Originally described by Karl Duncker (1945) for use within the psychology of problem solving, Jakob Nielsen used it for usability later on and described it as "*the single most valuable usability engineering method*" (1993), thinking aloud basically consists of letting the user use the system while simultaneously verbalizing their thoughts. The researcher can stimulate thinking out loud by asking the user questions such as 'what do you think this message means?'(Nielsen, 1993).

Two ways in which the think-aloud method is used mostly are concurrent (thinking out loud while using the system) and retrospective (after each task completion). Research has shown that users are not more positive towards one method or the other, but experience the retrospective version to be more of a disturbance for the test situation (van den Haak, Jong, & Schellens, 2003) . In this study, it was chosen to use concurrent think-aloud in this study, mostly to lessen the length of the entire testing procedure for the elderly participant.

2.3 Apparatus

For running Care@Home, a 46" Philips 8000 series Smart LED TV was used (46PFL8007T) combined with a pointer remote control for navigation. Participants were placed in front of the TV. Two camera's were used for recording. The first was a JVC camcorder (model no: GZ-EX315BE) which was placed next to the participant to film their screen actions. The second camera was a webcam (Philips webcam model no. SPC640NC/00) which was placed on the TV itself to film the participants interaction with the remote control and his or her facial expressions. This webcam was connected to a laptop (ACER Aspire 5749 series) which was placed next to the TV (see figure 3 below).

How do elderly people learn to work with an online system?

2.4 Procedure

Trials took place in a room in retirement home De Koperhorst which was also used to organize cooking- and computer classes for the elderly residents. The participant was welcomed and the goal and the procedure of the study were explained. S/he then had to fill in an informed consent and the questionnaires regarding demographics and previous experience. The participant was asked whether s/he would mind if his or her face would be filmed and if so, the position of the camera (C2, see figure 3 below) was changed so only his/her hands and the remote control would be visible. The researcher would then explain a little about the system and would allow the participant to go and take a look at the system for a minute while setting up the other camera (C1, see figure 3 below). Before starting, the participant was told again explicitly that this study was to test the system, and not the participants' skills, and that this was also the reason that the researchers would give as little help as possible during the tasks. As described above, a think-aloud protocol was used, while the researchers also asked questions during the trials.

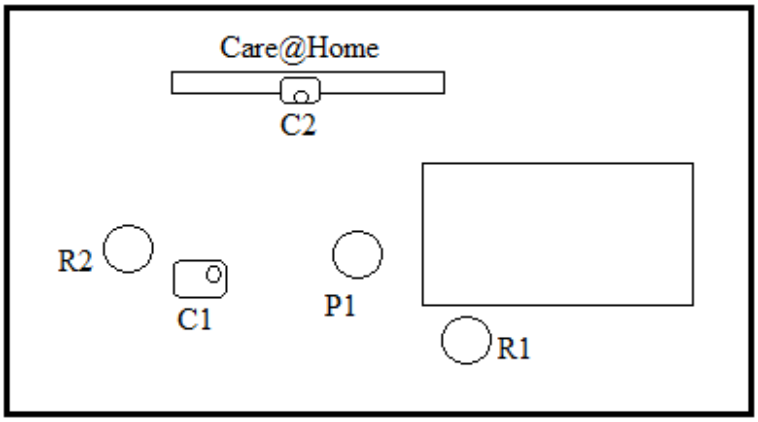


Figure : *location of cameras (C1, C2) and participant (P1) relative to the system. At the right there was a table where both one of the researchers (R1) and the participant would sit, also to fill in the questionnaires during the trials. The other researcher (R2) would control the cameras. Cameras filmed both the screen (C1) and the participant (C2).*

The participant then had to complete the tasks. After completion (or stopping) each task, the participant had to fill in the ASQ about that specific task and then continue with the next. If the participant was visibly tired (or indicated this by him/herself), s/he could take a short break in-between tasks. Also, some participants really struggled with the tasks – for these participants, a number of tasks were left out. After the last task, the first test round was

over. The participant could take a 30 minute break while one of the researchers would reset the system to default-values. The participant would then complete all of the tasks again, filling in the ASQ after each task. After the second round, the participant was done for that day. S/he was thanked for participating and received a €20,00 gift certificate for his or her help. As this was a longitudinal study, all participants were appointed for a third session after one week. This session was the same as the second one; Complete the tasks and fill in the ASQ questionnaire after each task.

2.5 Data gathering and analysis

First, camera recordings were imported into Morae (Techsmith) to make it easier to review them and add notes. First, Morae was used to carefully note the time on task for each task in every round for every single user. Furthermore, Morae made it possible to add notes to every recording, so incidents, or the problems participants ran into while trying to complete the tasks, could be noted down with ease. After reviewing all 60 recording (3 x 20 participants), incidents found were then clustered in order to find the underlying usability problems and their causes.

2.5.1 Matching process

After reviewing all recordings and noting down every incident in Morae, the total number of incidents found was 867. These incidents needed to be clustered together into usability problems, for which the method described by Lavery, Cockton and Atkinson (1997) was used (Haar, Schmettow, & Schraagen, 2013; Hornbæk & Frøkjær, 2008). In their research, Lavery and his colleagues found that analytical evaluations as they were, empirical testing based on heuristic evaluations or task analysis of design (without a checklist or questionnaire) were not good enough; Empirical testing did not find all usability problems, while task analysis was found to be ineffective as it ignored system output, leaving no possibility to detect poor feedback. Human computer interaction (HCI) principles needed to be. Eventually, they decided that since the definition of a usability problem as they saw it, being "*an aspect of the system and/or a demand on the user which makes it unpleasant, inefficient, onerous or impossible for the user to achieve their goals in typical usage situations*" contained multiple aspects, problems should be classified according to multiple components (cause, breakdown, outcome, context), using an analytical framework. Based on their work, the components used in this study were the following:

How do elderly people learn to work with an online system?

Time/Place:	*Not mentioned by Lavery et al (1997) but added as an extra factor for context.*
Context*:*	*What was the participant trying to do? What was his or her goal?*
Cause*:*	*Why did it happen? What is the type of the cause?*
Breakdown*:*	*Was there a breakdown between the user and the system? If so, how did the user react to it?*
Outcome*:*	*What was the effect of the incident on the users performance and work? Was the user able to continue?*

Some incidents were only found by one user, or even occurred only once. These incidents were still taken into account while matching, because these as well might pose a serious problem for system flow and for learning. Some incidents were also related to the specific situation of the day testing occurred (such as a failing internet connection because of renovation. These incidents were noted, though these were not taken into account as usability problems (Følstad, Law, & Hornbæk, 2012). Next to these components, another component was taken into account based on research by Hornbæk and Frøkjær (Hornbæk & Frøkjær, 2008). Their 'similar change' method included that incidents belonging to the same usability problems often also need the same design solution to fix them. It was shown that this method produced extra groups of (single) usability problems compared to other matching methods and therefore it was added as an extra component. Last, a component called 'problems resistance to learning'was added. This component shows how easily a problem solves itself or not, or how easily participants learn how to overcome it, and was also used as a basic measure for deciding which usability problems the Care@Home team should solve first (also see figure 4 below for the used format).

After filling in the components for all incidents found, incidents were grouped together based on these components like a technique described by Hornbæk and Frøkjær (Hornbæk & Frøkjær, 2008), based on the model by Lavery et al.(1997). The more components for incidents were alike, the more these incidents were alike and thus forming a usability problem. This way, a list of 129 usability problems was found. For the Care@Home platform improvement reports, usability problems found this way were written down based on the part

of the platform where they occurred the most often (main page, address book, agenda, contact, neighborhood) if possible, concluding with a list of problems that occurred throughout, or were not specifically related to a certain part of the platform. Improvement suggestions for these problems were provided as well. This list can be found in appendix B.

Observer/Nr	**User No.**	**Task/No** **Session/No**	**Start Time:** **0:00.00**
Incident No (of total)	Context	Cause	Breakdown
Time	Outcome	Design Change	Evaluation regarding the problem's resistance to learning

Figure 4*: Format used for rating each incident*

2.5.2 Usability problem classification

The classification models from Rasmussen (1983) and Zapf et al. (1992) that were described in the introduction were used to create a new guideline for classifying usability problems found in this study. In this guideline, the first difference was set between possible knowledge errors and the seven other possible action-based errors. The definitions of all these seven other types of errors were provided by Zapf et al. (1992) and Barendregt et al. (2006) – see appendix C for an overview. The next step in the guideline was to determine the level of regulation. Earlier work by Haar et al (2013) already provided a model based on the work of Rasmussen (1983) for classifying usability problems. As the skills, rules, knowledge theory by Rasmussen (1983) is quite similar to the level of regulation by Zapf et al (1992), this model was usable as well. After deciding on which level of regulation the usability problem was occurring, he next step was to check in which level of the action process a usability problem was taking place.

The guideline consisted of seven steps. Each step contained a number of questions regarding the choice that needed to be made between two options. These questions could be answered with a 'yes' or 'no', which would lead to two options. The option chosen would then lead either to a follow-up step (with new questions, leading to a next step or final category) or to a final category. A final category would contain a description of that category, control statements and examples to make sure that a usability problem was assigned to the

right category. The full guideline also included the needed data prerequisites for classification and a thorough description of every error and can be found in appendix C.

The classification guideline was used for all usability problems that were created with matching earlier. Two usability problems could not be classified either and were thus put in a 'unknown' category. Not being able to classify all found usability problems in a newly created classification system is not uncommon, as it happened in other classification studies as well, (Haar et al., 2013).

2.5.3 Frequency and persistency for problem ranking

One of the research questions entails which usability problems are the worst and best to be learned. Therefore, criteria need to be established to be able to compare usability problems. In usability research, severity ranking is used to prioritize which problems need to be tackled first (Dumas & Redish, 1999). There is a diversity of ways to perform a severity ranking, most deduced from the theory of Nielsen (1995), in which he states that the severity of a usability problem can be classified according to three factors (frequency, impact, persistence) with an equal weight of importance. Scoring high on these three factors means that a usability problem is severe.

For this study, frequency and persistence were measured while impact was not. It was chosen to do this because the full score of severity says something about the seriousness of a usability problem, but not purely about learnability. Persistence shows whether a problem occurs over time or not. It is very well possible that a problem occurs in the first round, but people remember it and do it right in the next round (e.g.; not knowing which button in the menu to press the first time, but remembering which one it was the second and third time). Persistence shows whether a problem can be overcome (learned) over time or not. Frequency shows how many times a problem occurs, or how many people encounter a certain problem. Combined with persistence, this gives an idea of whether a problem is hard to learn or easy, and whether this is the case for many people, just a few, or even just one person. Impact, on the other hand, measures what a usability problem does in regard to workflow, whether a user can continue to work with the system after the problem occurs or not. Impact states more about the seriousness of a problem in itself than it does about the learnability of it, therefore it was not used as a measure for problem ranking in this study.

2.5.4 Criteria for problem ranking

In order to rank all usability problems, criteria were set up in order to make a list. For persistency, a detection matrix was made for each of the three rounds. Combining these three matrices gave an overview of the presence of each problem for each user over time. If a problem occurred in one round, it was coded 1. If it did not, it was coded with a 0. This was done for all problems in each round. As a lot of these codes were the same, a pattern could be deducted concerning persistence. (Zandbergen, 2015). There were four groups of codes, which were in order of severity:

1) Overall persistency: problem occurs in each round (code 1-1-1)
2) Persistency, late onset persistency (code 1-0-1 or 0-1-1)
3) Semi-persistent but possibly learnable (code 1-1-0)
4) Non-persistent (code 1-0-0, 0-1-0 or 0-0-1)

This order was chosen because while persistency is an often overlooked factor in usability, it can have devastating effects on the user friendliness of a system. Usability problems do not have the same rate of appearing or disappearing (Kjeldskov, Skov, & Stage, 2010). If a user encounters the same problem every single time, even if it is a small problem, this may lead to a lowering of willingness to use a system. After each problem was given a persistency rate based on their code, frequency had to be decided. Frequency was based on the percentage of participants who encountered the problem in each round and was therefore also given for each round. An initial order of severity was decided based on the persistency group a problem was in, but within a group, rules for frequency were set up as well. Taken into account was the cutoff from Ruby and Chisnell (2008), who define a problem that is accounted by 30% or less than all participants as a problem that is not severe but needs improvement nonetheless. This was based on a single measure, and not on a longitudinal study. As this study was longitudinal and included a group of participants that have, in general, a different attitude towards and less experience with technology than younger participants, it was chosen to create not just one cutoff stage, but different ones. For extremely severe problems, these were the following:

1) Group 1 persistency and a frequency of a minimum of 30 percent of all participants experiencing this problem in at least two rounds.

2) Group 2 persistency frequency of a minimum of 30 percent of all participants experiencing this problem in at least two rounds.
3) Group 1 persistency and a frequency of a minimum of 30 percent of all participants experiencing this problem in one round, and a minimum of 25 percent experiencing this problem in another round.
4) Group 1 persistency and a frequency of a minimum of 30 percent of all participants experiencing this problem in one round, and a minimum of 25 percent experiencing this problem in another round.

The list of cutoff stages was further continued until group 3 and 4 persistency were also included, and until the frequency was just 5 percent in only one trial. This way, not only the worst usability problems could be deduced, but those learned the easiest as well. As all problems were listed in order of severity, rules based on the above were made for the entire set of 129 usability problems so they could be ranked accordingly. As more than one usability problem could adhere to one of the rules from one group, another set of rules for within each group was added to determine the ranking. Using these rules, all 129 usability problems were placed in a ranking order from most hard to easiest to learn (for the full list of usability problems, see appendix E). The second set of rules used were the following:

1) Problems that increase in frequency from round 1 to 2 and from round 2 to 3 time have priority over those that only increase in one round and decrease or stay the same in the other round.
2) Problems with a higher frequency in round 3 than in round 1 have a priority over problems that have a lower frequency in round 3 compared to round 1.
3) Problems with a frequency equal in round 1 and 3 have priority over problems with a frequency that is lower in round 3 than it is in round 1.
4) Problems with a frequency that first lowers from round 1 to 2 and then rises from 2 to 3 have priority over problems that first rises from round 1 to 2 and then lowers from 2 to 3.
5) Problems with a frequency that stay the same in round 2-3 have priority over problems that lower from round 2-3,
6) For these rules, higher frequency percentages have priority over lower frequency percentages.

How do elderly people learn to work with an online system?

3. Results

3.1 learning curves

Time on task was used to measure learnability. The time on task (until completed or until user gave up) was measured for each task on each round. Some users had such trouble working with the Care@Home system and the remote that certain tasks for them were skipped. This was also done due to time-related constraints. Using SPSS 22.0, time on task for these skipped tasks were taken into account as missing values. Time on task was also corrected for incidents beyond the users' control such as the internet shutting down, or empty batteries in the remote control. To provide an overview, time on task was plotted for each round and each task, taking the average time needed by all users who had completed that specific task. Figure 5 below shows that for almost every task, the time needed in the third round is lower than the time needed in the first round, with the time for some tasks staying relatively the same over time. As this only showed a general overview regarding whether tasks were learned or not, time on task was also plotted for each individual participant, for all tasks he or she had completed over all three rounds. This made it possible to see the individual learning process over time. It was very visible that individual learning differed greatly. For almost all participants, time spent on the tasks decreased over time. For some participants, however, time on task increased over time for most of the tasks they performed, or stayed the same with the exception of one or two tasks. Two examples of this are given in figure 6 and 7 below. While time on tasks overall seems to lower in the second round compared to the first, it increases again in the third round. The individual learning curves of all participants can be found in appendix F.

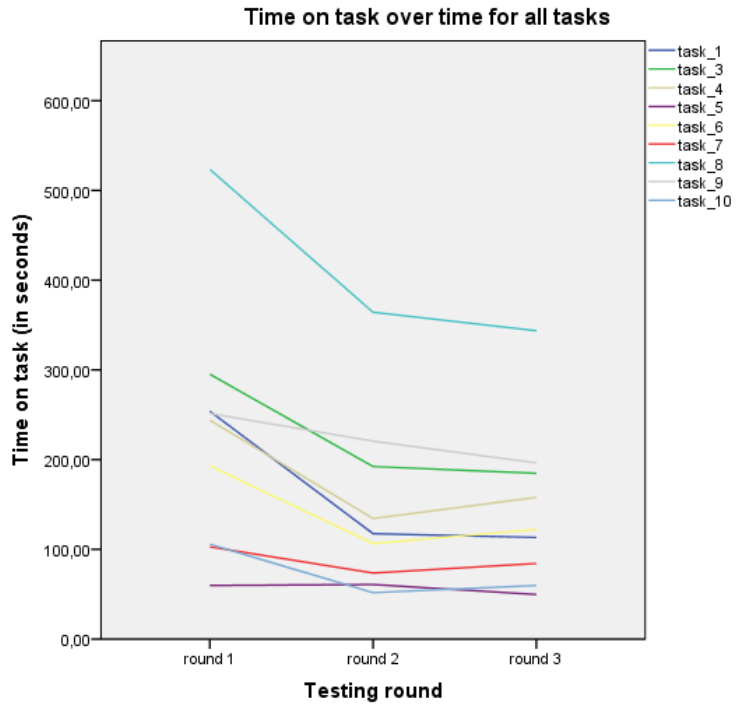


Figure 5: *Time on task (ToT) was plotted for all tasks , using the average ToT from all users who completed that task (missing values left out). Task 2 was left out earlier, so there was no data for it. As can be seen, task 8 took participants the longest to complete over all rounds.*

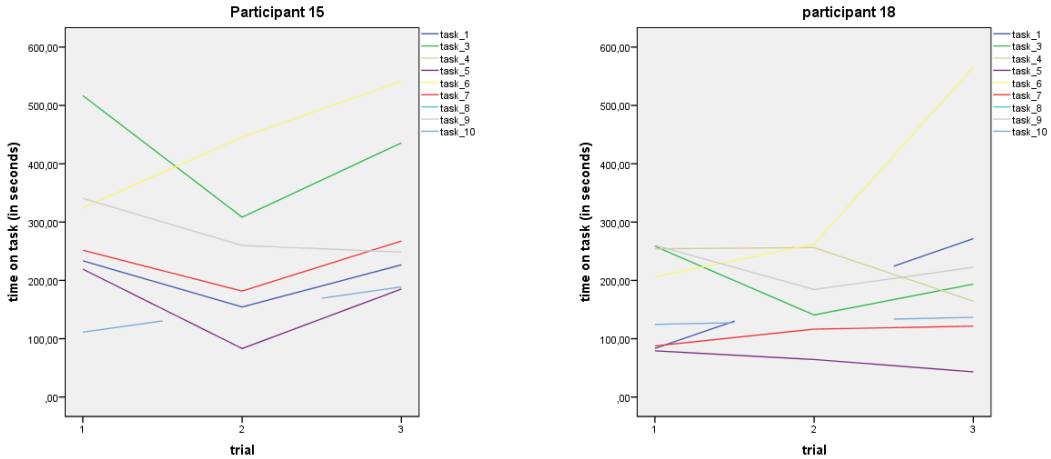


Figure 6 (left): *While the time on tasks overall shows a steady decline in round 2 compared to round 1, there is a rise in time on task in round 3, when compared with round 2 and mostly with round 1 as well. The gap in task 10 can be explained because the participant did not do this task in round 2. Task 4 was only done in round 1 and therefore, shows no line. Task 8 was left out completely.*

Figure 7 (right): *This graph shows a difference compared to the previous one in that time on task in round 2 is overall not lower when compared to round 1. In round 3, time on task overall is either higher compared to round*

1 and 2, or around the same. Tasks 1 and 10 were not completed in round 2, explaining the gaps in the graph. Task 8 was not completed in either round 2 and 3 and therefore shows no line at all.

To see how time on task changed over the course of the three trials, a generalized estimating equations model was performed. Assumptions were checked by using a guideline written by Zuur, Ieno and Elphick (2010, also see appendix G) and a model was chosen accordingly. It was chosen to use a gamma distribution with a log link and an autoregressive correlation matrix. Age, gender and previous experience in components were taken into account as well as it was possible that these would influence time on task over the course of time. The results can be seen in table 1 below:

Table 1: *GEE results for time on task*

Parameter	**B**	**Standard error**	**95% Confidence Interval**		**df**	**Sig.**	**QICC**
			Lower bound	Upper bound			
Intercept	1,534	,5011	3,665	5,629	1	,000	317,55
Trial 1	0[a]	.	.	.	.	.	
Trial 2	-,362	,0635	-,487	-,238	1	,000	
Trial 3	-,342	,0733	-,486	-,199	1	,000	
Gender (M)	0[a]	.	.	.	.		
Gender (F)	,041	-,187	-,187	,269	1	,727	
Age	,013	,001	,001	,025	1	,028	
Previous exp components	-,112	-,165	-,165	,059	1	,000	

a. Set to zero because this parameter is redundant. This is the parameter to which the others are compared.

As the results show, participants on average did became faster using the Care@Home system over time. Compared to trial 1, participants were on average 36% faster in trial 2. In trial 3 participants were 34% faster compared to trial 1. These findings show that time on task is the lowest in trial 2 and then rise a bit in trial 3, albeit that time on task in trial 3 is still lower than in trial 1. The results also show an effect of age. For each unit of age extra, time on task would increase with (exp(0,013) = 1,01) 1%. While this effect seems to be small at first, it means that an age difference of 40 years (so, compared to someone who is 40 years younger) would increase time on task by (1,01^40 = 1.49) 49%. . For every increase of previous experience in components with one unit, time on task would lower with (1 - (exp(-,112) = 0,89) =0,11) 11%.

How do elderly people learn to work with an online system?

3.2 Most severe mistakes

After applying the ranking criteria on all usability problems, the problems were listed in order. The initial plan was to present first 20 problems on the list. However, when ranking the problems according to the criteria, it was found that only the first 9 problems on the list had a frequency of 30 percent or more in more than 2 trials. Thereafter, the next 4 problems had a frequency of 30 percent or more in one trial but just 25 percent in a second trial (see figure 8 below). The 14th problem on the list had a frequency of 25 percent in two out of three trials. Problem 15 to 20 only had a frequency of 25 percent in one trial and 20 in a second trial. As the cutoff criteria for the most severe usability problems were determined on at least 30 percent in one trial, it was chosen to take the top 13 worst problems only, as these classified within this cutoff (Rubin & Chisnell, 2008a). These problems are listed in table 2 below and in figure 8 as well:

Table 2: *The most severe usability problems found*

List no.	**Description of usability problem**	**Prevalence in percentages (trial 1, trial 2, trial 3)**	**Type of problem (classification)**
1	User does not know where to find the movie with the exercises (it's under "mijn buurt" (my neighborhood))	60-65-65	sensorimotor
2	User is at the correct screen but goes back to the main menu because he/she believes it's not the right screen	30-25-60	recognition
3	User thinks events (agenda) are in the address book or contact, or thinks addresses are in the agenda (confuses the two)	40-30-55	thought
4	User goes to the wrong day (e.g. goes to today when it has to be in two days)	60-75-50	sensorimotor
5	User thinks the legend pictures are buttons	55-25-45	thought
6	User accidentally goes back to the previous screen	60-25-45	sensorimotor
7	User thinks he/she received new e-mails while this is not the case (or the other way around)	60-35-35	thought
8	The user cannot find the pointer	50-25-30	sensorimotor
9	User thinks there are events next week while there aren't any (misreads the pink/purple part of the legend)	50-35-30	thought

10	User clicks on the phone number itself instead of on the button "change phone number" next to it, in order to edit a contact's phone number	20-25-45	thought
11	User tries to use TAB to get to the next box for filling in time (event). Leads to a mistake as the system recognizes it only as an extra symbol (and three are too much)	25-30-25	habit
12	User forgets to hold the Fn key to type a number instead of a letter -> types a letter	35-5-25	memory
13	User does not know how to confirm adding someone to the address book	30-25-25	recognition

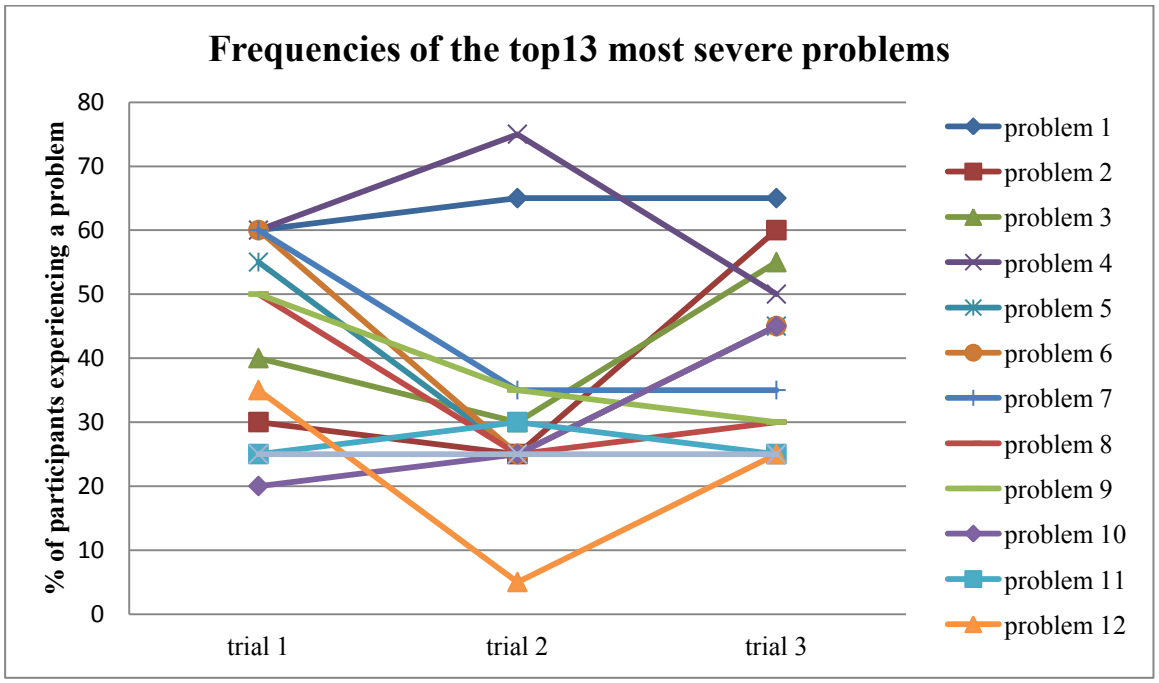


Figure 8: *the percentages of participants that encountered the most severe problems over time*

The problem ranked highest on the list regards where the movies about exercises at home could be found. In the first trial, 60 percent of users did not know where to find this page, which rose to 65 percent of users not being able to find this page in both trial 2 and 3. The page with the exercises was located under a menu called ‘my neighborhood’, where participants could also find news about their neighborhood and information about (upcoming) local events. The usability problem second on the list happened throughout the entire trial, as participants were at the right page to do one of the tasks but thought they were not and thus

went back to the main menu. This usability problem was one that happened to 30 percent of all participants in the first trial, 25 in the second trial and then rose to 60 percent of participants experiencing it in the third trial. The third usability problem had participants confusing what could be found in the agenda and in the address book, with 40 percent of users experiencing this in trial 1, 30 in trial 2 and then another rise to 55 percent. The problem ranked fourth happened when participants had to fill in an event happening in two days in the agenda. While it was first thought that the task itself was written in a confusing way, asking participants about what they did provided information that most of them failed to see what day was mentioned in the agenda and those who did were unsure on how to select another day. For this reason, 60 percent of participants encountered this problem in trial 1, rising to 75 percent in trial 2 and eventually a drop to 50 percent of all participants in the third trial.

Problem number 5 happened when participants had to use the agenda, as they thought the legend showing what all colors meant were buttons themselves. In the first trial, 55 percent of participants clicked the legend, which went better in the second trial with only 25 percent of participants clicking the legend as it they were buttons. In round 3 however, there was a rise again, as 45 percent of all participants clicked the legend again. The sixth problem happened in all tasks. Participants pressed the back button on the remote control and accidentally went back to the previous page. It happened to 60 percent of participants in trial 1, with a steep decline to 25 percent in trial 2, and a rise to 40 percent again in trial 3. The seventh problem regarded e-mail. While new e-mails were shown in bold letters, with a slightly discolored background compared to the older, already read e-mails, 60 percent of participants failed to see this in the first trials. This lowered to 35 percent in the second and third trial. Ranked number 8 on the list was the problem when participants were unable to locate the pointer on the screen. This either happened because the pointer was halfway off the screen, or because the color of the pointer did not have enough contrast compared to the background. It happened mostly in the first trial, with 50 percent of the participants having difficulty finding the pointer at some time, which lowered to 25 percent in trial 2 and had a slight rise to 30 percent in trial 3. The next usability problem on the list was regarding the agenda, when people either misread the legend or failed to read the legend at all to see whether there would be events in the next week. While half of the participants had trouble with this in trial 1, this lowered to 35 percent in trial 2 and eventually 30 percent in the final trial. Ranked 10th in the list was when users would have to change a phone number in their

address book. In order to do so, they had to click the button “change” next to the phone number, but users clicked on the number itself in order to change it. In trial 1, 20 percent of all users did so, rising to 25 percent in trial 2 and rising even more to 45 percent of participants making this mistake in trial 3.

The 11th usability problem on the list happened when users had to fill in the time for an event in the agenda. As this box was rather small, users wanted to use the TAB-button on the remote control keyboard to get to the next box for filling. This was not possible and created an extra symbol in the time box which was then seen as an error by the system as there were only two symbols allowed for time. This occurred to 25 percent of users in trial 1, then a slight rise to 30 percent in trial 2 and then a slight lowering to 25 percent again in trial 3. Usability problem 12 regarded typing as well, and happened when users wanted to type a number or a symbol instead of a letter. In order to do so, they had to press a small key labeled ‘Fn’ first and then press the key with the number or symbol they wanted to type. The problem was that users forgot to press this Fn key beforehand. 35 percent of users forgot it in the first trial, with a steep decline to only 5 percent in the second trial and then a rise back to 35 percent again in trial 3. Last, number 13 on the list of most severe usability problems happened when users did not know how to confirm adding someone to the address book. After participants found the person they want to add to the address book, they have to click the picture of that person in order to get to the confirmation screen. These users were aware that they had to confirm the adding (so they did not forget to do it) but did not know how. In trial 1, 30 percent of all users did not know how to confirm an added person, with a slight decline to 25 percent in trial 2, which stayed the same in trial 3.

3.3 Least severe mistakes

The usability problems with the lowest rates of frequency, impact and persistence were also noted. The last 20 usability problems on the ranking list were all problems that only happened to 5 percent of the participants in just one trial. A lot of these problems had the same rating both in persistency group and frequency percentage. It was chosen to rate those problems that had the 5 percent error in the first trial as the easiest to learn, as for these problems it was the most certain that they were learnable. Next were those problems that had the 5 percent frequency in the second trial, followed by those that had the 5 percent frequency in the third trial. Within these classes no difference could be made regarding problem

severity, so problems were listed in a random order. This means that problems 1 – 11, problems 12-17 and problems 18-20 basically have equal levels of severity. The top 20 problems that were the easiest to learn can be seen below in table 3:

Table 3: *The least severe usability problems found*

List no.	**Description of usability problems**	**Percentages (trial 1-2-3)**	**Type of problem (classification**
1	User has trouble seeing the cursor for typing (can't type because of this).	5-0-0	sensorimotor
2	User uses two hands to direct the remote control (e.g. against a tremor) which makes using the remote control too difficult.	5-0-0	sensorimotor
3	User thinks that the event he/she just added is not added yet.	5-0-0	recognition
4	User tries to make a mail first and then select a sender (can only be done the other way around).	5-0-0	habit
5	User is looking for an option to send a mail to multiple persons at the same time (there is no such function available).	5-0-0	habit
6	User accidentally types O instead of 0.	5-0-0	habit
7	User notices the hh/mm boxes but forgets to fill in the time anyway.	5-0-0	memory
8	User thinks the exercise movie can be found under 'video contact' (in contact).	5-0-0	thought
9	User clicks on the legend in the agenda because he/she thinks this will open a list of events.	5-0-0	thought
10	User has no idea how to hold the remote control (is literally saying that he/she does not know).	5-0-0	knowledge
11	User can't go back to the main screen when a movie is opened (only with the Esc. Button on the remote).	5-0-0	knowledge
12	User adds numbers on the wrong place (when changing a phone number, e.g. in the front or	0-5-0	sensorimotor

	in the middle instead of the last 3 numbers).		
13	User checks the address book for upcoming events because "that's the way I do it at home. I keep my appointments in my address book".	0-5-0	habit
14	User looks for the contact point of the remote to the TV (thus, compares the remote to a regular remote).	0-5-0	habit
15	User goes back to the main menu before going to the next screen (instead of directly selecting the right screen in the left menu).	0-5-0	thought
16	User thinks clicking 'postvak UIT' will send the e-mail he/she created.	0-5-0	thought
17	User adds a dot after the name of the contact he/she wants to add (which the system cannot find).	0-5-0	thought
18	User wants to send an e-mail to a contact that's not in the list (not possible yet).	0-0-5	habit
19	User goes to the wrong edit (e.g. "change details" or "change address" instead of "change telephone number").	0-0-5	thought
20	User edits the phone number of the wrong person (not related to remote control).	0-0-5	thought

While most of these problems seem clear, some need a little explaining: problem number 1 happened because when a participant wanted to type something in a box. This box needed to be selected and there would be a cursor in this box indicating that it was selected and s/he could start typing. While a lot of participants did not notice this cursor at all and just started typing right away, one participant could not see whether the cursor was in the box and therefore thought it was not possible to type at all.

Other problems were a combination of things the Care@Home system was not yet able to do and users' experience with doing things another way than the Care@Home system intended. For example problem number 4, where one participant was used to Gmail for e-mailing and wanted to send an e-mail in the way s/he usually did, by first typing the e-mail itself and then deciding who to send it to. This was not possible yet in the Care@Home

system so s/he had to figure out how to do it here. The same kind of thing occurred for problem 5, when a participant wanted to add multiple recipients at once "because I want to invite more than one person actually" (task 9), and problem 18, where a user wanted to send an e-mail to a contact that was not in the list, but this was not possible in the Care@Home system. Such problems were not just limited to the mail function but also occurred elsewhere, such as in the address book. One participants explained that at home, s/he would put both addresses and upcoming events all in one address book (problem 13). Having it divided into two different categories was therefore confusing. Another participant who was struggling with the remote control would eventually look for the contact point on the TV so s/he knew where to aim at, just like with a normal TV (problem 14). This was not possible because the Care@Home system did not have such a contact point.

Two other interesting usability problems were number 6 and number 11. Number 6 was not added under regular spelling errors but was given its own category. This was done because regular spelling errors were mostly made because of pressing the wrong button on the remote control keyboard unintentionally (sensorimotor error). In this case however, a participant noted the O consciously and pressed it because s/he thought it was a 0. When asked, this user proclaimed that s/he was used to having the 0 above the letters on a keyboard, which made this an error of habit. Last, problem 11 was interesting because it happened only to one participant in one trial, but could cause serious problems for usage of the end-product if it would have gone undetected. When a participant would go to the exercise movie ('my neighborhood, task 10), there was no button for returning to the previous screen. One could only return to the previous screen by pressing the escape button on the remote control but this was not mentioned on the screen or in the video. As no other participant had tried to go back to the previous screen (it was not part of the task), this was merely detected because one participant was curious about how to get back to the previous screen.

3.4 Classification and distribution of error types

While usability problems were ranked according to persistency and frequency, the distribution of types of mistake as classified by the classification guidelines were taken into account as well. Within the set of 129 usability problems, errors were found for each classification type. Figure 9 below shows how the types of errors were divided between the 129 different usability problems. As is visible, the most usability problems could be classified

as thought errors, followed by sensorimotor errors, habit errors, memory errors, knowledge errors, judgment errors, recognition errors and last, omission errors.

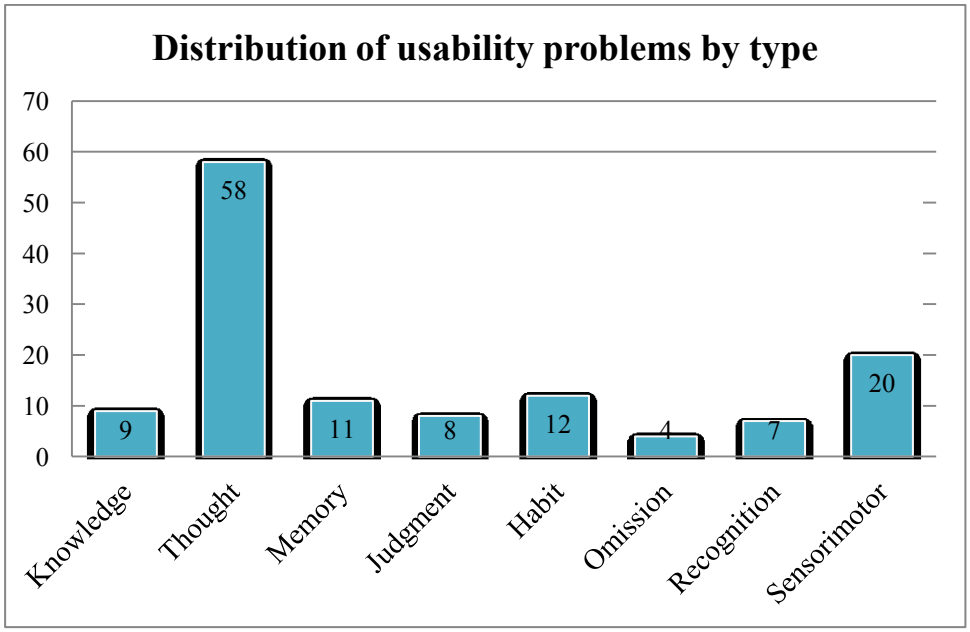


Figure 9*: distribution of usability problems by type within the set of 129 found usability problems*

In order to find out how the different types of errors were distributed among the ranking list, boxplots were used together with a summarized data table. This way, a complete overview of the range per classification type of error could be shown. This boxplot is shown below in figure 10, the table with summarized data can be found in appendix I As the Y-axis represents the ranking list, low numbers indicate errors that were high on the ranking list as given previously. The line with the label shows the median of error ranking, with the lower half consisting of more severe rated errors and the higher half of less severe rated errors. The boxplot shows that for errors of the type knowledge, thought, memory, judgment, omission, recognition and sensorimotor, the first quartile of errors within these types fall within the lower half of the error ranking. For knowledge, omission, recognition and sensorimotor errors, two quartiles fall within this lower range and for judgment errors, even the third quartile of errors within this type falls into the lower half of error ranking. It is worth noting that for knowledge errors, the second quartile merely falls within this the lower half, while for omission, recognition and sensorimotor errors, even a large part of the third quartile falls within the range of the lower half of ranking. For judgment errors, a part of the fourth quartile also falls within the lower half of ranking. However, ranging from 31 to 77, the range of

judgment errors is smaller than those of the other types of errors being mostly present in the lower half of the ranking.

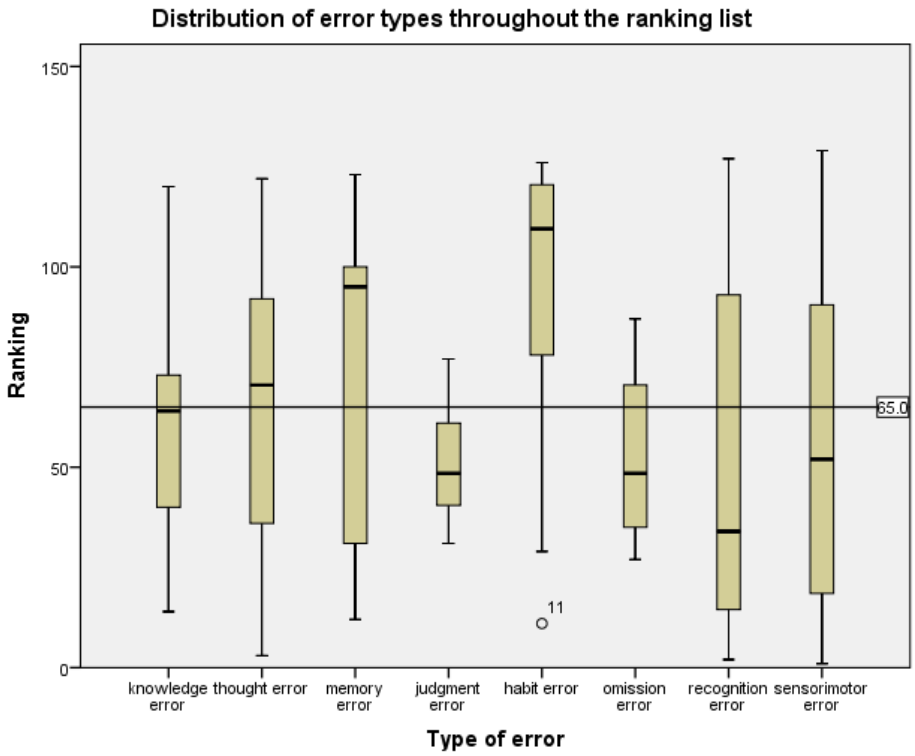


Figure 10: *The ranges of each type of error among the ranking list. The line with the label '65' represents the median of the ranking. Errors ranked lower than 65 are errors that are more severe than errors that are ranked higher than 65 on the list.*

Error types being mostly present in the higher half of the ranking are thought errors and memory errors, both having two quartiles above the median of the ranking and habit errors, having even three quartiles in the higher half of the ranking. For habit errors, one outlier was found (as shown below). It was chosen not to discard this outlier as this boxplot was used to show the complete range of error types throughout the entire ranking and discarding it would change the distribution of habit errors in this ranking.

3.5 Previous experience and its effects on time on task

Previous experience was given as two variables, one based on 8 questions that participants could answer about which devices they had already used in the past and whether they had ever been online and sent an e-mail before the experiment. Three participants (15%) currently owned a smart TV, but only one participant (5%) had ever worked with the smart

TV online functionalities (e.g. browsing the internet on the TV). Owning a computer, laptop or tablet seemed more relevant, as 13 participants (65%) owned one or more of these devices. Furthermore, when asked whether they had worked with a computer, laptop or tablet before (regardless of whether it was their own or not), 14 participants (70%) said that they had. These 14 participants were also asked where they had used a computer, laptop or tablet, with multiple answers being possible. 13 of them had used one of the aforementioned devices at home (65% of total) , seven had used it at their (former) job (35% of total) and two participants indicated that they also used a computer elsewhere, of which one said it was in a computer class and the other one did not explain where (10% of total). None of the participants had used a computer, laptop or tablet at their family or friends. 12 participants had surfed the internet before (60%), 11 had sent an e-mail to someone (55%).

The second measure was the number of hours participants would spend on average using a smart TV, tablet, laptop or pc, per month. Only one participant said s/he had worked with a smart TV in the last month with an average of 14 hours per week. Participants were more diverse in their amount of time working on a computer, laptop or tablet in the last month with answers ranging from not at all (8 participants, 40%) to an average of 28 hours a week (1 participant, 5%). The average amount of time spent on one or more of these devices was 7,67 hours (*SD* = 8,01).

As was already shortly mentioned in paragraph 3.1 and table 3.1, previous experience in components had a significant effect on the time spent finishing the task, in which time on task would lower by 11% for each extra component of experience. This model did not take into account possible interaction effects between previous experience and the second and third trial. Therefore, the same predictors were used for this model again, adding experience in hours as a predictor (despite the higher QICC value) using a gamma distribution with a log link and an autoregressive correlation matrix as well. The results can be found in table 4 below. This shows that there is one significant interaction effect for trial 2 and previous experience in hours. This means that, compared to trial 1, previous experience in hours in trial 2 is (exp(0.016) = 1,016) 1,6% higher. No other significant interaction effects were found.

Table 4: *Effects of previous experience on time on task including interaction effects*

Parameter	B	Standard error	95% Confidence Interval		df	Sig.	QICC
			Lower bound	Upper bound			
Intercept	4,643	,7055	3,260	6,026	1	,000	326,071
Trial 1	0[a]	.	.	.	.	.	
Trial 2	-,316	,0942	-,501	-,131	1	,001	
Trial 3	-,239	,0882	-,412	-,066	1	,007	
Gender (M)	0[a]	.	.	.	.		
Gender (F)	,032	,1037	-,171	,235	1	,756	
Age	,013	,0087	-,004	,030	1	,138	
Previous exp components	-,067	,0401	-,146	,011	1	,093	
Trial 1 * prev exp comp	0[a]	.	.	.	.	.	
Trial 2 * prev exp comp	-,060	,0410	-,141	,020	1	,143	
Trial 3*prev exp comp	-,067	,0419	-,149	,015	1	,110	
Previous exp hours	-,010	,0135	-,036	,016	1	,455	
Trial 1 * prev exp hours	0[a]	.	.	.	.	.	
Trial 2 * prev exp hours	,016	,0073	,002	,030	1	,030	
Trial 3 * prev exp hours	,011	,0095	-,008	,030	1	,250	

3.6 The role of previous experience on the most and least severe problems

The goal here was to find out whether the average level of previous experience differed between those participants that encountered the most severe problems and those that encountered the least severe problems. First, the average levels of experience of participants encountering the hardest mistakes were compared to the average levels of experience of those participants encountering the easiest problems to see whether there would be a difference in these levels. These comparisons showed that there was no significant difference in the hours spent using a computer, laptop, tablet or smart TV for participants experiencing the hardest problems (*Mdn* = 13.3) compared to those who experienced the easiest problems (*Mdn* = 19.4), $U = 82$, $z = -1.79$, *ns*, $r = -0.31$. Neither was there a difference in the scores related to

what kind of experience participants had between the participants encountering the hardest mistakes (*Mdn* = 17.85) and those participants encountering the easiest (*Mdn* = 16.45), *U* = 119, *z* = -0.407, *ns*, *r* =-0.07.

3.7 Previous experience and number of usability problems encountered

Even though there were no differences found between levels of previous experience between participants who encountered the hardest and the easiest mistake, levels of previous experience were also used to see whether participants with a lower level of experience encountered more usability problems than those with a higher level of experience. For further analysis, the number of usability problems encountered was taken as a dependent variable. These were taken from the matrices, so for each participant, all problems encountered in all three rounds together. Both scores on experience were taken as predictors. Age was taken into account as a predictor as well, as it was plausible that older people in general were expected to have less experience with modern technology (PewresearchCenter, 2014). Assumptions were checked using the protocol for data exploration by Zuur, Ieno and Elphic (2010) Taking into account repeated measures and possible effects of learning and fatigue, it was chosen to perform a generalized estimating equations with an autoregressive working correlation. Here, it was chosen to use a negative binomial distribution with a loglink instead of a Poisson distribution, as the variance was not equal to the mean (see appendix G for the full assumption check and appendix H for the SPSS feed for the models used).

Table 5: *Results of GEE for total number of usability problems found*

Parameter	B	Standard error	95% Confidence Interval		df	Sig.	QICC
			Lower bound	Upper bound			
Intercept	2,892	,0526	2,789	2,995	1	0,000	14,504
Exp hours	-,112	,0162	-,144	-,080	1	0,000	
Exp compartments	,008	,0305	-,051	,068	1	0,780	
Exp hours * Exp compartments	,020	,0031	,014	,026	1	0,000	

The analysis showed a main effect for experience as measured in hours of using a pc, laptop, tablet or smart TV in the previous month (see table 5). No effect was found for

experience measured in components. For experience in hours, it was found that for each extra single hour of using a pc, laptop, tablet or smart TV per month, the number of errors encountered would drop by (1- (exp(-0.112)) = 1- 0.89 = 0.11) 11%. An interaction effect between both measures of experience was found as well. It was found that participants with a higher score on experience as measured in hours also had a higher score on experience as measured in components. However, exponentiation of the the beta value showed that the effect itself was rather small, showing that for each extra hour of experience per month, the score on experience in components was raised by (exp(0.020) = 1,02) 2%.

3.8 Previous experience and types of error

The next step was to see whether and how previous experience would influence the number of different types of mistakes made. Again, data from the matrices was used. Because some types of errors would hardly occur as opposed to others, it was chosen to use the same approach as in the study by Zandbergen (2015) and converge the types of errors back to the three types as proposed by Rasmussen (1983). Sensorimotor errors were taken as a single category for problems on the skill-based level, thought errors, memory errors and judgment errors were taken together as the category of rule problems and habit errors, omission errors and recognition errors were taken together as the knowledge problem category. As the number of errors from the knowledge-based category as developed by Zapf et al (1992) was small, it was chosen not to use this group of errors as a dependent variable. The three remaining categories were used as dependent variables in different analyses. For all three datasets generalized estimating equations were performed with autoregressive working correlations to account for repeated measures and possible learning and fatigue effects. Gender, age and the two experience scores were used as predictors. For the datasets of rules and knowledge errors, a negative binomial distribution with a loglink was used in order to control fro overdispersion. The dataset for skill errors had a relatively small underdispersion. As the other option to deal with underdispersion, a Maxwell-Conway Poisson distribution, could not take into account repeated measures and possible learning or fatigue effects and the underdispersion was very small, it was chosen to use a Poisson distribution with a log linear instead. For all three datasets, models were compared using either age or experience measured in hours as a predictor, as these two predictors showed to be redundant. The model with the lowest QICC value was then chosen for each dataset (see appendix G for the full assumption check and choice of models).

Table 6*: Results of GEE for number of skill problems*

Parameter	B	Standard error	95% Confidence Interval		df	Sig.	QICC
			Lower bound	Upper bound			
Intercept	,406	1,6538	-2,835	3,647	1	,806	44,949
Exp hours	,014	,0208	-,027	,055	1	,502	
Exp compartments	-,374	,4443	-1,245	,497	1	,400	
Exp hours * Exp compartments	,004	,0056	-,007	,015	1	,494	

The model for the skill-error set shows no significant effects(see table 6). Therefore, no effects of experience either in hours of practice or in components on number of skill-related errors were found.

Table 7*: GEE for number of rules problems*

Parameter	B	Standard error	95% Confidence Interval		df	Sig.	QICC
			Lower bound	Upper bound			
Intercept	2,265	,0613	2,145	2,385	1	,000	20,190
Exp hours	-,038	,0301	-,097	,021	1	,211	
Exp compartments	,038	,0364	-,034	,109	1	,298	
Exp hours * Exp compartments	,001	,0057	-,010	,012	1	,891	

The model for the rules-related errors shows no effects for experience of both types, nor an interaction effect between the two types of experience.

Table 8: *GEE for number of knowledge problems*

Parameter	B	Standard error	95% Confidence Interval		df	Sig.	QICC
			Lower bound	Upper bound			
Intercept	,851	,1279	,600	1,102	1	,000	33,219

Exp hours	-,248	,0231	-,293	-,202	1	,000
Exp compartments	,043	,0539	-,063	,148	1	,428
Exp hours * Exp compartments	,051	,0039	,044	,059	1	,000

The model for knowledge errors shows that there is a main effect of experience in hours on number of knowledge errors (see table 8). Exponentiation of the beta-value shows that for each extra single hour of using a pc, laptop, tablet or smart TV per month, the number of knowledge errors would drop by (1-(exp(-0.248)) = 1 – 0.78 = 0.22) 22%. An interaction effect between both types of experience was found here as well, showing that participants with a higher score on experience in hours also had a higher score on experience measured in components; for each extra hour of experience in hours per month, the score on experience in components was raised by (exp(0.051) = 1,05) 5%.

How do elderly people learn to work with an online system?

4. Discussion

4.1 Findings

4.1.1 Learning and learning curves

The results show that while there is great variety in participants' rate of learning, most participants became faster over time using the Care@Home system. One noticeable thing was that time on task in general tended to be highest for the first trial, lowest for the second trial and then (a bit) higher again in the third trial, but still under the time of the first trial. This can be explained by how the trials were set up, as the second trial took place right after the first but the third was always one week apart from the second. The retention span of elderly people is lower than that of younger people (Hawthorn, 2000). It is possible that right after the first trial, they were able to retain all information needed to complete the second trial faster as it followed right after the first with little distractions. Second, participants actively rehearsed the information in the second trial, making it easier to learn and make less mistakes as the information was freshly presented right after the first trial with hardly any distraction. In the time between the second and third trial, the information was not being rehearsed actively and (logically) there were multiple distractions. Therefore, not all information was transferred to the long-term memory (Baddeley & Hitch, 1974, 2010) or information was transferred but did not have strong enough synaptic connections to be retrieved properly (Byrne, 2015), which also coincides to some extent with a decline in encoding information with increasing age, especially detailed information (Balota et al., 2000; McDonough et al., 2014). As the most of the participants in the end were able to learn, the times on task for the third trial were in general lower than that for the first.

While most participants became faster and thus learned over time, it seemed there were also some participants that did not learn how to work with the Care@Home system. These participants stayed on the same level, or even became slower over time. Data showed that these were the oldest participants with no to very little previous experience related to computers. This also related with the small effect of age and the effects of experience in components that were found related to time on task. One possible explanation for these results is that these participants suffered more from effects of fatigue over time: For some participants, the second round and sometimes the third round even had to be shortened (e.g. leave tasks out) because some tasks were too difficult for them (e.g. participants did not want

to complete a task because they found it too difficult in the previous round) or because of time restraints (e.g. because the building would close up or because the next participant was already waiting).

Another explanation is the motivational aspect. Though motivation was not taken into account as a measure, participants seemed to enjoy taking part in the Care@Home trials. Still some said that they would never use a likewise system in their own home because they did not need it and because they felt it was too much of a hassle to learn how to work with it properly. Even though this is just anecdotal, here was seemed to be a lack of intrinsic motivation, which is a strong predictor of learning (Maehr & Meyer, 1997; Washburne, 1936). This factor combined with having no previous computer-related experience at all could inhibit learning to some extent as well, as learning goes faster when participants do have some previous experience, as the new information encoded can then be easily connected to other information already available in the brain in order to create a bigger understanding of how everything relates to one another (Austin et al., 2001). This effect was also shown in multiple studies where elderly learners were compared based on different factors, of which one was previous experience (Charness et al., 2001; Czaja & Sharit, 1993; Laberge & Scialfa, 2005). It may be possible that these participants will be able to learn (as they did already become faster over time in some tasks) to work with Care@Home or a likewise system but at a much slower rate than other participants who have a higher level of motivation and at least some previous experience with computers.

4.1.2 Type of problems

Of all 129 problems found, the majority were thought problems (58) and then sensorimotor (20). These results make sense, as thought errors are errors that are made when a participants developed an incorrect plan or goal to do a task (Barendregt et al., 2006; Zapf et al., 1992). It is comparable with the level of knowledge problems by Rasmussen (1983), on which error mostly occur when users have not worked with a system or likewise system before and have to use their full conscious control in order to carry out a task. While a great amount of users had some experience in the computer-related domain, none but one users had experience with the online functions of a Smart TV. Participants thus had to come up with new plans and goals for execution, and these often failed. Other problems that were found often were sensorimotor problems, which were related to all stadia of orientation (planning,

execution, feedback). This has multiple causes. First, the system was still in a testing phase, so the version used was not yet optimally designed for elderly. Some buttons were really small and color contrasts were not always visible enough, which caused errors. Another source of error was the remote control which could also be used as a keyboard (see appendix I). The buttons on this keyboard were very small and in order to type a capital letter or punctuation mark, another (Fn) button needed to be pressed before typing the wanted letter or punctuation mark. After testing, it was decided that no matter what the new solution would be, this keyboard could not be included in the final version of Care@Home. Second, elderly people overall experience a decline in sight and hearing over the years which might make sensorimotor mistakes more common if the design of the system did not take this into account (Cheng & Lin, 2012; Ivers, Cumming, Mitchell, & Attebo, 1998). Some elderly users also experience trembling, which makes sensorimotor errors more common if the system used small buttons or a remote control that is hard to handle (Cham, Studenski, Perera, & Bohnen, 2008; Seidler et al., 2010; van Dyck et al., 2008). Furthermore, elderly users are slower than younger people when it comes to motor learning in general (Cai et al., 2014; Ketcham & Stelmach, 2004).

The problems least common were omission problems (4), recognition problems (7) and judgment problems (8). Omission and recognition problems both take place on the level of flexible action patterns, comparable to the rules level of Rasmussen where users need some conscious control to perform an action but also use their previous experience as a base of automatism for some parts (Rasmussen, 1983). Mistakes on this level were more likely expected from participants with a lot of previous experience. As most participants did not have a lot of previous experience, mistakes of these types were less common. Another reason that both recognition and judgment problems (the latter being on the knowledge level of conscious control) are less common is because of when they occur: These two types of problems are those that happen after a (sub)task as a result of feedback. The Care@Home system as tested did not include a lot of feedback after actions were performed, so mistakes could hardly be made on this level. Furthermore, some of the users did not finish every task, mostly because of thought errors made earlier on or because they just gave up.

How do elderly people learn to work with an online system?

4.1.3 Most and least severe mistakes

The most severe mistake was caused by a badly labeled submenu, “my neighborhood”, which included movies about local news but also exercise movies. Other menu labels that participants found confusing were address book and agenda, as a lot of participants would use their agenda as an address book as well in real life. Users, especially those who are inexperienced, tend to compare online activities to real-life situations in order to get a better understanding of it (Dadlani, Sinitsyn, Fontijn, & Markopoulos, 2010; Vastenburg, Visser, Vermaas, & Keyson, 2008). When the difference between the system and the everyday surroundings are large, errors will occur.

A lot of problems that were found were based on legends: Participants misread the legend for the agenda and e-mail continuously while a lot of them also thought the legend were clickable buttons for making a new e-mail. The agenda in particular was prone to error, as users had to make a lot of clicks to be able to actually read the agenda or put in a new event. These findings coincide with literature about spatial learning in elderly, which declines over time (Holzinger et al., 2007). They also coincide with findings related to a decline in sight related to contrast (Vastenburg et al., 2008). Other severe mistakes were related to the user going back to the previous screen because s/he was not sure whether the s/he was at the correct screen. This can be explained by the elderly suffering from a decline in spatial map learning in a new environment, combined with confusing labels for the submenu’s and a lack of information within some menu’s (Head & Isom, 2010)

Last, a large cause of error was lack of consistency, which became visible in a task where participants had to change a phone number. While other menu’s let participants click the thing to be changed (e.g. adding an event to a day in the agenda), this menu included a button. Participants tended to click the phone number itself instead. This lack of consistency was also found in some menu’s presenting information and some that did not. A lot of participants did not know how to confirm someone to the address book because there was no instruction on how to do this (click on the person’s photo). Last, mistakes were made because of the remote control and because of errors within the system (e.g. not being able to use the TAB key and not limiting a box’s input to two characters but instead giving an error message because the user entered more than two characters). Summarizing, the most severe mistakes were made based on unclear menu labeling and not taking into account the users day-to-day

surroundings, used screen colors, failure of consistency in the design (e.g. for use of buttons or providing information) and the design of the remote control.

A ranking of all usability problem types in the complete list of 129 usability problems showed that judgment errors mostly occurred in the first half of the list, or the more severe half. This can be explained by the active user paradox: This paradox claims that participants won't take time to read instructions or carefully find out how something works, but try to work with something new right away, even though the former way would save them time in the end (Carroll & Rosson, 1987). It also related to heuristics: if there is a lack of information, users will use rules that worked in situations they think are likewise, but often turn out to be inappropriate for the new situation, producing a bias – or in this case, errors (Kahneman, 2011). Another possibility to take into account is that, as there are only four judgment problems, it is a coincidence that three of them are in the more severe half of the list.

The mistakes that were the easiest to overcome for users differed greatly from the hardest mistakes in that they were more likely mistakes made because a participant wanted to try something beyond the actual task description (trying to get back to the main menu from the exercise video) or because a participant tried to do something the way s/he was used to (e.g. sending a mail to multiple persons at the same time,. These mistakes were more likely to be advanced-user mistakes. It was expected that such problems would occur less, as non-experienced users most likely would not explore the Care@Home system in a likewise way as users with more previous experience. The ranking showed that habit problems in particular are more common at the lower half of the list, representing the less severe usability problems. This could be explained by the fact that habit problems occur at the level of flexible action patterns by Zapf et al. (1992), which can be compared to the rules level by Rasmussen (1983). These mistakes are made based on what participants are used to do in their daily setting. Therefore this type of mistakes is seen as more likely to happen to participants with some previous experience. However, just as for judgment problems, as the number of habit problems was relatively small, it might also be coincidental that most of them are in one half of the list.

4.1.4 Previous experience

Comparing the average previous experience of the participants who experienced the most severe problems to that of the participants who encountered the least severe mistakes

showed no difference, not on experience in hours nor in components. One explanation is that there just was no difference in level of experience. Another plausible explanation is that the way of comparing makes it impossible to find a difference. Average scores were computed by taking the average of all participants who encountered a usability problem, per problem. For the least severe problems, this meant that the scores of only one participant (the one that encountered that problem) was used, which may have flawed the distribution: If some of these problems happened to participants with no previous experience, then that would have an impact on the score in comparison to the average score of the most severe problems (which were always encountered by a multitude of participants and therefore there was a mean based on more than one score).

As no difference between the levels of previous experience could be found between those participants encountering the most severe problems and those encountering the least severe problems, it was chosen to see whether previous experience would have an effect on the number of errors, as more experienced participants were expected to make less mistakes than participants with less experience. As expected, an effect was found, showing that for each single hour of extra previous experience using a pc, laptop, tablet or smart TV, the number of mistakes lowered with 11%. The effect was found only for experience in hours and not for experience as measured in compartments. This contrasts the findings of time on task, where not previous experience in hours but experience in components had a significant lowering effect. One explanation is that for time on task, experience in hours did not matter (as the design of the system was different from what most participants were used to so they had to search anyway) but having any experience with technological devices at all was, as these participants were less familiar with general concepts of the online world, such as e-mailing, which made them need more time to complete tasks. Another possibility is that the questionnaire for experience was self-made and thus not validated in advance, it might be possible that the questionnaire regarding experience in compartments was not right for measuring previous expertise.

Another finding was an interaction effect between the second trial and previous experience in hours, where, compared to in trial 1, previous experience in hours was significantly raised by 1,6% for the second trial. This finding makes sense because as participants were testing the platform in the first trial, this added up to their level of previous

experience as well. No such effects were found for trial 3 compared to 1, which may be explained by the amount of time between trial 1 and 3 (one week) compared to trial 1 and 2 (right after each other) in which users may already have forgotten about parts of the system.

Because this study wanted to find out whether there was a different effect of experience for different types of mistakes as well, it was chosen to run extra tests. These tests were performed on the categories of mistakes as proposed by Rasmussen (1983) and showed that there was a significant effect of hours of previous experience on mistakes of the knowledge type. For each extra hour of previous experience, number of knowledge mistakes lowered with 22%. Again, an interaction effect between the two types of experience was found but it was still very small (5%). The relationship between number of knowledge mistakes and previous experience in hours can be explained, as knowledge mistakes are mistakes on the level of full consciousness (Rasmussen, 1983). This type of mistake is expected when people work with a system for the first time, as they have to develop an understanding of it (Barendregt et al., 2006; Zapf et al., 1992). If users have worked with a likewise system before, they make less mistakes of this type, as they are quicker to learn how to work with it. The relationship between previous experience in hours and the number of knowledge errors also shows that it is to some extent possible for elderly people to lower the number of knowledge errors by practice when using a smart TV.

No effects were found for previous experience on number of mistakes on either the rules or the skills level. For the rules level mistakes, no effect was found from either experience in hours, experience in compartments nor age or any interaction effects. As rules type mistakes are based partly on experience (rules) and partly on conscious control, there were no expectations about what would influence types of mistakes here. Even though no effects were found, it's not certain that this type of mistake is not influenced by a factor. It may be that there are other factors that weren't accounted for in this study. For the skills level mistakes, a model using age as a predictor was used as well, as it was expected that age might have a possible effect on skill level mistakes, as these are often related to motor-skills and those tend to decline over time (Barendregt et al., 2006; Seidler et al., 2010; van Dyck et al., 2008; Zapf et al., 1992). While the model using age as a predictor instead of experience in hours was a better fit, no such effect was found. This might be because there really was no effect. Another possibility is that effects are not completely measurable because the group of

data was very small. For mistakes of the rules and knowledge level, the categories from Zapf et al's model (1992) were added up to create a new variable. For skill problems, no merging of groups was needed as sensorimotor errors are the same as skill problems (Rasmussen, 1983; Zapf et al., 1992). Therefore, the group of skill problems was smaller than the groups of rules and knowledge problems.

4.2 Research implications

The rationale of this study was to find out more about whether and how elderly people learn to work with technical systems, what types of usability problems they encounter while doing so and how experience influences this. While doing so, some interesting implications were found for future research. First of all, the study showed that elderly people in general are able to learn, albeit not all of them and most certainly not at the same rate. It was also found that longitudinal usability testing was very useful to find the mistakes that are not obvious at first sight. There is a tendency to perform usability in merely one round because it is cheaper. However, not all problems are found in this first round and not all problems have the same rate of appearing and disappearing (Jeffrey Rubin & Dana Chisnell, 2008; Kjeldskov et al., 2010; Schnittker, Schmettow, Verhoeven, & Schraagen, 2016). Next to that, fixing problems after product-release is often way more expensive than beforehand (Boehm & Basili, 2001).

In this study, 20 elderly people participated over the course of three trials, challenging the '5 people find 85% of all usability problems' where the claim lies that if multiple small tests are performed with no more than 5 participants, up to 85% of all usability problems can be found (Nielsen, 2000). While this rule is still well-used to this day, there are a lot of studies discussing this view, saying this viewpoint is too simplistic as it won't detect less obvious problems and it is no guarantee that the most severe problems will be in the (up to) 85% detected (Sauro, 2013). Other studies challenged this view of homogeneity as well, together with notes of completeness of observations and independence of trials (James R. Lewis, 2001; Schmettow, 2012). In this study, the point of independency in trials is not taken into account. As all problems are found by the same evaluator, it may be that this influences the chance of finding the same problem in multiple rounds. It would be interesting to see whether this is the case or not, and if not (so, if chances stay the same over multiple trials), what the actual benefit of more than one trial would be. Another interesting point would be to

find out what the best number of a longitudinal design would be, depending on the number of trials and the percentage of usability problems that need to be detected.

The classification guideline as used was a proper way to see the distribution of usability problem per category. One problem was that a number of usability problems could not be classified and the data for these problems was discarded for this study (not for the Care@Home reports). These problems were mostly related to technical issues. Such technical issues are very important and should be among the first problems to be fixed because, if even possible, users have to find a way to work around the technical issue, which tends to cause more usability problems, especially for inexperienced users. Furthermore, it was found that knowledge errors were the most common usability problem, which coincides with the idea of new and relatively technology-inexperienced users trying to work with an online system (Kjeldskov et al., 2010; Rasmussen, 1983). The same results were found in another study with elderly people as well (Zandbergen, 2015).

Elderly people are a user group in itself and their results may differ from other user groups, especially over time. Elderly people tend to learn slower than younger people, which makes it possible that more knowledge problems will be found when testing with elderly participants, as it takes them longer to learn the basics of a system. It may be that testing a technical system with people who are more experienced with technology in general gives a completely other pattern of usability problems, for example more problems in the rule-based domain. This might even be the case with different age groups of elderly, as for example elderly aged 75 and up are different than elderly aged 55-74, not only in levels of technical experience but also in levels of certain memory functioning and therefore rate of learning (Aine et al., 2011; Charness et al., 2001; PewresearchCenter, 2014; Rönnlund et al., 2005). It will be interesting to see the development of these differences, as people overall age healthier, and the people who will be 75 in 20 years from now will probably be more used to technology than the 75 year-olds nowadays.

When designing, it should be taken into account that inexperienced elderly tend to approach an unknown technical system by comparing it to their everyday surroundings (Holzinger et al., 2007). In the current study, discrepancies between the design of Care@Home and such surroundings caused a multitude of usability problems (e.g. in the agenda and address book). Adapting the design to the daily life and everyday surroundings of

elderly does not only lessen the amount of usability problems, it also heightens the level of acceptation(Van Veldhoven et al., 2008; Vastenburg et al., 2008). However, the average user should be kept in mind as well: If the expected user is one with a lot of technical experience, making the design too easy might not work at all.

Last, this study found that there was a link between knowledge problems and previous experience. No such link was found for skill problems or rules problems. While this may also be because of other reasons (small dataset for skill problems, experience questionnaire not validated in advance), this may be something usability researchers and developers need to take into account: If skill and rules problems are not impacted by previous experience, these are the problems that won't lessen or go away over time by extra expertise, or learning how to work with the system. It is therefore needed to find out if other factors impact these types of problems and whether these can be trained.

4.3 Study limitations

In this study, a number of limitations were encountered. First of all, the questionnaire about experience was not tested for validity and reliability due to time constraints of the Care@Home project. Therefore, it might be possible that the questionnaire did not represent true experience as this study intended or something else. Second, there was no interrater reliability for the incidents nor the incident matching, as this was done by one person only. The only slight form of interrater-reliability available were the notes from two other researchers who were alternately helping during the usability sessions. The rating of these problems were also done by one person, which makes it possible that some problems are not seen (Boehm & Basili, 2001) and that others are not ranked properly. This is also called the evaluator effect (Jacobsen, Hertzum, & Bonnie, 1998). During the sessions there were some drawbacks as well. First of all, the think-aloud method was used, which is a strong method but has a drawback of being extra affirmative to finding problems a usability tester has a suspicion of in advance (Nørgaard & Hornbæk, 2006). Furthermore, a concurrent think aloud (CTA) protocol was used instead of a retrospective one (RTA). While it was chosen to do so for proper reasons, CTA has a drawback of interrupting task performance, which was measured here as time on task and number of mistakes made (van den Haak et al., 2003). During classification, 9 out of 138 usability problems needed to be discarded because they could not be classified. These were mostly usability problems related to technical failures, but

were still encountered a lot by the participants. Classification itself was done by first grouping the incidents together into usability problems and then classification. While this is not wrong in itself, it is done the other way around in a different study using this classification guideline, making it possible to compare the number of incidents and the number of usability problems of one classification category (Zandbergen, 2015). Last, only previous experience and age were used as predictors in the GEE models. As no results were found for the error categories of skills and rules problems, it may be that there are other factors missing that were not measured, such as motivation or different age classes. While this was not done in this study due to time constraints (length of trials could not be too long, so no extra questionnaires) and small group of participants (division in different age groups would make the groups too small), these and other factors may account for learning differences as well. Finally, a Poisson distribution was used for the set of skill-related problems while there was underdispersion. While this would normally have to be accounted for in a different way, such as a COM-Poisson distribution, it was chosen not to do so as the underdispersion was very small and it was needed to take into account effects of repeated measures, fatigue and learning as well which could not be done with the COM-Poisson distribution (Kokonendji, 2014; Shmueli, Minka, Kadane, Borle, & Boatwright, 2005).

How do elderly people learn to work with an online system?

5. Future research

While this study found out that the majority of elderly participants was able to learn how to work with an online system over time, precise predictors of learning were not examined. As the elderly population will grow steadily in the upcoming years, more technical devices will be developed as well. For this reason, it might be interesting to address which factors are responsible for elderly learning and how these factors can be used to develop a system that is not only usable for the majority of elderly but also for the small group of elderly that has trouble learning.

It was found that previous experience as measured in this study only influenced the number of knowledge problems encountered. Not only is it a good idea to create a validated and reliable questionnaire regarding previous experience and use this for retesting these findings, it may be that other, possibly age-related, factors influence the number of skill or rules problems as well. Finding such factors makes it possible to predict how and in which amount problems may develop for different user groups: from this study, it seems that people with a lot of previous experience encounter a smaller amount of knowledge problems. For finding other variables, it may not only be needed to look for completely new things. As mentioned before, one possibility is to renew the experience questionnaire, for example by making the type of previous experience more defined. Another factor that could be re-examined is the age group, as younger elderly (60-74) may learn in different ways than older elderly (75+). This study was not able to define such age groups because the user sample was too small to do so. This brings up another point, as this study used a longitudinal design. First of all, the influence of sample size on a longitudinal design is, which makes it a good subject for future research. Second, longitudinal research made it possible to rank problem severity not only by frequency and impact, but by persistence as well. Future research may take into account the importance of persistent, as a lot of usability research currently tends to overlook it, while a highly persistent problem may have an influence on the willingness of users to work with a system at large.

Finally, while the classification system as defined in this study proved quite helpful for classifying usability problems, it was far from perfect as not all usability problems could be defined and it was not determined whether one should classify incidents first and then merge

them into usability problems or merge incidents and classify the usability problems coming from that. Future research could focus on implementing possible new categories, or better ways on how to deal with those problems or incidents that cannot be classified in the current version.

How do elderly people learn to work with an online system?

6. Conclusions

The goal of this study was to find out if elderly people are able to learn how to work with a technical system, how to learn, which are the most and least severe usability problems they encounter while doing so and what the role of experience is for these problems. It was found that while nearly all elderly users got faster over time (and thus, were able to learn), some users did not get faster or got even slower. It seems thus that not all elderly users are immediately able to learn how to work with a technical system, even though the exact underlying factors of this are unclear at this point. The learning curves over time showed that elderly participants in general needed the most time for trial 1, then became faster in trial 2 and then slowed down a bit in trial 3, but still completed trial 3 faster than trial 1. These findings are probably related to how the trials were set up, with the first two in a row and the last trial one week later.

For finding the most and least severe problems encountered, this study did not only use problems frequency and impact, but persistency as well to rank usability problems in order of severity. It was shown that the most severe problems often based on menu's with an unclear or ambiguous labeling, severe inconsistencies in the design of the system, design that did not take the users everyday surroundings into account, poor use of screen color and contrasts and finally, the use of a remote control. The majority of the recommendations for the Care@Home system were based on improving these factors, as an improved, more consistent design with better use of colors and contrasts would solve a lot of other, less severe usability problems as well (see appendix B). It was decided that the remote control would not be used in a new design as it was impossible to improve it (it could not be customized as it came with the Smart TV), instead it was opted to maybe use a tablet instead. The least severe usability problems encountered where mostly related to users with more experience who were trying to do things beyond the task description, often based on how they were used to do things for themselves at home. While these problems were not the most severe, it was a good reminder to keep the entire spectrum of the user group in mind and not only design for the inexperienced users but also provide extra's for those users that were more experienced.

While there did not seem to be a difference between the experience levels of those who experienced the most severe problems compared to those who experienced the least

severe problems, previous experience (in hours) did have an influence on the total number of problems encountered. It was also found that previous experience in components had an effect on time on task. This may indicate that for task completion time, it matters whether someone has had any previous experience with technology at all or not while for the number of errors encountered the amount of time spent using technological devices matters more.

Splitting the problems into different categories based on Rasmussen (Rasmussen, 1983) showed that this effect of previous experience was related to knowledge type of problems only and not to rules or skill-related problems. This makes sense, as knowledge problems are related to users who are new to a system, and most participants in this study were relatively inexperienced with technical devices (Rasmussen, 1983; Zapf et al., 1992). One could say that if someone knows in advance that the target user group is relatively inexperienced, not only will they encounter this type of problems more often, they will be able to lessen the number of problems by practice. As there was no connection between previous experience and the other two types of problems, practice will not lower the numbers of those problems encountered, so researchers might want to take that into account as well when they rank problems according to severity. However, it may very well be possible that there are other factors having a likewise influence on rules and skills problems. If these are factors can be found, a fuller understanding of elderly learning and the elderly as a user group can be reached.

How do elderly people learn to work with an online system?

References

Aine, C. J., Sanfratello, L., Adair, J. C., Knoefel, J. E., Caprihan, A., & Stephen, J. M. (2011). Development and decline of memory functions in normal, pathological and healthy successful aging. *Brain Topography*, *24*(3-4), 323–339. doi:10.1007/s10548-011-0178-x

Austin, K., Orcutt, S., & Rosso, J. (2001). How people learn: Introduction to learning theories. In L.-D. Hammond (Ed.), *The learning classroom; Theory into practice - a telecourse for teacher education and professional development* (pp. 1–22). Stanford, California: Stanford University School of Education. Retrieved from http://www.seas.upenn.edu/~eas285/Readings/Hammond_HowPeopleLearn.pdf

Baddeley, A. (2001). The concept of episodic memory. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *356*(1413), 1345–1350. doi:10.1098/rstb.2001.0957

Baddeley, A., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (pp. 47 – 89). Academic press. doi:10.1016/S0079-7421(08)60452-1

Baddeley, A., & Hitch, G. J. (2010). Working memory. *Scholarpedia*, *5*(2), 3015. doi:10.4249/scholarpedia.3015

Balota, D. a, Dolan, P. O., & Duchek, J. M. (2000). Memory changes in healthy young and older adults. *The Oxford Handbook of Memory*, *The Oxford*, 395–409. Retrieved from http://www.psych.wustl.edu/coglab/publications/BalotaDolanDuchekMemchapter2000.pdf

Barendregt, W., Bekker, M. M., Bouwhuis, D. G., & Baauw, E. (2006). Identifying usability and fun problems in a computer game during first use and after some practice. *Int. J. Hum.-Comput. Stud.*, *64*(9), 830–846. doi:10.1016/j.ijhcs.2006.03.004

Bauer, E., Toepper, M., Gebhardt, H., Gallhofer, B., & Sammer, G. (2015). The significance of caudate volume for age-related associative memory decline. *Brain Research*, *1622*, 137–148. doi:10.1016/j.brainres.2015.06.026

Boehm, B. W., & Basili, V. R. (2001). Software Defect Reduction Top 10 list. *IEEE*

Computer, *34*(1), 135–137.

Broadbent, E., Kuo, I. H., Lee, Y. I., Rabindran, J., Kerse, N., Stafford, R., & MacDonald, B. a. (2010). Attitudes and reactions to a healthcare robot. *Telemedicine Journal and E-Health : The Official Journal of the American Telemedicine Association*, *16*(5), 608–13. doi:10.1089/tmj.2009.0171

Byrne, J. H. (2015). Neuroscience online. *Chapter 7: Learning and Memory*. Retrieved February 27, 2016, from http://neuroscience.uth.tmc.edu/s4/chapter07.html

Cacioppo, J. T., Hawkley, L. C., Crawford, L. E., Ernst, J. M., Burleson, M. H., Kowalewski, R. B., … Berntson, G. G. (2002). Loneliness and health: potential mechanisms. *Psychosomatic Medicine*, *64*(3), 407–417.

Cai, L., Chan, J. S. Y., Yan, J. H., & Peng, K. (2014). Brain plasticity and motor practice in cognitive aging. *Frontiers in Aging Neuroscience*, *6*, 1–12. doi:10.3389/fnagi.2014.00031

Campen, C. (2011). Kwetsbare ouderen. *Sociaal Cultureel Planbureau*, 218. Retrieved from http://library.wur.nl/WebQuery/clc/1956247

Cansino, S., Hernández-Ramos, E., Estrada-Manilla, C., Torres-Trejo, F., Martínez-Galindo, J. G., Ayala-Hernández, M., … Rodríguez-Ortiz, M. D. (2013). The decline of verbal and visuospatial working memory across the adult life span. *Age*, *35*(6), 2283–2302. doi:10.1007/s11357-013-9531-1

Caplin, M. (1969). Resistance to learning. *Peabody Journal of Education*, *47*(1), Atherton J S. doi:10.1080/01619566909537673

Carroll, J. M., & Rosson, M. B. (1987). Paradox of the active user. In J. M. Carrol (Ed.), *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction* (pp. 80–111). Cambridge, MA: MIT Press.

Centraal Bureau voor de Statistiek. (2012). Ouderen beginnen pas op latere leeftijd te vereenzamen. Retrieved May 20, 2015, from http://www.cbs.nl/nl-NL/menu/themas/dossiers/levensloop/publicaties/artikelen/archief/2012/2012-ouderen-vereenzaming-dns-pub.htm

Cham, R., Studenski, S. A., Perera, S., & Bohnen, N. I. (2008). Striatal dopaminergic denervation and gait in healthy adults. *Experimental Brain Research*, *185*(3), 391–398. doi:10.1007/s00221-007-1161-3

Charness, N., Kelley, C. L., Bosman, E. a, & Mottram, M. (2001). Word-processing training and retraining: effects of adult age, experience, and interface. *Psychology and Aging*, *16*(1), 110–127. doi:10.1037/0882-7974.16.1.110

Cheng, C.-H., & Lin, Y.-Y. (2012). The effects of aging on lifetime of auditory sensory memory in humans. *Biological Psychology*, *89*(2), 306–12. doi:10.1016/j.biopsycho.2011.11.003

Chou, W. H., Lai, Y.-T., & Liu, K.-H. (2013). User requirements of social media for the elderly: a case study in Taiwan. *Behaviour & Information Technology*, *32*(9), 920–937. doi:10.1080/0144929X.2012.681068

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*, 73–101. doi:10.1146/annurev.psych.093008.100427

Craik, F. I. M. (1994). Memory changes in normal aging. *Curr. Dir. Psychol. Sci.*, *3*(5), 155–158.

Czaja, S. J., & Sharit, J. (1993). Age differences in the performance of computer-based work. *Psychology and Aging*, *8*(1), 59–67. doi:10.1037/0882-7974.8.1.59

Dadlani, P., Sinitsyn, A., Fontijn, W., & Markopoulos, P. (2010). Aurama: Caregiver awareness for living independently with an augmented picture frame display. *AI and Society*, *25*, 233–245. doi:10.1007/s00146-009-0253-y

Dumas, J. S., & Redish, J. C. (1999). Tabulating and analyzing data. In *A practical guide to usability testing* (pp. 309 – 330). Exeter, UK: Intellect Books. Retrieved from https://books.google.nl/books?id=4lge5k_F9EwC&pg=PA325&lpg=PA325&dq=dumas+and+redish+severity+ranking&source=bl&ots=vqgd7Eb6uG&sig=3fkxaHAc6yIdnUMkGwuRzIr4IPY&hl=nl&sa=X&ved=0CDAQ6AEwAmoVChMInv7nqp6IyQIVQXMPCh0iUwD-#v=onepage&q=dumas and redish severity r

Duncker, K., & Lees, L. S. (1945). On problem-solving. *Psychological Monographs*, *58*(5), i–113.

Eisenberger, R., Rhoades, L., & Cameron, J. (1999). Does pay for performance increase or decrease perceived self-determination and intrinsic motivation? *Journal of Personality and Social Psychology*, *77*(5), 1026–1040. doi:10.1037/0022-3514.77.5.1026

Eppinger, B., Herbert, M., & Kray, J. (2010). We remember the good things: Age differences in learning and memory. *Neurobiology of Learning and Memory*, *93*(4), 515–521. doi:10.1016/j.nlm.2010.01.009

Fleischman, D. A. (2007). Repetition priming in aging and Alzheimer's disease: An integrative review and future directions. *Cortex*, *43*(7), 889–897. doi:10.1016/S0010-9452(08)70688-9

Følstad, A., Law, E. L., & Hornbæk, K. (2012). Outliers in usability testing : How to treat usability problems found for only one test participant? In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design* (pp. 257–260).

Friedman, D., Nessler, D., & Johnson, R. (2007). Memory encoding and retrieval in the aging brain. *Clinical EEG and Neuroscience : Official Journal of the EEG and Clinical Neuroscience Society*, *38*(1), 2–7. doi:10.1177/155005940703800105

Gottlieb, J. (2012). Attention, Learning, and the Value of Information. *Neuron*, *76*(2), 281–295. doi:10.1016/j.neuron.2012.09.034

Haar, A., Schmettow, M., & Schraagen, J. M. . (2013). *Developing of a Qualitative Classification Method for Usability Errors after Rasmussen*. University of Twente.

Hawthorn, D. (2000). Possible implications of aging for interface designers. *Interacting with Computers*, *12*(5), 507–528. doi:10.1016/S0953-5438(99)00021-1

Head, D., & Isom, M. (2010). Age effects on wayfinding and route learning skills. *Behavioural Brain Research*, *209*(1), 49–58. doi:10.1016/j.bbr.2010.01.012

Holzinger, A., Searle, G., & Nischelwitzer, A. (2007). On some aspects of improving mobile

applications for the elderly. In C. Stephanidis (Ed.), *Universal Access in Human Computer Interaction: Coping with Diversity, Pt 1* (Vol. 4554, pp. 923–932). Berlin: Springer-Verlag Berlin. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-73279-2_103

Hornbæk, K., & Frøkjær, E. (2008). Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers*, *20*(6), 505–514. doi:10.1016/j.intcom.2008.08.005

Hurtienne, J., Horn, A.-M., Langdon, P. M., & Clarkson, P. J. (2013). Facets of prior experience and the effectiveness of inclusive design. *Universal Access in the Information Society*, *12*(3), 297–308. doi:10.1007/s10209-013-0296-1

Ivers, R. Q., Cumming, R. G., Mitchell, P., & Attebo, K. (1998). Visual Impairment and Falls in Older Adults: The Blue Mountains Eye Study. *Journal of the American Geriatric Society*, *46*(1), 58–64. doi:10.1111/j.1532-5415.1998.tb01014.x

Jacobsen, N. E., Hertzum, M., & Bonnie, J. (1998). The evaluator effect in usability studies: problem detection and severity judgments. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, *42*(9), 1336–1340. doi:10.1177/154193129804201902

Jeffrey Rubin, & Dana Chisnell. (2008). *Handbook of Usability Testing: Howto Plan, Design, and Conduct Effective Tests* (2nd ed.). John Wiley & Sons.

Jones, B. D., & Bayen, U. J. (1998). Teaching Older Adults To Use Computers: Recommendations Based on Cognitive Aging Research. *Educational Gerontology*, *24*(7), 675–689. doi:10.1080/0360127980240705

Kahneman, D. (2011). *Thinking, fast and slow*. New york: Farrar, Straus and Giroux.

Kazdin, A. E. (Ed.). (2000). *Encyclopedia of Psychology: 8 Volume Set*. American Psychological Association.

Ketcham, C. J., & Stelmach, G. E. (2004). Movement control in the older adult. In R. W. Pew & S. B. Van Hemel (Eds.), *Technology for adaptive aging* (pp. 64–92). Washington, D. C.: The National Academies Press.

Kjeldskov, J., Skov, M. B., & Stage, J. (2005). Does time heal?: a longitudinal study of usability. *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*. Canberra, Australia: Computer-Human Interaction Special Interest Group (CHISIG) of Australia.

Kjeldskov, J., Skov, M. B., & Stage, J. (2010). A longitudinal study of usability in health care: does time heal? *International Journal of Medical Informatics*, *79*(6), e135–43. doi:10.1016/j.ijmedinf.2008.07.008

Klencklen, G., Després, O., Dufour, A., & Despres, O. (2012). What do we know about aging and spatial cognition? Reviews and perspectives. *Ageing Research Reviews*, *11*(1), 123–135. doi:10.1016/j.arr.2011.10.001

Kokonendji, C. C. (2014). Over- and Underdispersion Models. In N. Balakrishnan (Ed.), *Methods and Applications of Statistics in Clinical Trials: Planning, Analysis and Inferential Methods* (2nd ed., pp. 506–526). John Wiley & Sons, Inc.

Kruschke, J. K. (2011). Models of Attentional Learning. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 120–152). Cambridge university press.

Kürten, J., De Vries, M. H., Kowal, K., Zwitserlood, P., & Flöel, A. (2012). Age affects chunk-based, but not rule-based learning in artificial grammar acquisition. *Neurobiology of Aging*, *33*(7), 1311–7. doi:10.1016/j.neurobiolaging.2010.10.008

Laberge, J. C., & Scialfa, C. T. (2005). Predictors of Web navigation performance in a life span sample of adults. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(2), 289–302. doi:10.1518/0018720054679470

Larson, R. W. (2000). Toward a psychology of positive youth development. *American Psychologist*, *55*(I), 170–183. doi:10.1037//0003-066X

Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, *16*(4-5), 246–266. doi:10.1080/014492997119824

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer*

Interaction, *7*(1), 57 – 78. doi:10.1080/10447319509526110

Lewis, J. R. (2001). Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated From Small Samples. *International Journal of Human-Computer Interaction*, *13*(4), 445–479. doi:10.1207/S15327590IJHC1304_06

Lin, Y. G., McKeachie, W. J., & Kim, Y. C. (2001). College student intrinsic and/or extrinsic motivation and learning. *Learning and Individual Differences*, *13*(3), 251–258. doi:10.1016/S1041-6080(02)00092-4

Lutz, W., Sanderson, W., & Scherbov, S. (2008). The coming acceleration of global population ageing. *Nature*, *451*(7179), 716–719. doi:10.1038/nature06516

Maehr, M. L., & Meyer, H. a. (1997). Understanding motivation and schooling: Where we've been, where we are, and where we need to go. *Educational Psychology Review*, *9*(4), 371–409. doi:http://dx.doi.org.library.capella.edu/10.1023/A:1024750807365

Martinez, R., Cabecinhas, R., Loscertales, F., Loscertales Abril, F., & Martínez Pecino, R. (2011). University Senior Students on the Web. *Comunicar*, *19*(37), 89–95. doi:10.3916/c37-2011-02-09

McDonough, I. M., Cervantes, S. N., Gray, S. J., & Gallo, D. A. (2014). Memory's aging echo: Age-related decline in neural reactivation of perceptual details during recollection. *NeuroImage*, *98*, 346–358. doi:10.1016/j.neuroimage.2014.05.012

Meadmore, K. L., Dror, I. E., & Bucks, R. S. (2009). Lateralisation of spatial processing and age. *Laterality*, *14*(1), 17–29. doi:10.1080/13576500802022265

Nielsen, J. (1993). *Usability engineering*. San Fransisco, CA: Morgan Kaufmann Publishers Inc.

Nielsen, J. (1995). Severity Ratings for Usability Problems. *Papers and Essays*, *54*. Retrieved from http://www.useit.com/papers/heuristic/severityrating.html

Nielsen, J. (2000, March). Why you only need to test with 5 users. Retrieved April 11, 2016, from http://www.useit.com/alertbox/20000319.html

Nørgaard, M., & Hornbæk, K. (2006). What Do Usability Evaluators Do in Practice ? An

Explorative Study of Think-Aloud Testing. In *ACM Conference on Designing Interactive Systems* (pp. 209–218). Pennsylvania, USA. Retrieved from http://www.kasperhornbaek.dk/papers/DIS2006_UsabilityEvaluation.pdf

Ormrod, J. E. (2013). Motivation and affect. In *Educational Psychology: Developing Learners* (pp. 384–386). Pearson Education.

PewresearchCenter. (2014). Older Adults and Technology Use, 1 – 27. Retrieved from http://www.pewinternet.org/files/2014/04/PIP_Seniors-and-Tech-Use_040314.pdf

Pinquart, M., & Sörensen, S. (2001). Influences on Loneliness in Older Adults : A Meta-Analysis. *Basic and Applied Social Psychology*, *23*(4), 245–266. doi:10.1207/S15324834BASP2304

Pugh, K. J., & Bergin, D. A. (2006). Motivational Influences on Transfer. *Educational Psychologist*, *41*(3), 161–180. doi:10.1207/s15326985ep4103_2

Purdie, N., & Boulton-Lewis, G. (2003). The learning needs of older adults. *Educational Gerontology*. doi:10.1080/713844281

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *Systems, Man and Cybernetics, IEEE Transactions on*, *SMC-13*(3), 257–266. doi:10.1109/tsmc.1983.6313160

Reason, J. (1990). *Human error*. Cambridge university press. Retrieved from http://books.google.com/books?hl=nl&lr=&id=WJL8NZc8lZ8C&pgis=1

Ricker, J. (2012). Chapter 5: Remembering and forgetting. *PSY 101 – Introduction to Psychology*. Retrieved February 10, 2016, from http://sccpsy101.com/home/chapter-5/

Rodrigues, P. F. S., & Pandeirada, J. N. S. (2015). Attention and working memory in elderly: the influence of a distracting environment. *Cognitive Processing*, *16*(1), 97–109. doi:10.1007/s10339-014-0628-y

Rönnlund, M., Nyberg, L., & Bäckman, L. (2005). Stability, Growth, and Decline in Adult Life Span Development of Declarative Memory: Cross-Sectional and Longitudinal Data From a Population-Based Study. *Psychology and Aging*, *20*(1), 3–18. doi:10.1037/0882-

7974.20.1.3

Rubin, J., & Chisnell, D. (2008). Developing the Test Plan. In 2nd (Ed.), *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests* (pp. 65–92). New York, NY, USA: John Wiley & Sons, Inc.

Ryan, R., & Deci, E. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, *25*(1), 54–67. doi:10.1006/ceps.1999.1020

Salat, D. H., Buckner, R. L., Snyder, A. Z., Greve, D. N., Desikan, R. S. R., Busa, E., … Fischl, B. (2004). Thinning of the cerebral cortex in aging. *Cerebral Cortex*, *14*(7), 721–730. doi:10.1093/cercor/bhh032

Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological Psychology*, *54*(1-3), 35–54. doi:10.1016/S0301-0511(00)00052-1

Sauce, B., Wass, C., Smith, A., Kwan, S., & Matzel, L. D. (2014). The external-internal loop of interference: Two types of attention and their influence on the learning abilities of mice. *Neurobiology of Learning and Memory*, *116*, 181–192. doi:10.1016/j.nlm.2014.10.005

Sauro, J. (2013). 5 Reasons You Should And Should Not Test With 5 Users. *Measuring U*. Retrieved May 22, 2016, from http://www.measuringu.com/blog/five-for-five.php

Schmettow, M. (2012). Sample size in usability studies. *Communications of the ACM*, *55*(4), 64. doi:10.1145/2133806.2133824

Schnittker, R., Schmettow, M., Verhoeven, F., & Schraagen, J. M. C. (2016). Combining situated Cognitive Engineering with a novel testing method in a case study comparing two infusion pump interfaces. *Applied Ergonomics*, *55*, 16–26. doi:10.1016/j.apergo.2016.01.004

Schugens, M. M., Daum, I., & Spindler, M. (1997). Differential effects of aging on explicit and implicit memory. *Aging, Neuropsychology and Cognition*, *4*(1), 37–41. doi:10.1080/13825589708256634

Seidler, R. D., Bernard, J. A., Burutolu, T. B., Fling, B. W., Gordon, M. T., Gwin, J. T., … Lipps, D. B. (2010). Motor control and aging: Links to age-related brain structural, functional, and biochemical effects. *Neuroscience and Biobehavioral Reviews*, *34*(5), 721–733. doi:10.1016/j.neubiorev.2009.10.005

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *54*(1), 127–142. doi:10.1111/j.1467-9876.2005.00474.x

Spaan, P. E. J., & Raaijmakers, J. G. W. (2011). Priming Effects from Young-Old to Very Old Age on a Word-Stem Completion Task: Minimizing Explicit Contamination. *Aging, Neuropsychology, and Cognition*, *18*(1), 86–107. doi:10.1080/13825585.2010.511146

Stijacic Cenzer, I. (2012). Loneliness in Older Persons - A Predictor of Functional Decline and Death. *Archives of Internal Medicine*, *172*(14), 1078. doi:10.1001/archinternmed.2012.1993

van den Haak, M., Jong, M. De, & Schellens, P. J. (2003). Retrospective vs . concurrent think-aloud protocols : testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339 – 351. doi:10.1080/0044929031000

Van der Linden, M., Bredart, S., & Beerten, A. (1994). Age-related differences in updating working memory. *British Journal of Psychology*, *85*, 145–152. doi:10.1111/j.2044-8295.1994.tb02514.x

van Dyck, C. H., Avery, R. A., MacAvoy, M. G., Marek, K. L., Quinlan, D. M., Baldwin, R. M., … Arnsten, A. F. T. (2008). Striatal dopamine transporters correlate with simple reaction time in elderly subjects. *Neurobiology of Aging*, *29*(8), 1237–46. doi:10.1016/j.neurobiolaging.2007.02.012

Van Veldhoven, E. R., Vastenburg, M. H., Keyson, D. V., Veldhoven, E. R. Van, Vastenburg, M. H., & Keyson, D. V. (2008). Designing an interactive messaging and reminder display for elderly. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5355 LNCS*, 126–140. doi:10.1007/978-3-540-89617-3-9

Vastenburg, M. M. M., Visser, T., Vermaas, M., & Keyson, D. D. (2008). Designing Acceptable Assisted Living Services for Elderly Users. In E. Aarts, J. L. Crowley, B. Ruyter, H. Gerhäuser, A. Pflaum, J. Schmidt, & R. Wichert (Eds.), *Ambient Intelligence: lecture notes in computer science* (pp. 1–12). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-89617-3

Voelcker-Rehage, C. (2008). Motor-skill learning in older adults-a review of studies on age-related differences. *European Review of Aging and Physical Activity*, *5*(1), 5–16. doi:10.1007/s11556-008-0030-9

Ward, E. V., Berry, C. J., & Shanks, D. R. (2013). Age effects on explicit and implicit memory. *Frontiers in Psychology*, *4*, 1–11. doi:10.3389/fpsyg.2013.00639

Washburne, J. N. (1936). The definition of learning. *Journal of Educational Psychology*, *27*(8), 603–611. doi:10.1037/h0060154

White, H., McConnell, E., Clipp, E., Bynum, L., Teague, C., Navas, L., … Halbrecht, H. (1999). Surfing the Net in Later Life: A Review of the Literature and Pilot Study of Computer Use and Quality of Life. *Journal of Applied Gerontology*, *18*(3), 358–378. doi:10.1177/073346489901800306

Wisniewski, Z., & Polak-sopinska, A. (2009). HCI Standards for Handicapped. *Access in Human-Computer Interaction. Addressing Diversity. Part I: Held as Part of HCI International 2009*, *5614*, 672–676. doi:10.1007/978-3-642-02707-9

World Health Organization. (2014). Facts about Ageing. *Ageing and Lifecourse*. Retrieved September 1, 2015, from http://www.who.int/ageing/about/facts/en/

Zandbergen, R. (2015). *Predicting persistency of usability problems based on error classification*. University of Twente.

Zantinge, E. (RIVM). (2014). Hoeveel mensen zijn eenzaam? *Volksgezondheid Toekomst Verkenning, Nationaal Kompas Volksgezondheid.* Bilthoven: RIVM. Retrieved November 20, 2015, from http://www.nationaalkompas.nl/gezondheid-en-ziekte/functioneren-en-kwaliteit-van-leven/eenzaamheid/hoeveel-mensen-zijn-eenzaam/

Zapf, D., Brodbeck, F. C., Frese, M., Peters, H., & Prümper, J. (1992). Errors in working with

office computers: A first validation of a taxonomy for observed errors in a fiedl setting. *International Journal of Human-Computer Interaction*, *4*(4), 311–339.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*, 3–14. doi:10.1111/j.2041-210X.2009.00001.x

Appendix A – Questionnaires

All questionnaires used were in Dutch, as this was the native language of all participants. Questionnaires were printed in font size 16 for better readability.

Informed consent

Ik verklaar hierbij dat het voor mij duidelijk is wat het doel en de methode is van dit onderzoek. Hier ben ik door de onderzoeker door voorgelicht, en eventuele vragen die ik had zijn naar tevredenheid beantwoord.

Ik stem geheel in met deelname aan dit onderzoek. Daarbij ben ik me er van bewust dat ik het recht heb om tijdens het onderzoek mijn deelname in te trekken zonder daar een reden voor hoeven op te geven.

Als mijn onderzoeksresultaten gebruikt worden in wetenschappelijke publicaties of op een andere manier, zal dit volledig geanonimiseerd gebeuren. Mijn persoonsgegevens zullen niet door derden worden ingezien zonder mijn uitdrukkelijke toestemming.

Als ik in de verdere toekomst nog meer informatie wil over dit onderzoek kan ik mij wenden tot Ouderenfonds Nederland. Al dus getekend op …. / …. / 201.. door:
_____.

Naam proefpersoon: Naam onderzoeker:

_____ _____

Ik, als onderzoeker, verklaar hierbij dat ik toelichting heb gegeven op het onderzoek dat volgt. Ik verklaar hierbij dat ik vragen die de proefpersoon nog heeft over het onderzoek naar vermogen zal proberen te beantwoorden:

Handtekening proefpersoon: Handtekening onderzoeker:

_____ _____

How do elderly people learn to work with an online system?

Demographics

1) Wat is uw naam?

_____

2) Wat is uw leeftijd (in jaren)?

_____

3) Bent u een man of een vrouw? (omcirkel wat van toepassing is)
   Man / Vrouw

4) Wat is uw nationaliteit?
   - Nederlands
   - Anders, namelijk _____

5) Heeft u al eens eerder meegewerkt aan een onderzoek van Ouderenfonds Nederland?
   Ja / Nee

Experience

De volgende vragen gaan over uw ervaring met Smart TV's, computers of tablets. Omcirkel of kruis het rondje aan met het antwoord wat voor u het meest van toepassing is.

1) Ik ben in het bezit van een Smart TV
   Ja / Nee

2) Ik heb wel eens gewerkt met een Smart TV (dit kan ook bij andere mensen thuis zijn, bijvoorbeeld bij familie of vrienden).
   Ja / Nee

3) Ik ben in het bezit van een computer, laptop of tablet.
   Ja / Nee

How do elderly people learn to work with an online system?

4) Ik heb wel eens gewerkt met een computer, laptop of tablet.
   Ja / Nee

   Als u hierboven ‘Ja’ heeft ingevuld, gelieve dan de volgende 2 vragen ook in te vullen. Als u ‘Nee’ heeft geantwoord, ga door naar vraag 5.

   a) Waar heeft u gewerkt met een computer, laptop of tablet? (meerdere antwoorden zijn mogelijk)
      O - In mijn huis
      O - Bij familie thuis
      O - Bij vrienden thuis
      O - Op het werk
      O - Anders, namelijk: _____

   b) Als u op het werk heeft gewerkt met een computer, laptop of tablet, gebruikte u deze dan voor het grootste deel van uw werkzaamheden?

      O - Ik gebruikte geen computer op mijn werk.
      O - Ik gebruikte de computer voor een klein deel van al mijn werkzaamheden.
      O - Ik gebruikte de computer voor ongeveer de helft van al mijn werkzaamheden
      O - Ik gebruikte de computer voor meer dan de helft van al mijn werkzaamheden

5) Heeft uw wel eens gebruik gemaakt van het Internet?
   Ja / Nee

6) Heeft u wel eens een e-mail naar iemand verstuurd?
   Ja / Nee

Hierna volgen een aantal vragen over uw gemiddelde gebruik van Smart TV’s, computers en tablets. Vul hier een getal in of kruis het antwoord aan wat het beste bij u past.

How do elderly people learn to work with an online system?

7) Hoeveel uren per week werkt u gemiddeld met een Smart TV?
   (denk hierbij aan het aantal uren per week in de afgelopen maand).

   _____

   Als u nog nooit met een smart TV hebt gewerkt, zet dan een 0

8) Hoeveel ervaring zou u zelf zeggen dat u heeft met een Smart TV?
   O - Ik heb geen ervaring met een Smart TV
   O - Ik heb een beetje ervaring met een Smart TV
   O - Ik heb een gemiddelde ervaring met een Smart TV
   O - Ik heb een bovengemiddelde ervaring met een Smart TV
   O - Ik ben zeer ervaren met een Smart TV

9) Hoeveel uren per week werkt u gemiddeld met een computer, laptop of tablet? (denk hierbij aan het aantal uren per week in de afgelopen maand).

   _____

   Als u nog nooit met een computer, laptop of tablet hebt gewerkt, zet dan een 0.

10) Hoeveel ervaring zou u zelf zeggen dat u heeft met een computer, laptop of tablet?
   O - Ik heb geen ervaring met een computer, laptop of tablet
   O - Ik heb een beetje ervaring met een computer, laptop of tablet
   O - Ik heb een gemiddelde ervaring met een computer, laptop of tablet
   O - Ik heb een bovengemiddelde ervaring met een computer, laptop of tablet
   O - Ik ben zeer ervaren met een computer, laptop of tablet

ASQ Questionnaire

Over het geheel genomen ben ik tevreden met het gemak waarmee ik de taak kan voltooien.

Helemaal mee oneens								*Helemaal mee eens*

How do elderly people learn to work with an online system?

Over het geheel genomen ben ik tevreden met de tijd die het me gekost heeft om de taak te voltooien.

Helemaal mee oneens								*Helemaal mee eens*

Over het geheel genomen ben ik tevreden met de ondersteunende informatie (help-functie, berichten op het scherm en andere documentatie) die ik kreeg om de taak te voltooien.

Helemaal mee oneens								*Helemaal mee eens*

(Optioneel) ruimte voor opmerkingen:

Appendix B – Care@Home research

While this study already gave an overview of the learning curve, time on task and the number of problems encountered, some other tests were performed purely for the Care@Home project. A list of which system functionalities needed to be improved was provided as well. This appendix will give an overview these tests and their results, together with the improvement list.

User satisfaction

User satisfaction was measured by using the ASQ after each task in each trial. In general, ASQ's were high, meaning that users were very satisfied with the Care@Home system. There were some expectations on the ASQ scores over time. A higher ASQ score meant that a user was more satisfied with the system. Expectations were the following

A. The ASQ scores will be higher over time
B. The ASQ scores will be higher when time on task is lower
C. There are effects of age and gender on the ASQ scores (direction undefined)

In order to find out whether this was the case, a generalized estimating equations was performed. Assumptions were checked beforehand using the protocol as written by Zuur et al.(2010), see Appendix G. A gamma distribution with a log link was used, as the distribution of the residuals showed a negative skew and the values could only be positive to infinite. An autoregression correlation matrix was used as well.

Table B1: results generalized estimating equations for average ASQ scores

Parameter	B	Standard error	95% Confidence Interval		df	Sig.	QICC
			Lower bound	Upper bound			
Intercept	0,693	,4637	-,216	1,602	1	,135	98,758
Trial 1	0[a]	.	.	.	.	.	
Trial 2	,165	,0092	-1,245	,497	1	,000	
Trial 3	,162	,0059	,092	,237	1	,002	
Gender (M)	0[a]	.	.	.	.	.	
Gender (F)	,027	,1228	-,187	,241	1	,807	
Time on Task	-,001	,0002	-,001	,000	1	,000	
Age	,010	,0052	,000	,020	1	,047	
Prev. exp 2	,041	,0321	-,014	,096	1	,142	

a. Set to zero because this parameter is redundant. This is the parameter to which the others are compared.

ASQ scores between the three trials which confirm hypothesis A. Compared to trial 1, scores in trial 2 were on average 16.5% higher. Average ASQ scores in trial 3 were on average 16.2% higher than in trial 1. That the average ASQ scores do not rise more in trial 3 compared to 1 than in 2 compared to 1 can be explained by the finding related to learning. The overall learning curves show that participants tend to improve in time on task in trial 2 compared to 1, and then worsen again in trial 3 compared to trial 2. Still, they stay below the level in trial 1. It is very well possible that participants found the third trial more difficult than the third and based their ASQ scores on this.

Some prove for hypothesis B was found as well, as there was an effect of time on task on average ASQ scores. First of all, the effect was measured in reverse (so, what happens to the average ASQ score when time on task rises with one unit?). This showed that for each extra single unit of time on task, the ASQ would lower with (1- (exp(-0.001) = 0,99 =0.01) 1%. However, time on task was measured in seconds, so one unit was equal to one second. Translating this into results: if time on task rises with one second, the average ASQ score would lower with 1% - spending one extra minute on a task would have severe consequences on the ASQ scores.

For hypothesis C, only partial support was found. No effect on the average ASQ score was found for gender but there was an effect for age. Still, this was a very small effect: For every unit age would increase, the ASQ score would rise with (exp(0.010) = 1,01) 1%. This means that the 'older' group of elderly participants were more satisfied with the system than the 'younger' participants. There might be different reasons for this: First of all, it may be that the 'older' group of elderly users underestimated their own performance before the trials and found that it wasn't so hard at all (Zandbergen, 2015). Another possibility is that the 'younger' group of elderly had far more critique on the Care@Home system because they had more previous experience with other online systems which they could compare it to.

List of recommendations for Care@Home

Based on the usability problems found, a primary list with possible improvements was made for the technical partners of the Care@Home program. The list was made up in order of importance, so things on top of the list were the most urgent.

How do elderly people learn to work with an online system?

- Make sure the technical failures are gone:
  - Improvement of system lag, as it makes people select the wrong buttons beyond their own fault.
  - Make sure that there is no entry to parts of the system that aren't ready yet (as the system will crash when users do this by accident)
  - Fix the XML parser failures that occur when using the mailbox and the agenda functions.
  - Make it impossible for the pointer to float offscreen.
  - Restriction on character boxes where only 2 characters are allowed; make it impossible to add more than 2 characters
  - All English messages should be translated into Dutch
  - Dutch spelling errors should be corrected
- The remote control as it currently works should go. Replace it by a remote control with bigger buttons, enlarge the buttons on the TV screen. Best option would be to go without completely and use something more steady instead as a remote control, like a tablet.
- There should be more consistency overall:
  - Every page should include the name at the top (so the user knows at which page s/he currently is).
  - General graphics should be the same for every page. No changes in border size, no differences in fonts or sizes, etc.
  - Every page needs a visible way for the user to get back to both the previous way and the home page. Home page preferably by clicking the home button at the top left corner.
  - Changing data should be doable the same way in every menu. Make the thing that needs to be changed look like a button and make it clickable in order to change (as users did when they wanted to change the phone number; they clicked on 'phone number' itself, and not on the button 'change' next to it).
  - In the submenu's, pictograms and descriptions don't look like one thing and users do not know they can click the (bigger) pictogram as well. The design should be changed in such a way that the description and the pictogram are connected to each other.

  - Confirmation screens should always be the same. So the same words on the button(s) used, and those buttons always on the same side (e.g ‘yes’ is always the left button and ‘no’ always the right button).
  - Buttons for confirmations, adding and such should always be at the same place: Below the thing that needs to be confirmed.
- The color contrasts used should be more visible.
  - Make the background of e-mails that are unread visibly another color than those that are read.
  - Make the borders more visible, probably thicker as well.
- The labels used for menu descriptions should be revised
  - Split the ‘my neighborhood’ and make an extra menu for video exercises.
  - Either the address book or the agenda should be merged into one or the names should really be differentiated. Maybe put address book under “contacts”.
- Make the agenda look like a real agenda! This way, no legend is needed and it will shorten the clickpath participants need to follow in order to use it.
  - Make it possible to write something in the agenda simply by clicking a day.
  - Let participants add time by using a picture of a clock; this would work great when a tablet is used as a remote control as this would be easy to use with a touchscreen.
  - There should be only two boxes the participant needs to fill in; event description and time. There should be an option for an event to be ‘the whole day’ as well.
- Remove the legends throughout the system. They cause more confusion than they add clarity.
- Give more feedback, for example when an event is added to the agenda, when someone is added to the address book or when an e-mail is sent out. Do so by using a smaller screen that appears over the normal menu. Make it possible to turn this feedback off as well (for the advanced users).
- Remove the subfunction “memory” from the agenda. Include it in the new agenda by asking the user whether s/he would like to be reminded of the event after s/he has put it in the agenda
- For adding or viewing someone: make it possible to do so by clicking a person’s photo as well, not just the name.
- Make the font of unread e-mails bold (and the background contrast more obvious).

How do elderly people learn to work with an online system?

- Enlarge the boxes where the recipient and e-mail subject have to be filled in. Also put them less close together.
- Remove the ‘cool topic’ as the start up screen. Change it into a welcome message, possibly with a notion of new e-mails received (and make these clickable from that screen as well)
- Make it possible to send an e-mail to more people at once.
- Make it possible to send e-mails to people outside of the Care@Home system as well
- It should be able to send e-mails without a subject as well.
- Make it possible to write an e-mail and select a recipient afterwards as well.

Appendix C – Classification guideline

Introduction

Usability problems can be very useful when done correctly; In most cases when a product is being developed, multiple trials of usability tests are performed to find and fix as many problems as possible before the product is being launched into the real world market. However, sometimes a user makes a certain mistake when working with a product for the first time, learns, and thus does not make the same mistake when working with the system later on. Knowing which usability problems will solve themselves over time, as explained before, and which ones will stay can save a lot of time and money: Problems can be classified in less trials of usability testing and there needs to be done less work about fixing found usability problems as just those problems that are classified as "will stay over time" need specific focus to be fixed and those that solve themselves don't.

This document contains a basic guideline for classifying usability problems accordingly. First, the theory behind this usability classification guideline will be explained shortly, consisting of a mix from previous classification theories. Second, there will be some ideas on how to process your data before you can use this classification guideline properly. Next, there is a detailed description of the two dimensions and the eight types of problems of which the guideline consists, along with multiple control questions and examples to compare with your own set of usability problems.

Theory

Theory by Reason

Classifying mistakes has been done previously, as this explains a lot about how the human brain works. To start out with, Reason (1990) stated two basic types of mistakes:

1) **Execution failures – consisting of slips and lapses** : Here, a user knows what to do (intention or plan is correct) but the execution is not. Logically, this only happens in situations which are known for the user. The difference between a **slip** and a **lapse** is that:

    a) A **slip** concerns a situation in which the execution was incorrect.
    b) A **lapse** concerns a situation in which there was no execution at all.

2) **Planning failures**: A user does not know what to do (the intention or plan is incorrect) with a rather logical consequence that the execution is incorrect as well (most of the time, sometimes users take a good guess). These mistakes occur at settings that are rather unknown for the user.

Theory by Rasmussen

This basic classification can then be compared to aspects of another classification model that was created by Rasmussen (1983). In his model, Rasmussen defines that there are multiple dimensions of regulation control, or how conscious the action patterns are that the users express. This model contains the following, ranging from highest level of conscience to lowest level of conscience ("automatic" behaviour):

1) **Knowledge based behaviour and mistakes:** At this level, plans are made and regulation is mostly conscious, so there are no automatic processes but rather a serial step-by-step way of thinking and applying rules (like when following a step-by-step manual for putting together furniture from IKEA). A mistake will mostly belong to this category when a user has no rules known to the situation. Therefore, mistakes in this category are very diverse. A known cause is often an overload of information in a (too) short amount of time.
2) **Rule based behaviour and mistakes:** At this level, there is a lower level of conscious processing than at the intellectual level. Processing is mostly done in schemata by using ready-made programs which have to be specified by parameters and only work in certain situations (for example when you know how to bake a basic cake but don't know how to bake a chocolate cake). Processes here can be conscious but don't need to be**.** The user uses rules that worked in an earlier, other (mostly likewise) setting in a current setting. He or she uses the roles correctly but they do not work. The goal or plan as defined by the user is incorrect.
3) **Skill based behaviour and mistakes:** This level of behavior has the lowest level of regulation, as a lot of processes here are automated and can thus be performed without conscious attention. Regulation here cannot change action programs, at best only stop the performance coming from it. Mistakes do occur here when the situation is familiar to the user. The intention or plan is then correct but the execution is not.

The work of Reason can also be compared to that of Zapf et al.(1992). In their work (based on the German Action Theory), Zapf et al start with the comparison with the three levels of action regulation as proposed by Rasmussen:

How do elderly people learn to work with an online system?

1) **Intellectual level:** Comparable with knowledge based level by Rasmussen where conscious processing occurs almost all the time.
2) **Flexible action patterns**: Comparable to the rule based level as proposed by Rasmussen where conscious processing happens in schemata but is not needed all the time.
3) **Sensorimotor level:** Comparable to the skill-based level from Rasmussen where processes are almost automated and barely need conscious processing

Theory by Zapf et al.

Around the same time, Zapf et al. made a model similar to the one by Rasmussen, but extended it by adding a **knowledge base for regulation** which is used for developing plans and goals in the first place. This base consists of a) knowledge of facts, b) knowledge of procedures and c) understanding in the sense of mental models.

When comparing the work from Zapf et al with that from Reason and Rasmussen, as discussed previously, two things are noticeable. First of all, Zapf et al add a third dimension. Next to planning and monitoring problems, they add a category for usability problems based on feedback. Second, while Reason does not imply that slips and lapses (which basically differ in regulation level) can also occur during planning (only during execution), Zapf et al combine the dimensions of both regulation level and planning level. This creates a possibility to define type of error by two dimensions:

This base creates a possibility to define type of usability problems by two dimensions:

1) *Where in the process did the usability problem occur?*

    a) Planning
    b) Monitoring
    c) Feedback

2) *What is the level of regulation for this usability problem (as defined by Rasmussen)?*

    a) Knowledge level
    b) Rules level
    c) Skills level

Combining these dimensions gives, as described by Zapf et al, eight types of problems which are summarized in the table below:

Knowledge base for regulation			
Knowledge errors			
	Goals/Planning	**Monitoring**	**Feedback**
Knowledge level	Thought problems	Memory problems	Judgment problems
Rules level	Habit problems	Omission problems	Recognition problems
Skills level	Sensorimotor problems (slips/lapses)		

As can be seen above, sensorimotor mistakes happen only at the skills level of behaviour but in all three phases of the process. Mistakes in this category still need to be divided in slips and lapses by questioning for each usability problem found whether the point of execution was reached or not. Next is a short explanation for each type of usability problem or mistake from the table given above:

- **Knowledge problems:** The user cannot make a correct plan for execution because he or she does not know all the (sub)parts or commando's from the system that's being used. These problems can occur because the instructions about the program or task are inadequate, and can be traced back to the knowledge base for regulation.

Errors that occur in the knowledge level of regulation mostly are complex as there are a multitude of errors possible:

- **Thought problems:** As can be seen, these problems occur while setting up a goal or preparing a planning. While the user knows all the parts of the system (albeit in a very conscious way of processing), the plan or goal that is set up beforehand is incorrect.
- **Memory problems:** Happen during the task monitoring. The plan or goal is correctly set up, but while working with the system the user forgets part of the plan and thus forgets to execute this what either leads to a) possible execution of a task while forgetting a part or b) execution not possible because the part of the plan that was forgotten was necessary for execution.
- **Judgment problems:** Happen during the feedback phase, so after an the user has given the system input. The user receives feedback from the system but either does not understand this feedback or interprets this the wrong way.

As explained above, problems on the rules level happen when the actions that are performed are relatively well-known:

- **Habit problems:** Mistakes of these category occur at the beginning of a task. For example, a participant might say that "well, it looks like [something similar] so I figured it will work that way. This type of problem can occur when, for example, a user switch to a new program for an old task or after an interface redesign of a known program.

How do elderly people learn to work with an online system?

- **Omission problems:** Happen during monitoring, when a (sub)plan is executed incorrect even when it normally is done.. For example, when sending an e-mail one does not click 'send' but goes straight back to the main menu even when this went right three times before.
- **Recognition problems:** Happen during the feedback phase, when feedback provided by the system is misinterpreted, or misunderstood, even when someone did understand it before. It is really important to note that the difference between recognition problems and judgement problems is that judgement problems have to do with newly received feedback while recognition errors have to do with interpreting feedback that has been received (and understood) before.

Last, there is the level of skill-based problems. There is only one category here. As skill-based behavior is mostly performed at a less conscious level (automated), it is very hard to make a difference in whether the mistake occurred at the planning, monitoring or feedback phase:

- **Sensorimotor problems:** Mistakes where the plan or intention was fully correct but the execution failed. For example, when a participant presses the wrong button but immediately says that this was not his or her intention. The assimilation bias can also be found in sensorimotor problems, as an earlier learned automatism from another situation can lead to the execution of this automatism in the wrong situation. This level of behaviour can be divided into slips and lapses afterwards, depending on the outcome of the action:

  a) **Slip:** When execution goed right after some time (e.g. correcting a spelling mistake during typing)**.**
  b) **Lapse:** Execution end up in incorrect action (e.g. when a participant accidentally goes back to the main screen and knows this is incorrect but is not able to get back to the working screen).

Now that the theory behind usability problem classification has been explained thoroughly, it is time to define what should be done with the data before usability problems can actually be classified into one of the above categories

Prerequisites for data

Most usability tests give a lot of data to work with. Here are some ideas to get started:

1) These guidelines have been developed and tested to classify usability problems. Even though these guidelines might be able to classify individual errors, we strongly advise you to sort out your data by classifying and creating usability problems beforehand.
   Usability problems are created by grouping individual errors, or incidents, together to get a more general description or underlying idea of what went wrong.
   There are multiple methods available for doing getting these Usability problems. For our research, one of the methods as described by Lavery, Cockton and Atkinson (1997) was chosen. Here, similarities are found between individual errors based on multiple aspects of an incident. This is just one method and there are many more (for a comparison see, for example, Hornbæk & Frøkjær, 2008). If you choose another method, make sure that there is still enough details saved from the incidents to be able to classify the problems in the correct way.
2) You should pay attention to the following things (if applicable) when grouping incidents together (because this will make it easier to classify later on in the process):
   a. Intention: Did the participant had a plan for execution or not?
   b. Comparison: Did the participant compare the task with something familiar? (e.g. "oh, this looks just like my old computer, so I should probably do this...")
   c. Feedback: Was the participant able to know what the feedback meant? Has it appeared earlier to him or her during usability testing?
   d. Reappearance: Did the participant made the same mistake before?
3) Before testing, it might be a good idea to measure previous experience as well. As you will read later on, certain types of mistakes also depend on previous experience with a (likewise) system or device.

Following this, you should end up with a list of carefully described usability problems consisting of a collection of similar incidents.

Step-by-step Guidelines

Now that you have your list if usability problems, we will describe the guideline for classifying each problem into a category. Rather similar to the theory described above, distinction will first of all be made between the knowledge mistake and the rest of the schedule, as the knowledge problem in the taxonomy is placed separately. Afterwards, the seven problems of action theory that are left will be first broken down in the three different regulation levels (Dimension 1) which will in turn be broken down to the specific problem categories using the definition of the steps in the action process below (Dimension 2). For each usability problem, follow the steps below.

How do elderly people learn to work with an online system?

> ***Note :*** *If you feel you aren't able to classify a problem by following the steps, please read the problem classification with examples and control statements at the end of the document to compare your problem and find the best match.*

Start the steps

Step 1

In this step, a check will be made whether a usability problem is one in the range of knowledge base regulation (a knowledge problem) or not

Relevant questions:

- Did the user miss any knowledge about the buttons, functions, etc. making it impossible to complete the action successfully? *(Yes: Choice 2; No: Choice 1)*
- Did the user receive adequate instruction*? (Yes: Choice 1; No: Choice 2)*

Choices:

1. **Action regulation:** The user received an adequate instruction and had enough knowledge to possibly successfully perform the action: continue to STEP 2
2. **Knowledge base for regulation:** The user didn't receive (part of) an instruction or didn't know about certain buttons, functions, etc. It was impossible for the user to complete the action successfully: continue to CATEGORY KNOWLEDGE PROBLEMS.

Dimension One (Regulation level):

Step 2

If the usability problem is not a knowledge problem, the next step is to find out in which level of regulation it occurred: either knowledge-based, rule-based or skill-based.

Relevant questions:

- Did the outcome comply with the intention of the user? (Even if the outcome wasn't successful/useful for the task?) *(yes; choice 2, no; choice 1)*
- Does the user exclaim out loud that this wasn't what he meant to happen? *(yes; choice 1, no; choice 2)*

- Did the user accidentally press the wrong button/link or next to a button/link? *(yes; choice 1, no; choice 2)*

Choices:

1. **Skill level:** The user performed this action with little conscious thinking or almost automatically. The user seemed familiar with the situation. His or her plan of action was correct, even though the execution was not necessarily: continue to CATEGORY SENSORIMOTO PROBLEMS
2. **Rules level or Knowledge level:** The user used quite a lot of conscious control for this action, the situation was rather unknown or new to him or her, and most likely there was an incorrect plan of action: continue to STEP 3

Step 3

If your problem was not a skill-based sensorimotoric one, then it is either a knowledge-based or a rule-based problem

Relevant questions:

- Did the user use an (implicit) if/then statement or rule (if I do this.....then this will happen) in his plan for the action? *(yes; choice 1, no; choice 2)*
- Did the user encounter this same problem before in the same manner? *(yes; choice 1, no; choice 2)*
- Did the user find that this situation resembled something that he knew from another situation or recognise the situation? *(yes; choice 1, no; choice 2)*
- Did the user need to form a new plan for this action? *(yes; choice 2, no; choice 1)*
- Does the user say that he is going to try something new (but there is an intention/plan)? *(yes; choice 1, no; choice 2)*

  Note: If there is no intention or plan, it can never lead to a problem. So there has to be a plan or intention!

How do elderly people learn to work with an online system?

Choices:

1. **Rules level:** The user performed the action on the Rules level. There was a (schematic) plan that was probably based on earlier experiences with a (likewise) system, but this planning was incorrect, leading to an incorrect execution (in most cases): continue to STEP 4: (Dimension Two: Rule-based)
2. **Knowledge level:** The user performed the action on the Knowledge level. The user made a new plan, which was executed step-by-step. There might have been an information overload, as the user did get an introduction to the system but this might be a lot of information at once: continue to STEP 6 (Dimension Two: Knowledge-based)

Dimension Two (steps in action process):

Rule-based

Step 4

Your usability problem is rule-based. The next question is whether it took place during the planning, monitoring or during the feedback phase.

Relevant questions:

- Was the plan that the user formed adequate? *(yes; choice 2, no; choice 1)*
- Did the problem occur before the execution of the action? *(yes; choice 1, no; choice 2)*
- Was the action based on a habit of the user (from another situation)? *(yes; choice 1, no; choice 2)*
- Did a feature of the application lead to a wrong assumption/plan? *(yes; choice 1, no; choice 2)*

Choices:

1. **Planning:** The usability problem occurred in the phase of planning, so before the action was executed: continue to CATEGORY HABIT PROBLEMS
2. **Monitoring or Feedback:** The plan for execution was right, but the execution went wrong or feedback interpretation or usage after action performance went wrong: continue to STEP 5

Step 5

Your usability problem is rule-based and took place either during monitoring of during the feedback phase. This step is to find out when:

How do elderly people learn to work with an online system?

Relevant questions:

Monitoring:

- Did the user forget to execute a part of the plan? *(yes; choice 1, no; choice 2)*
- Did the error occur during a sub action? *(yes; choice 1, no; choice 2)*
- Was this part of the plan well known? *(yes; choice 1, no; choice 2)*

Feedback:

- Did the user complete the task? *(yes; choice 2, no; choice 1)*
- Did the user have trouble understanding or interpreting feedback by the program? *(yes; choice 2, no; choice 1*)
- Was this known/earlier encountered feedback? *(yes; choice 2, no; choice 1)*
- Was there a lack of feedback that confused the user? *(yes; choice 2, no; choice 1)*
    a. *(If/then construction: if I finish, then there will follow feedback. If this doesn't follow this is an Recognition problem)*
- Was there feedback present that the user didn't see which led to a problem? *(yes; choice 2, no; choice 1)*

Choices:

1. **Monitoring:** The problem encountered took place during execution of a (sub)plan and not afterwards. Therefore, it is an omission problem: continue to CATEGORY: OMISSION PROBLEMS
2. **Feedback:** The problem encountered took place after the execution of a (sub)action. It either happened because feedback was misinterpreted or because there was a lack of feedback. It is a recognition problem: continue to CATEGORY: RECOGNITION PROBLEMS.

Knowledge-based

Step 6

Your usability problem is knowledge-based. The next question is whether it took place during the planning, monitoring or during the feedback phase.

How do elderly people learn to work with an online system?

Relevant questions:

- Was the plan that the user formed adequate? *(yes; choice 2, no; choice 1)*
- Did the problem occur before the execution of the action? *(yes; choice 1, no; choice 2)*
- Did a feature of the application lead to a wrong assumption/plan? *(yes; choice 1, no; choice 2)*

Choices:

1. **Planning:** The usability problem occurred in the phase of planning, so before the action was executed. It is a thought problem: continue to CATEGORY: THOUGHT PROBLEMS
2. **Monitoring or Feedback:** The plan for execution was right, but the execution went wrong or feedback interpretation or usage after action performance went wrong: continue to **Step 7**

Step 7

Your usability problem is a knowledge-based problem that had a good action plan. It did went wrong either during monitoring or feedback. This step is to check at which point it went wrong.

Relevant questions:

Monitoring

- Did the user forget to execute a part of the plan? *(yes; choice 1, no; choice 2)*
- Did the error occur during a sub action? *(yes; choice 1, no; choice 2)*

Feedback

- Did the user complete the task? *(yes; choice 2, no; choice 1)*
- Did the user have trouble understanding or interpreting feedback by the program? *(yes; choice 2, no; choice 1)*
- Was this new/unknown feedback? *(yes; choice 2, no; choice 1)*

Choices:

1. **Monitoring:** The problem encountered took place during execution of a (sub)plan and not afterwards. Therefore, it is a memory problem: continue to **Memory problem**
2. **Feedback:** The problem encountered took place after the execution of a (sub)action. It either happened because feedback was. It is a judgement problem: continue to **Judgment problems**

How do elderly people learn to work with an online system?

Problem Categories

Below each category description you will find a short set of control statements and examples. Use these to make sure your usability problem belongs to the right category in case of doubt.

Knowledge problems

The user cannot make a correct plan for execution because he or she does not know all the (sub)parts or commando's from the system that's being used. These problems can occur because the instructions about the program or task are inadequate, and can be traced back to the knowledge base for regulation.

- *The user has not performed the task with the tested device before*
- *The user has not worked with a (very) similar device before*
- *The user states that he or she has no idea how to do this, since it is unlike anything witnessed before*
- *Possible to check by questionnaires about previous experience with the device tested, or similar devices.*
- *The user didn't receive the correct instructions about buttons, touch screen or functions beforehand to perform the task*

Thought problems

As can be seen, these problems occur while setting up a goal or preparing a planning. While the user knows all the parts of the system (albeit in a very conscious way of processing), the plan or goal that is set up beforehand is incorrect.

- *The user received adequate instructions.*
- *The user shows an incorrect plan of action when thinking out loud.*
- *The user wants to try something to see if it will work.*

Memory problems

Happen during the task monitoring. The plan or goal is correctly set up, but while working with the system the user forgets part of the plan and thus forgets to execute this what either leads to a) possible execution of a task while forgetting a part or b) execution not possible because the part of the plan that was forgotten was necessary for execution.

- *The user has a correct plan of action before actually doing something but forgets to execute a part of it.*
- *The plan that the user wants to execute is newly formed/no prior experience with the plan.*
- *It is only a memory problem if the user forgot to perform the action. If he tried to perform it but he failed this is another type of problem (For example: when trying to click the save button, but it doesn't react and you don't know what is happening → judgment problem).*
- *If the user tried to click it but failed in the action and doesn't notice it, this is a sensorimotor problem and NOT a memory problem.*

Judgment problems

Happen during the feedback phase, so after an the user has given the system input. The user receives feedback from the system but either does not understand this feedback or interprets this the wrong way.

- *The user notices the feedback (either by responding to it verbally or behaviourally) but does not know what to do with it, or act wrong on it.*
- *The user indicates to not understand this feedback ("Huh? What is this about?" or something likewise)*
- *The user did not receive this feedback before, and if he or she did receive it, not understand it then either.*

Habit problems

Mistakes of these category occur at the beginning of a task. Participants want to perform an action or plan that in itself is not wrong but the moment of using this action is wrong. For example, a participant might say that "well, it looks like [something similar] so I figured it will work that way. This type of problem can occur when, for example, a user switch to a new program for an old task or after an interface redesign of a known program.

- *The user is familiar with the system or task, or a likewise system or task.*

- *The action the user performs in itself is not wrong. The action could have been correct in another situation. The place or situation is wrong*
- *The user exclaims this will probably work like a similar situation he knows, or that he wants to try if this is the same as another situation.*

Omission problems

Happen during monitoring, when a (sub)plan is executed incorrect even when it normally is done.. For example, when sending an e-mail one does not click 'send' but goes straight back to the main menu even when this went right three times before.

- *The user already talks or thinks out loud about the next step that has to be performed (e.g. a user sending an e-mail thinking out loud: "I will have to go to outbox to check whether I have sent it" who consequently forgets to hit the send button and goes straight to outbox, only to discover that the mail was not send).*
- *The user has performed the task correctly before.*
- *The plan was adequate for the task.*

Recognition problems

Happen during the feedback phase, when feedback provided by the system is misinterpreted, or misunderstood, even when someone did understand it before. It is really important to note that the difference between recognition problems and judgement problems is that judgement problems have to do with newly received feedback while recognition errors have to do with interpreting feedback that has been received (and understood) before.

- *The user shows no indication of noticing a feedback message from the system halfway a task (or during a subtask) when it appears. The user continues without the feedback.*
- *Feedback is present but the user didn't notice it, due to the feeling that he was already finished and didn't need to pay attention anymore.*
- *The user has shown intention of noticing this feedback message earlier.*

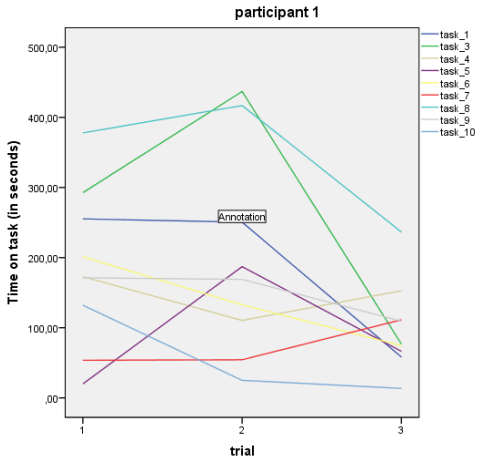
- *The user has shown before to know the meaning of this feedback message, either in this system or a likewise system.*
- *The user indicates that he or she is missing feedback: either by indicating directly ("it would have been nice if the system would tell me what to do next") or indirectly ("I don't know what to do next..?").*
- *If the user doesn't notice a lack of feedback due to automatized behaviour (for example: trying to check a box and click on continue, but the box is still empty and the page doesn't react) it is a sensorimotor problem if the user understands what went wrong and a recognition problem if he doesn't understand the feedback.*

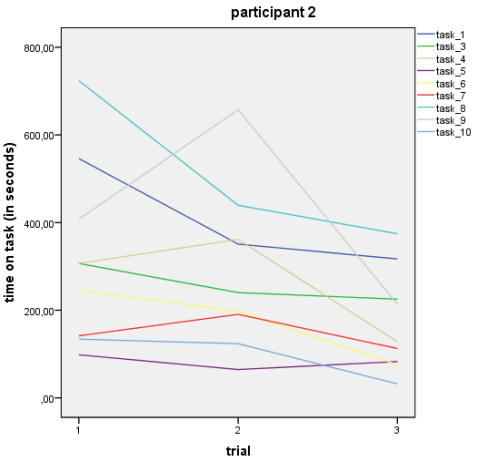
Sensorimotor problems

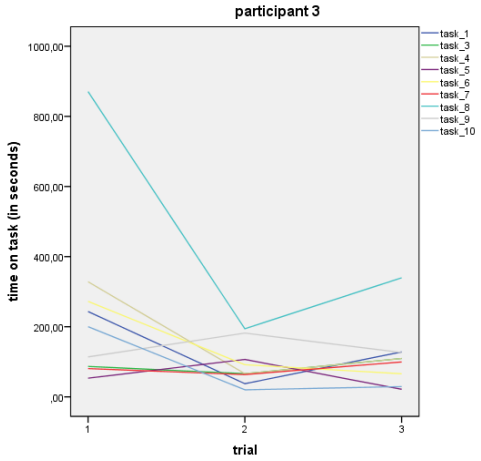
Mistakes where the plan or intention was fully correct but the execution failed. For example, when a participant presses the wrong button but immediately says that this was not his or her intention. The assimilation bias can also be found in sensorimotor problems, as an earlier learned automatism from another

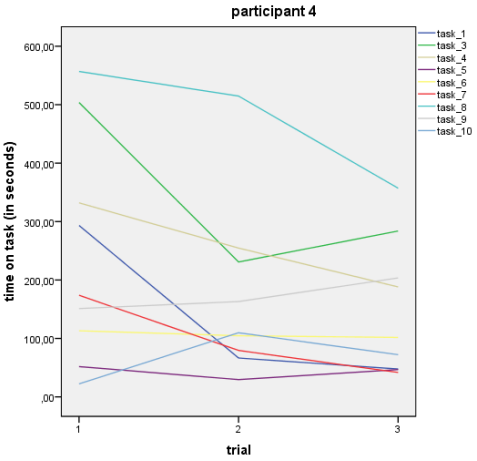
- *The user states that he or she knows what to do, or describes a (correct!) plan of action.*
- *The user immediately indicates that the thing that went wrong was a mistake, or even explains what he or she intended to do instead (note: this description must be correct!)*
- *The error is a physical one: for example, knowing what the next step is but accidentally pressing a wrong button because they are too close*
- *As there is no separate feedback level for the sensorimotor level, it is possible that a user tries to click a button, misses, and doesn't notice this due to automatized behaviour. This also qualifies as a sensorimotor problem.*

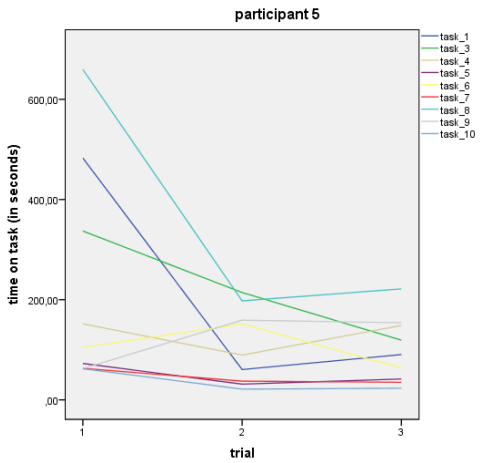
Appendix D – Individual learning curves

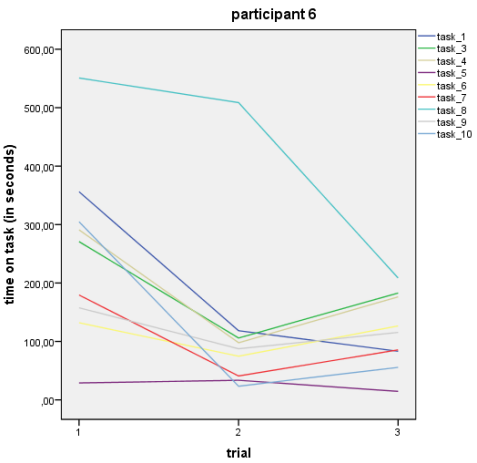

participant 1
Time on task (in seconds)
500,00
400,00
300,00
200,00
100,00
,00
Annotation
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

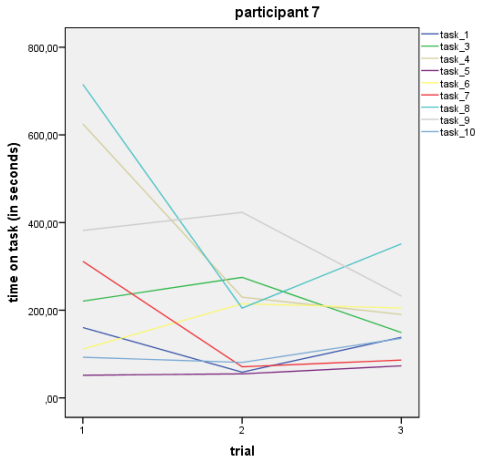

participant 2
time on task (in seconds)
800,00
600,00
400,00
200,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

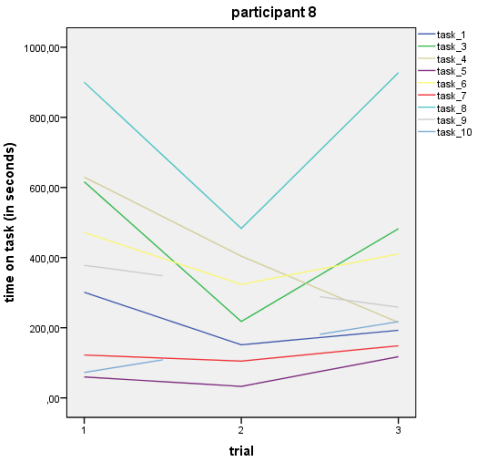

participant 3
time on task (in seconds)
1000,00
800,00
600,00
400,00
200,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

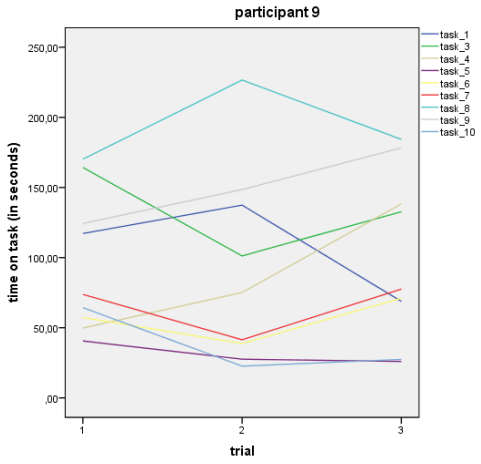

participant 4
time on task (in seconds)
600,00
500,00
400,00
300,00
200,00
100,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

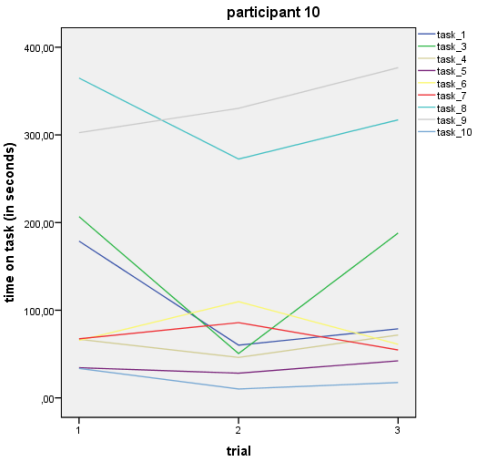

participant 5
time on task (in seconds)
600,00
400,00
200,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

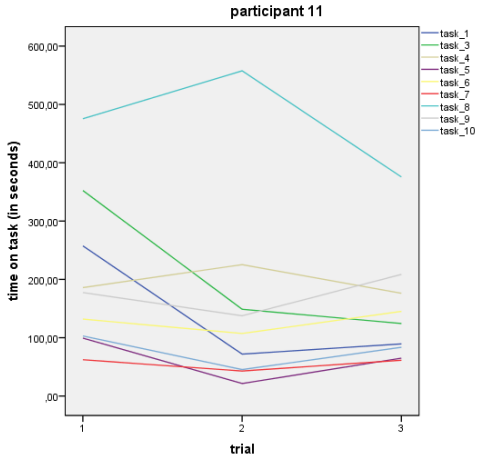

participant 6
time on task (in seconds)
600,00
500,00
400,00
300,00
200,00
100,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

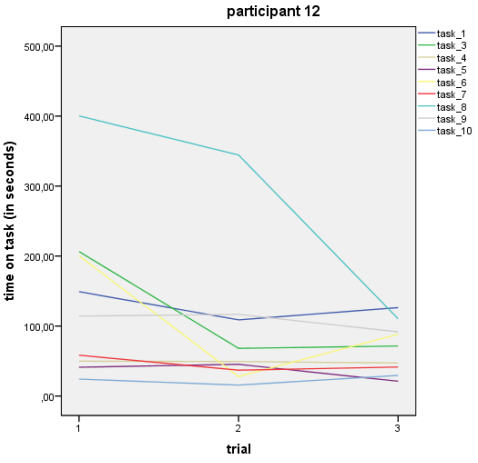

participant 7
time on task (in seconds)
800,00
600,00
400,00
200,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

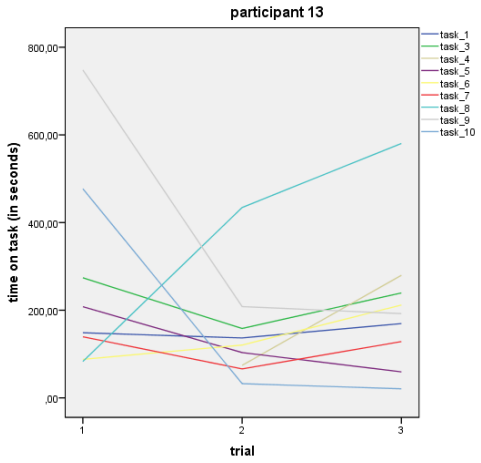

participant 8
time on task (in seconds)
1000,00
800,00
600,00
400,00
200,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

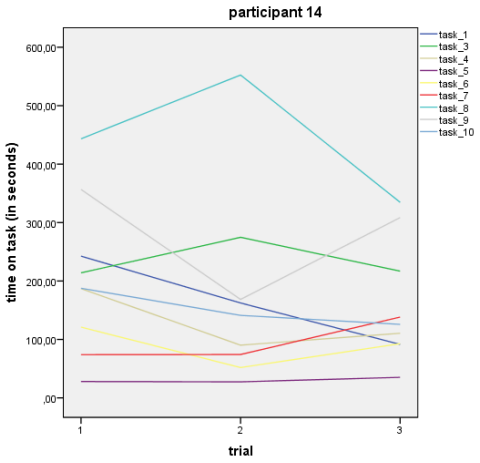

participant 9
time on task (in seconds)
250,00
200,00
150,00
100,00
50,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

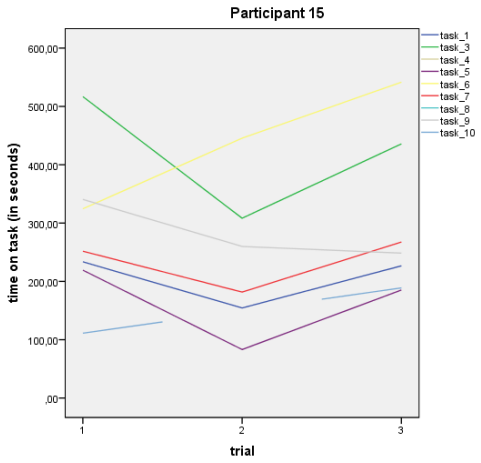

participant 10
time on task (in seconds)
400,00
300,00
200,00
100,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

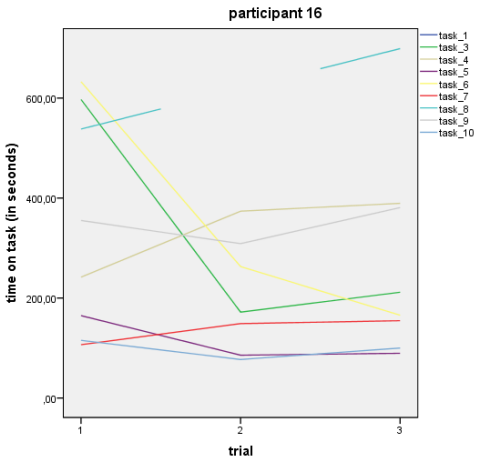

participant 11
time on task (in seconds)
600,00
500,00
400,00
300,00
200,00
100,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

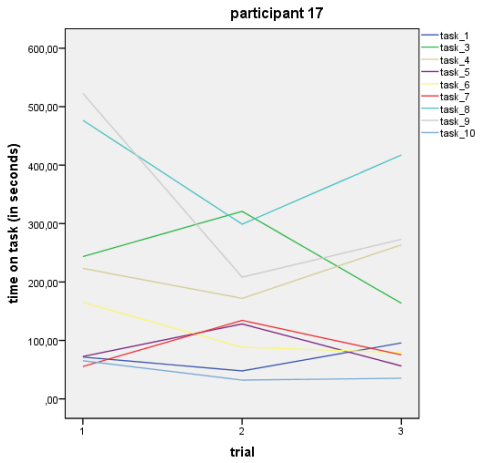

participant 12
time on task (in seconds)
500,00
400,00
300,00
200,00
100,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

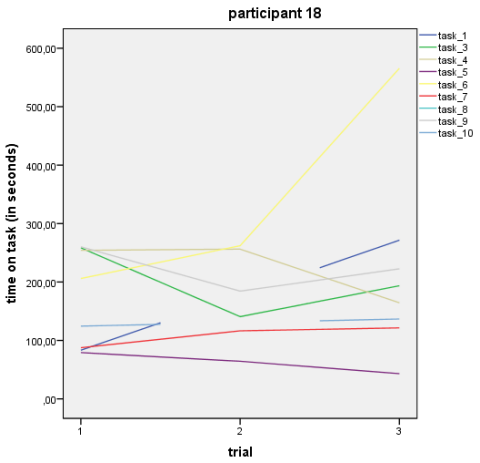

participant 13
time on task (in seconds)
800,00
600,00
400,00
200,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

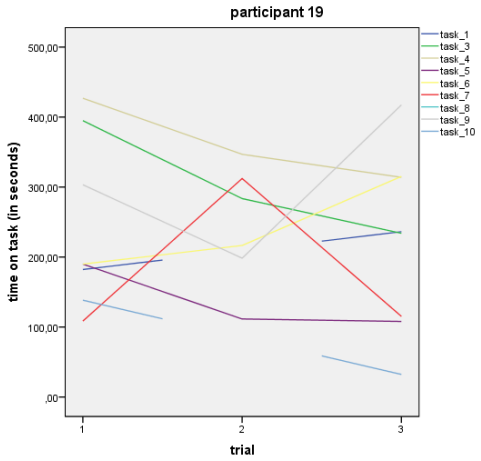

participant 14
time on task (in seconds)
600,00
500,00
400,00
300,00
200,00
100,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

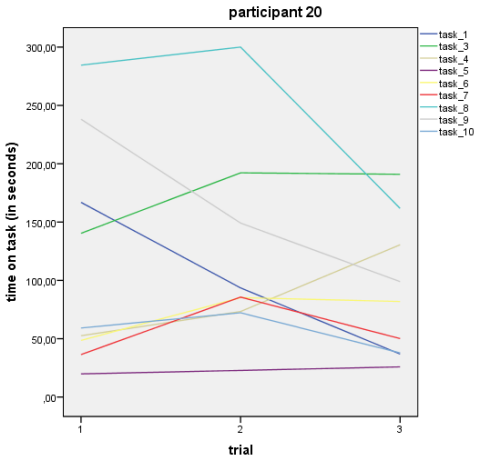

Participant 15
time on task (in seconds)
600,00
500,00
400,00
300,00
200,00
100,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10


participant 16
time on task (in seconds)
600,00
400,00
200,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10


participant 17
time on task (in seconds)
600,00
500,00
400,00
300,00
200,00
100,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10


participant 18
time on task (in seconds)
600,00
500,00
400,00
300,00
200,00
100,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10


participant 19
time on task (in seconds)
500,00
400,00
300,00
200,00
100,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10


participant 20
time on task (in seconds)
300,00
250,00
200,00
150,00
100,00
50,00
,00
1
2
3
trial
task_1
task_3
task_4
task_5
task_6
task_7
task_8
task_9
task_10

Appendix E – List of all usability problems in order of severity

Problem number (ranking)	Problem ID	Description	Prevalence in percentages (trial 1, trial 2, trial 3)	Type of problem (classification)
1	115	User does not know where to find the movie with the exercises (it's under "mijn buurt" (my neighborhood))	60-65-65	sensorimotor
2	109	User is at the correct screen but goes back to the main menu because he/she believes it's not the right screen	30-25-60	recognition
3	12	User thinks events (agenda) are in the address book or contact, or thinks addresses are in the agenda (confuses the two)	40-30-55	thought
4	129	User goes to the wrong day (e.g. goes to today when it has to be in two days)	60-75-50	sensorimotor
5	15	User thinks the legend pictures are buttons	55-25-45	thought
6	112	User accidentally goes back to the previous screen	60-25-45	sensorimotor
7	67	User thinks he/she received new e-mails while this is not the case (or the other way around)	60-35-35	thought
8	117	The user cannot find the pointer	50-25-30	sensorimotor
9	10	User thinks there are events next week while there aren't any (misreads the pink/purple part of the legend)	50-35-30	thought
10	19	User clicks on the phone number itself instead of on the button "change phone number" next to it, in order to edit a contact's phone number	20-25-45	thought
11	98	User tries to use TAB to get to the next box for filling in time (event). Leads to a mistake as the system recognizes it only as an extra symbol (and three are too much)	25-30-25	habit
12	68	User forgets to hold the Fn key to type a number instead of a letter -> types a letter	35-5-25	memory
13	108	User does not know how to confirm adding someone to the address book	30-25-25	recognition
14	4	It is not clear to the user how to get the pointer in the box in order to type	25-20-25	knowledge
15	37	User selects the page where he/she is already at in the left menu	20-20-25	thought

16	106	It is not clear to the user that he or she is already in the box where he/she can type	20-10-25	recognition
17	124	User has trouble distinguishing the borders of the boxes in the left menu	15-25-20	sensorimotor
18	16	User thinks (sending) e-mail can be found under address book (or the other way around).	25-10-20	thought
19	14	User clicks "wijs taak toe" or "voeg toe" or "agenda" in the left menu instead of "creëer gebeurtenis" at the bottom -> everything that the user filled in disappears.	30-10-20	thought
20	114	User cannot find a number on the remote control	45-10-20	sensorimotor
21	116	User accidentally presses the wrong button on the screen (remote control slip)	25-20-20	sensorimotor
22	18	User clicks on the day itself to see whether there is an event (instead of overlooking the agenda at once)	20-25-15	thought
23	71	User forgets to add a subject when typing an e-mail (which gives an error message when he/she tries to send it)	30-20-15	memory
24	26	User adds more than 2 digits in the time box, which creates a scroll bar which in turn causes an error message (agenda - adding event)	35-20-10	thought
25	110	User makes a spelling error (types wrong letter)	20-20-20	sensorimotor
26	11	User wants to click on the person 'Gert Dijkstra' but it does not respond (only by clicking the exact name or photo)	20-20-15	thought
27	99	The user starts typing but the box is not selected so nothing happens	10-15-20	omission
28	53 α	User thinks there are events on a day on which there are no events planned	25-5-15	thought
29	87α	User holds the Fn button when it isn't necessary (so, numbers or punctuation marks appear instead of letters - Compares to shift)	25-5-15	habit
30	69	User forgets to fill in the time, does not notice the hh/mm boxes altogether (when creating an event)	20-10-15	memory
31	79	User is confused by 'no results found' in the search results section (two sections, one always displays no results found)	25-10-15	judgment

32	78	User forgets to include a space in between two names (so no result is found)	50-10-15	memory
33	20	User keeps the remote control upside down (which makes it very difficult to aim right)	15-25-15	thought
34	107	User is not sure whether he/she send out an e-mail	40-15-15	recognition
35	61	User types O instead of 0 (they are close and look the same on the remote)	20-15-10	thought
36	55	User thinks the button 'terug naar adresboek' is an instruction instead of a button (reads it but does not use it. Uses the button on the left menu instead)	35-15-10	thought
37	86	User does not comprehend the English error message (e.g. when forgetting content in e-mail, filling in time wrong for an agenda event, etc.)	20-15-5	judgment
38	27	User sees the contact under 'resultaten" or "suggesties" and thinks he/she is added to the address book (does not confirm)	15-0-20	thought
39	1	User does not know whether there are events next week	20-0-15	knowledge
40	8	User clicks multiple time everywhere on the screen to see which parts reacts (aka: options where to go next)	15-15-15	knowledge
41	64	User tries to send an e-mail from the agenda	15-5-10	thought
42	32	User clicks below or on the "suggesties" or "resultaten" bar because he/she thinks this will lead to adding another contact	0-15-15	thought
43	101	User clicks OK without editing the phone number	5-10-15	omission
44	83	User clicks multiple times on a page that is already clicked and loading	10-10-15	judgment
45	2β	User does not know which one of the two movies is the one with the exercises.	10-5-15	knowledge
46	45β	It is not clear to the user that both the small box with the name of the subpage and the bigger pictogram (which is easier to aim at) will lead to the same page	10-5-15	thought
47	24	User uses the numbers that are on the front of the remote control (which don't	15-10-10	thought

		do anything)		
48	85	User clicks ´no´ in the confirmation screen for deleting an e-mail (or a person) (instead of yes, delete)	20-10-10	judgment
49	80γ	User does not know how to delete a phone number (while stating how to do it earlier)	35-10-10	judgment
50	111γ	User accidentally goes back to the main screen	35-10-10	sensorimotor
51	118	User clicks one of the arrow buttons on the remote control (opens a small, black-transparant box on the tv screen)	45-10-10	sensorimotor
52	70	User forgets to confirm when adding / deleting a person to / from the address book	15-10-5	memory
53	125	User accidentally selects the wrong person	20-10-5	sensorimotor
54	102	User forgets to type one of the letters (e.g. when trying to find a contact for the address book)	0-10-15	omission
55	34	User wants to click on "address book" in the left menu to start searching for the contact to add, after typing the name (makes the name disappear)	15-0-10	thought
56	46ω	User thinks he/she already clicked a button because the darker over roll color has a delay (so the button is darker even if the pointer is not on it)	5-10-10	thought
57	52ω	User clicks exactly on the border between two boxes (nothing happens)	5-10-10	thought
58	42	User finds a contact for the address book and clicks 'address book' in the left menu because he/she thinks this is a confirmation for putting the contact in the address book	10-10-10	thought
59	128	User fills in his/her very own birthday instead of the imaginary one we put in the test (agenda)	10-10-5	sensorimotor
60	81	User does not know how to fill in the time/The hh/mm 24h notion is not clear to the participant	5-5-15	judgment
61	63δ	User clicks on 'onderwerp' itself and then starts to type	5-10-5	thought
62	82δ	User is confused by the name of the e-mail recipients who he/she emailed (not the name but cahuser[no]): thinks he/she made a mistake	5-10-5	judgment

63	48δ	User only fills in one of the two boxes for time (putting an event in the agenda)	5-10-5	thought
64	6ε	User does not know that the two space bars on the remote control have the exact same functionality	10-5-5	knowledge
65	25ε	User clicks on the day in the agenda but this does not work (only clicking the day opens it)	10-5-5	thought
66	28ε	User clicks "wijzig details" instead of "verwijder contactpersoon" (so, edit instead of deleting)	10-5-5	thought
67	96ε	User thinks "maak boodschap" is for ordering groceries, or something likewise, Compares it to the AH/Jumbo	10-5-5	habit
68	121ε	User holds the wrong button on the remote control for navigation (e.g. not the OK button but the one next to it) -> nothing happens	10-5-5	sensorimotor
69	7ζ	User does not know how to get back from "wijzig naam" to the previous screen	15-5-5	knowledge
70	38ζ	User clicks "wijs taak toe" instead of "voeg toe" for adding a new event to the agenda (so for opening an empty event sheet, **not** confirming)	15-5-5	thought
71	43ζ	User fills in 99 at hours (for agenda event)	15-5-5	thought
72	23	User clicks "cool topic" because he/she thinks it will open into another screen (e.g. e-mail or events)	35-5-5	thought
73	5	User does not know how he/she can check an event that he/she put in the agenda his/herself	0-5-15	knowledge
74	29	User thinks "postvak UIT" is for creating new e-mails	5-5-20	thought
75	30	User clicks "geen resultaten gevonden" because he/she thinks that is the way to find or commit a person (in the address book)	0-5-10	thought
76	43	User fills in 99 at hours (for agenda event)	5-0-20	thought
77	84	User does not know how to send an e-mail (after typing, does not know that he/she has to click the "verzend bericht" button under the mail)	0-15-5	judgment
78	49	User forwards the wrong e-mail	0-10-5	thought
79	22η	User clicks the arrow to go to next	10-0-5	thought

		week (in the agenda)		
80	113η	User has trouble aiming the remote control at a small piece	10-0-5	sensorimotor
81	123η	User is confused because he or she cannot find things from his/her daily life back in the agenda	10-0-5	sensorimotor
82	13	User does not know how to add a new made event to the agenda (by clicking "creëer gebeurtenis")	20-0-5	thought
83	36θ	User fills in the time and clicks on the box again (makes the time disappear)	5-5-5	thought
84	58θ	User clicks the big envelope on the banner to watch his/her new e-mail	5-5-5	thought
85	65θ	User deletes the wrong e-mail	5-5-5	thought
86	66θ	User cannot find the C1000 e-mail (the one that has to be forwarded) in the e-mail list	5-5-5	thought
87	100θ	User forgets to confirm when deleting an e-mail	5-5-5	omission
88	39ι	User forwards an e-mail instead of creating a new one	0-5-5	thought
89	90ι	User looks for his contacts under 'contacts' because "that's the way my computer at home has it organized" (so, not expecting to find a mailing option)	0-5-5	habit
90	104ι	User is confused as he/she cannot directly see who the received e-mail is from (cahuser1, 2 or 3)	0-5-5	recognition
91	35κ	User thinks the light blue rollover color has a particular meaning as well, just as the legend colors (in the agenda)	5-0-5	thought
92	40κ	User clicks 'herinnering' instead of 'agenda' (subsections of agenda in the left menu)	5-0-5	thought
93	47κ	User presses the OK button the entire time (unnecessary and it makes the hand tired)	5-0-5	thought
94	59κ	User clicks the banner to see the events	5-0-5	thought
95	74κ	User forgets to add a subject when creating an event in the agenda	5-0-5	memory
96	105	User indicates that he/she does not know whether he/she has edited the phone number	5-15-0	recognition
97	31	User thinks that the exercise movie is in the agenda	15-5-0	thought
98	72λ	User forgets to keep his/her finger on	10-5-0	memory

		the OK button for pointer navigation		
99	73λ	User clicks "creëer gebeurtenis" without filling in anything -which gives an error message	10-5-0	memory
100	120λ	User presses a button on the remote that brings him/her back to the main screen	10-5-0	sensorimotor
101	77	User goes to 'herinnering' instead of agenda and deletes a so-called memory	20-0-0	memory
102	75	User wants to send an e-mail without content (error message)	15-0-0	memory
103	41	User thinks clicking "suggesties" or "resultaten" above the person will put a person into the address book (so, as a confirmation)	0-010	thought
104	21	User opens an existing e-mail instead of creating a new one	0-10-0	thought
105	60ν	User has trouble understanding the legend as he/she is colorblind	10-0-0	thought
106	88ν	User clicks a box twice for typing but de-selects the box for typing this way (normally, double clicking works better for selecting a button. For typing, just click once)	10-0-0	habit
107	97ν	User isn't aware that the Fn button has to be either held all the time or pressed again for every mark (compares it to caps lock)	10-0-0	habit
108	57ς	User clicks the big envelope on the top banner to send an e-mail	5-5-0	thought
109	119ς	User can't find the letter that he/she wants to type (on the remote control)	5-5-0	sensorimotor
110	50ψ	User edits the phone number of the wrong person (not RC related!)	0-0-5	thought
111	62ψ	User goes to the wrong edit (e.g. "wijzig details" or "wijzig adres" instead of "wijzig tel. Nr.)	0-0-5	thought
112	91ψ	User wants to send an e-mail to a contact that's not in the list (not possible yet)	0-0-5	habit
113	33φ	User adds a dot after the name of the contact he/she wants to add (which the system cannot find)	0-5-0	thought
114	51φ	User thinks clicking 'postvak UIT' will send the e-mail he/she created	0-5-0	thought
115	56φ	User goes back to the main menu before going to the next screen (instead	0-5-0	thought

		of directly selecting the right screen in the left menu)		
116	94φ	User looks for the contact point of the remote to the TV (thus, compares the remote to a regular remote)	0-5-0	habit
117	95φ	User checks the address book for upcoming events because "that's the way I do it at home. I keep my appointments in my address book"	0-5-0	habit
118	122φ	User adds numbers on the wrong place (when changing a phone number, e.g. in the front or in the middle instead of the last 3 numbers)	0-5-0	sensorimotor
119	3τ	User can't go back to the main screen when a movie is opened (only with the Esc. Button on the remote)	5-0-0	knowledge
120	9τ	User has no idea how to hold the remote control (is literally saying that he/she does not know)	5-0-0	knowledge
121	17τ	User clicks on the legend in the agenda because he/she thinks this will open a list of events	5-0-0	thought
122	54τ	User thinks the exercise movie can be found under 'video contact' (in contact)	5-0-0	thought
123	76τ	User notices the hh/mm boxes but forgets to fill in the time anyway	5-0-0	memory
124	89τ	User accidentally types O instead of 0 (apart from spelling mistakes, as this is the utmost common typo)	5-0-0	habit
125	92τ	User is looking for an option to send a mail to multiple persons at the same time (there is no such function available)	5-0-0	habit
126	93τ	User tries to make a mail first and then select a sender (can only be done the other way around)	5-0-0	habit
127	103τ	User thinks that the event he/she just added is not added yet	5-0-0	recognition
128	126τ	User uses two hands to direct the remote control (e.g. against a tremor)	5-0-0	sensorimotor
129	127τ	User has trouble seeing the cursor for typing	5-0-0	sensorimotor

α = same prevalence values for problems with the ID's 53 and 87
β = same prevalence values for problems with the ID's 2 and 45
γ = same prevalence values for problems with the ID's 80 and 111
ω = same prevalence values for problems with the ID's 56 and 52
δ = same prevalence values for problems with the ID's 48, 63 and 82
ε = same prevalence values for problems with the ID's 6, 25, 28, 96 and 121

ζ = same prevalence values for problems with the ID's 7, 38 and 43
η = same prevalence values for problems with the ID's 22, 113 and 123
θ = same prevalence values for problems with the ID's 36, 58, 65, 66 and 100
ι = same prevalence values for problems with the ID's 39, 90 and 104
κ = same prevalence value for problems with the ID's 35, 40, 47, 59 and 74
λ = same prevalence values for problems with the ID's 72, 73 and 120
ν = same prevalence values for problems with the ID's 60, 88 and 97
ς = same prevalence values for problems with the ID's 57 and 119
τ = same prevalence values for problems with the ID's 3, 9, 17, 54, 76, 89, 92, 93, 103, 127 and 127
φ = same prevalence values for problems with the ID's 33, 51, 56, 93, 94 and 122
ψ = same prevalence values for problems with the ID's 50, 62 and 91

Appendix F – Ranking of usability problem per type

Case Summaries[a]

				Ranking
Type	knowledge error	1		14
		2		39
		3		40
		4		45
		5		64
		6		69
		7		73
		8		119
		9		120
		Total	N	9
			Mean	64,78
			Median	64,00
	thought error	1		3
		2		5
		3		7
		4		9
		5		10
		6		15
		7		18
		8		19
		9		22
		10		24
		11		26
		12		28
		13		33
		14		35
		15		36
		16		38
		17		41
		18		42
		19		46
		20		47
		21		55

How do elderly people learn to work with an online system?

22	56
23	57
24	58
25	61
26	63
27	65
28	66
29	70
30	71
31	72
32	74
33	75
34	76
35	78
36	79
37	82
38	83
39	84
40	85
41	86
42	88
43	91
44	92
45	93
46	94
47	97
48	103
49	104
50	105
51	108
52	110
53	111
54	113
55	114
56	115
57	121
58	122

How do elderly people learn to work with an online system?

	Total	N	58
		Mean	65,19
		Median	70,50
memory error	1		12
	2		23
	3		30
	4		32
	5		52
	6		95
	7		98
	8		99
	9		101
	10		102
	11		123
	Total	N	11
		Mean	69,73
		Median	95,00
judgment error	1		31
	2		37
	3		44
	4		48
	5		49
	6		60
	7		62
	8		77
	Total	N	8
		Mean	51,00
		Median	48,50
habit error	1		11
	2		29
	3		67
	4		89
	5		106
	6		107
	7		112
	8		116
	9		117

	10		124
	11		125
	12		126
	Total	N	12
		Mean	94,08
		Median	109,50
omission error	1		27
	2		43
	3		54
	4		87
	Total	N	4
		Mean	52,75
		Median	48,50
recognition error	1		2
	2		13
	3		16
	4		34
	5		90
	6		96
	7		127
	Total	N	7
		Mean	54,00
		Median	34,00
sensorimotor error	1		1
	2		4
	3		6
	4		8
	5		17
	6		20
	7		21
	8		25
	9		50
	10		51
	11		53
	12		59
	13		68
	14		80

How do elderly people learn to work with an online system?

	15		81
	16		100
	17		109
	18		118
	19		128
	20		129
	Total	N	20
		Mean	56,40
		Median	52,00
Total	N		129
	Mean		65,00
	Median		65,00

a. Limited to first 150 cases.

Appendix G – Assumption checkup GEE

GEE 1 : Time on task

1. Situation and variables

The prediction is that there time on task is lower in the second and third trial than in the first one. We therefore have one variable, time on task (μ = 188,366, σ = 158,708, σ^2 = 25188,210) measured in seconds. It is also expected that participants' age, gender and previous experience possibly have an influence as well on time on task.

2. Data exploration

Are there outliers?

As there were participants who did not complete all tasks, time on task was zero for some participants. These values were treated as missing values in SPSS. The data was already corrected for those moments where the system crashed or went offline. Boxplots were made for the residuals of time on task to check for outliers. Checking the data showed that these outliers did not happen under special circumstances. As outliers for time on task are to be expected and they are relevant for the findings, it was chosen not to remove any data from the set.

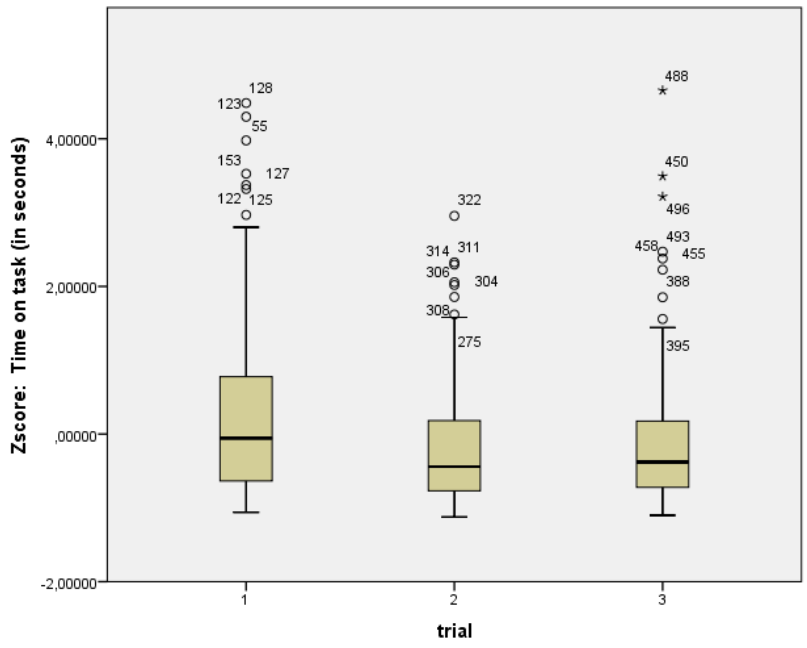


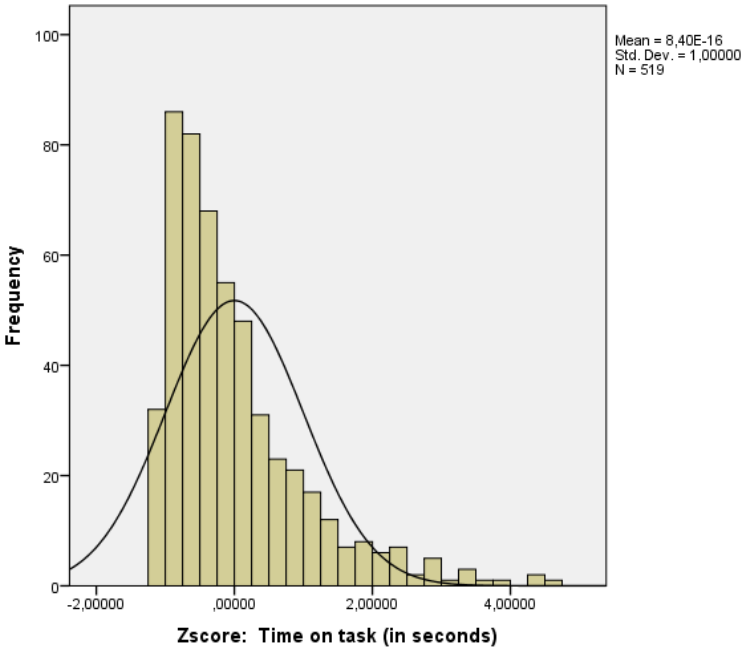
Figure G1: *Residual boxplots for time on task*

Is there homoscedasticity?

Figure G1 above shows that there does not seem to be homoscedasticity. The variance of the residuals is not the same in each group. This creates the possibility of a least squares estimation underestimating the standard errors which results in an increase of type I error.

Is there a normal distribution?

A histogram was made using the residuals of time on task (figure G2). This histogram shows that there is a strong positive skew when compared to how the normal distribution should be: A normal distribution cannot be assumed



q

Figure G2: *Histogram for residuals of time on task variable*

Are there a lot of zeros in the data?

The zero's in the time on task data were taken into account as missing data. The variables trial, gender and age do not have any zero's. There are some zero's in the variables for previous experience.

Is there collinearity?

Scatterplots showed that there seemed to be collinearity between age and experience in hours (figure G3 below). Relating this to real life makes sense: 'older' elderly tend to use

web browsing technology less than 'younger' elderly (PewresearchCenter, 2014). Models will have to be compared using either one of the two variables to determine the best fit. No collinearity was found between the other variables.

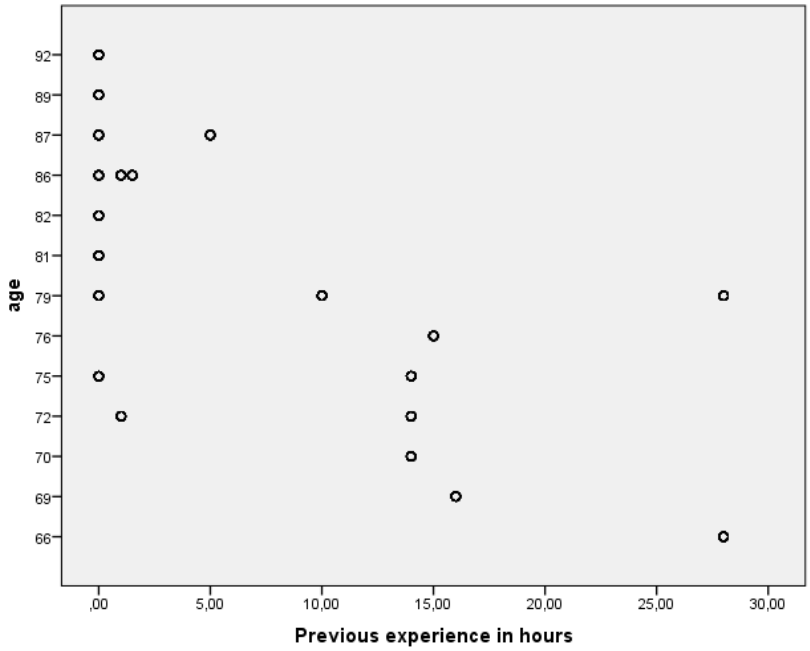


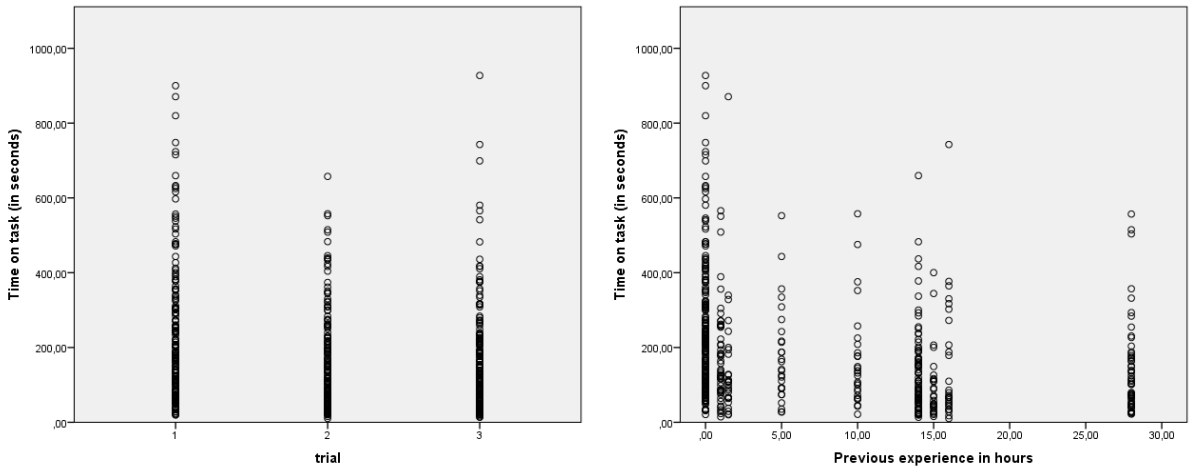
Figure G3: *Scatterplot for age x previous experience in hours*

What is the relationship between X and Y?

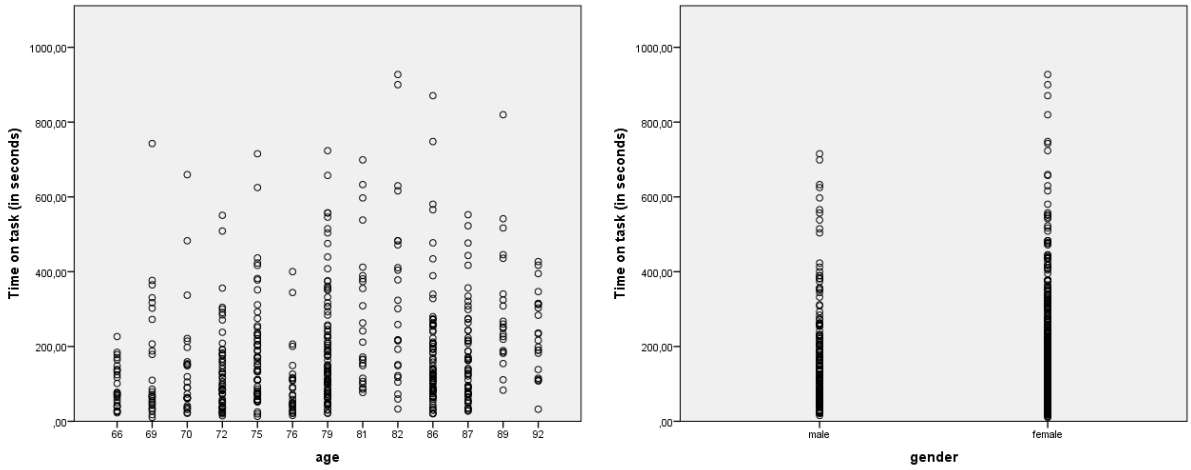
Scatterplots showed that it seems time on task lowers in trial 2 and then rises in trial 3 again (figure G4). There also seems to be a correlation between time on task and both previous experience in hours and age (figures G5 and G6). Time on task lowers with more previous experience and time on task rises with a higher age. Furthermore, the scatterplots seem to show that female participants had a higher time on task than male participants (but this may also be because there were more female participants than male in the first place – also see figure G7). No clear relationship between time on task and experience in compartments could be seen.

Are observations of the response variable independent?

It has to be taken into account that the design used is a repeated measures design, using the same participants over three trials. There may be effects of learning and fatigue as well. An autoregressive correlation matrix can be used to take this into account.



Figures G4 en G5 (left to right): *ToT X trial and ToT X Previous experience in hours*



Figures G6 and G7 (left to right)*: ToT x age and ToT x gender*

3. Final model used

Generalized estimating equations was used. As the time on task were measured in seconds and could only display values between 0 to infinite, it was chosen to use a gamma distribution with a log link. To account for repeated measures and possible effects of fatigue and learning, an autoregressive correlation matrix was used. Because the variables of age and previous experience in hours seemed to be redundant, models using either one of these variables were compared to see which model had a better fit (by comparing the QICC values). As the model with age (Table G1) had a lower

QICC value, it was chosen to use this model over the model using previous experience in hours (Table G2)

Table G1: *GEE results for time on task with age as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	1,534	,5011	1	,000	317,556
Trial 1	0[a]	.	.	.	
Trial 2	-,362	,0635	1	,000	
Trial 3	-,342	,0733	1	,000	
Gender (M)	0[a]	.	.		
Gender (F)	,041	-,187	1	,727	
Age	,013	,001	1	,028	
Previous exp 2	-,112	-,165	1	,000	

a. Set to zero because this parameter is redundant. The other variables are compared to this one

Table G2: *GEE results for time on task with previous experience in hours as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	5,767	,1061	1	,000	320,153
Trial 1	0[a]	.	.	.	
Trial 2	-,360	,0621	1	,000	
Trial 3	-,339	,0702	1	,000	
Gender (M)	0[a]	.	.		
Gender (F)	,023	-,198	1	,839	
Previous exp 1	-,113	-,160	1	,435	
Previous exp 2	-,006	-,022	1	,000	

a. Set to zero because this parameter is redundant. The other variables are compared to this one

GEE 2: Total number of problems encountered

1. **Situation and variables**

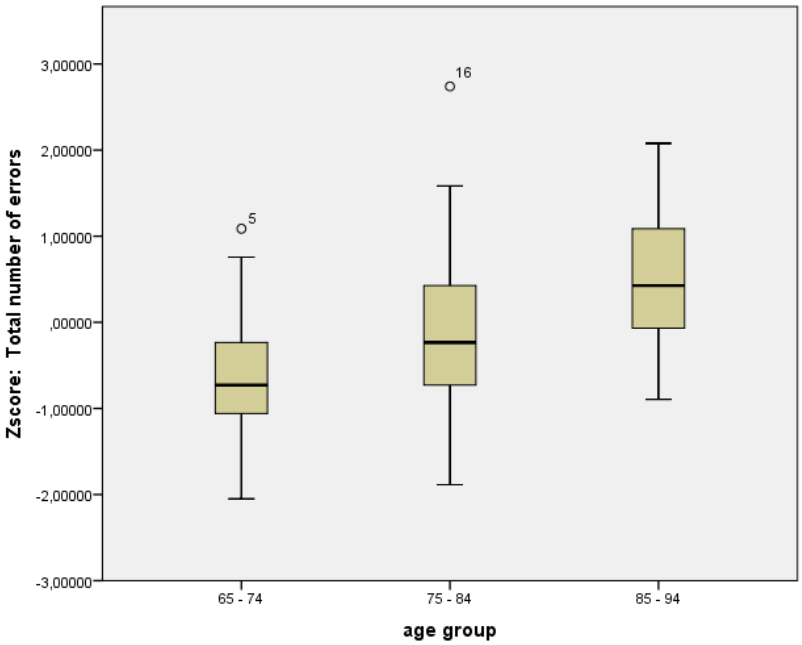
   The prediction is that experience influences the number of problems encountered (μ = 15.42, σ = 6,054, σ^2 = 36,654). We therefore have 2 variables based on previous experience, one based on hours of using a computer, laptop, tablet or smart TV on average each month. The second score is based on usage or owning one of these devices, and performing certain acts (e.g. sending an e-mail or surfing the web). Other possible predictors, or covariates of number of problems encountered are age and gender.

2. Data exploration

Are there outliers?

The data was first checked for impossible values, but no such values were found. As not all tasks were completed by every participant, some time on task values and ASQ scores were missing. These were scored as missing values in SPSS. The residual score from average ASQ was used to check for outliers.

Figure G8*: residual boxplot for total number of problems*



Is there homoscedasticity?

Viewing the boxplot (figure G8) shows that the variance of the residuals are almost equally distributed. Still, the variance within age group 85 – 94 seems to differ and thus shows heteroscedasticity,

Is there a normal distribution?

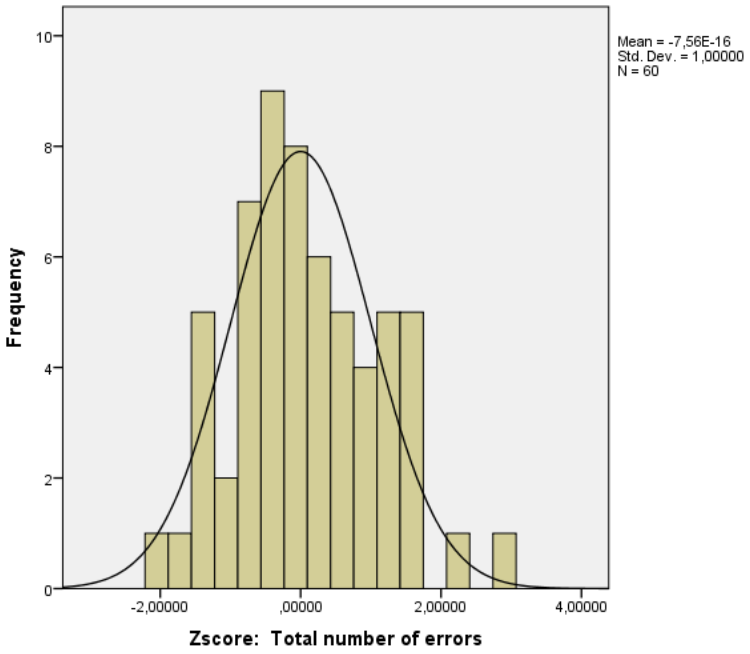
A residual histogram (figure G9) was created for the total number of errors. Comparing the histogram with the line of a normal distribution through it shows that the data has a positive skew. Therefore, a normal distribution cannot be assumed.

Are there a lot of zeros in the data?

This has to be considered for all X variables. Logically, age has no zero's in this group of participants. The first experience score is based on hours, a count score, and there are four

zero's present. The data also show that for the second experience score, there are eight zero's as well, with an overlap of the four zero's found in the first experience score and the second experience score. For this reason, it is also plausible to consider using a zero-inflated model and at least compare it to other possible models used

Figure G9*: residual histogram for total number of problems*



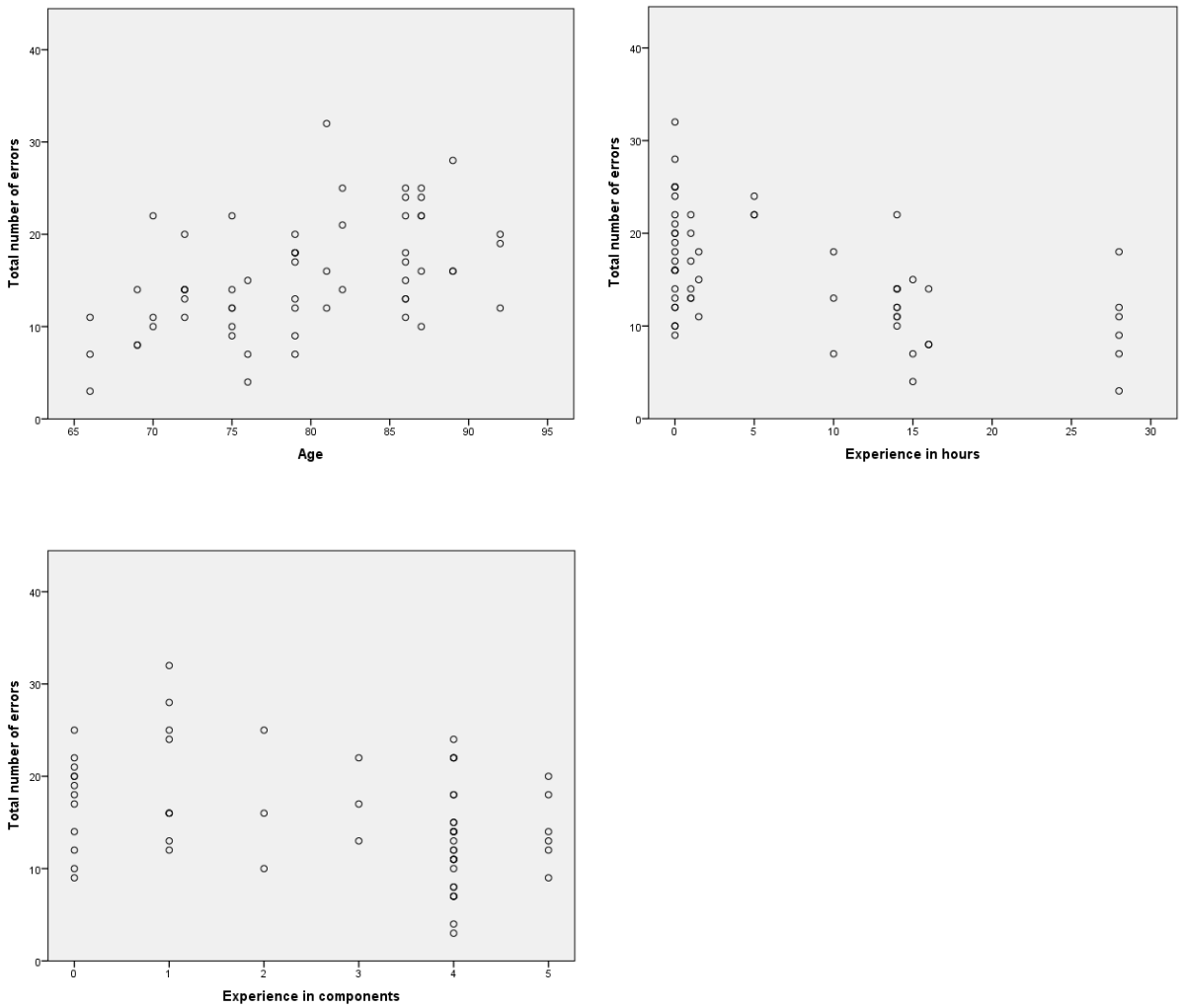
Is there collinearity?

Scatter plots show that age and exp 1 seem redundant: they correlate with each other (see scatterplot G3). Models with either one variable will have to be compared in order to find out which variable fits the data best. No collinearity was found between either age and the score of experience based on hours nor on parts.

What are the relationships between X and Y?

Scatterplots showed that the number of problems seem to have some correlation with age (figure G10) and also experience in with hours of practice (figure G11). The correlation of number of problems with exp 2 (Figure G12) seems unclear. It is possible that there is a slight correlation between number of problems and experience in components, but it is also very well possible that experience in components is not a good predictor of the number of problems.

How do elderly people learn to work with an online system?



Figures G10, G11 and G12 (left to right, top to bottom): *Correlations of the X variables*

Are observations of the response variable independent?

As the total number of problems encountered consist of the sum of problems as seen over the course of three trials of measuring (repeated measures), it is possible that there are effects of learning or perhaps because of fatique. These effects can be accounted for by using autoregressive modelling.

3. Final model used

It was chosen to use Generalized Estimating Equations. The data for 'number of problems' was count data, but as the assumptions for using a Poisson distribution could not be met (e.g. overdispersion, or $\sigma^2 \geq \mu$), it was chosen to use a negative binomial distribution with a log link. An autoregressive working correlation matrix was used to take into accounts the effects of repeated measures as well as possible

effects of learning or fatigue stemming from it. As the covariates age and experience in hours (exp1) seemed to correlate strongly, two models were plotted each using one of the covariates. The QICC values showed that the model in which exp1 was used was a better fit than the one using age as a covariate, therefore age was not used as a covariate.

Table G3: *GEE results for total number of problems with age as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	1,634	0,9394	1	0,082	15,173
Age	0,015	0,0113	1	0,181	
Exp2	-0,293	0,3334	1	0,379	
Age * Exp2	0,003	0,0041	1	0,437	

Table G4: *GEE results for total number of problems with previous experience in hours as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	2,892	0,0526	1	0,000	14,504
Exp1	-0,112	0,0162	1	0,000	
Exp2	0,008	0,0305	1	0,780	
Exp1 * Exp2	0,020	0,0031	1	0,000	

GEE 3: Number of problems encountered per type

1. **Situation and variables**

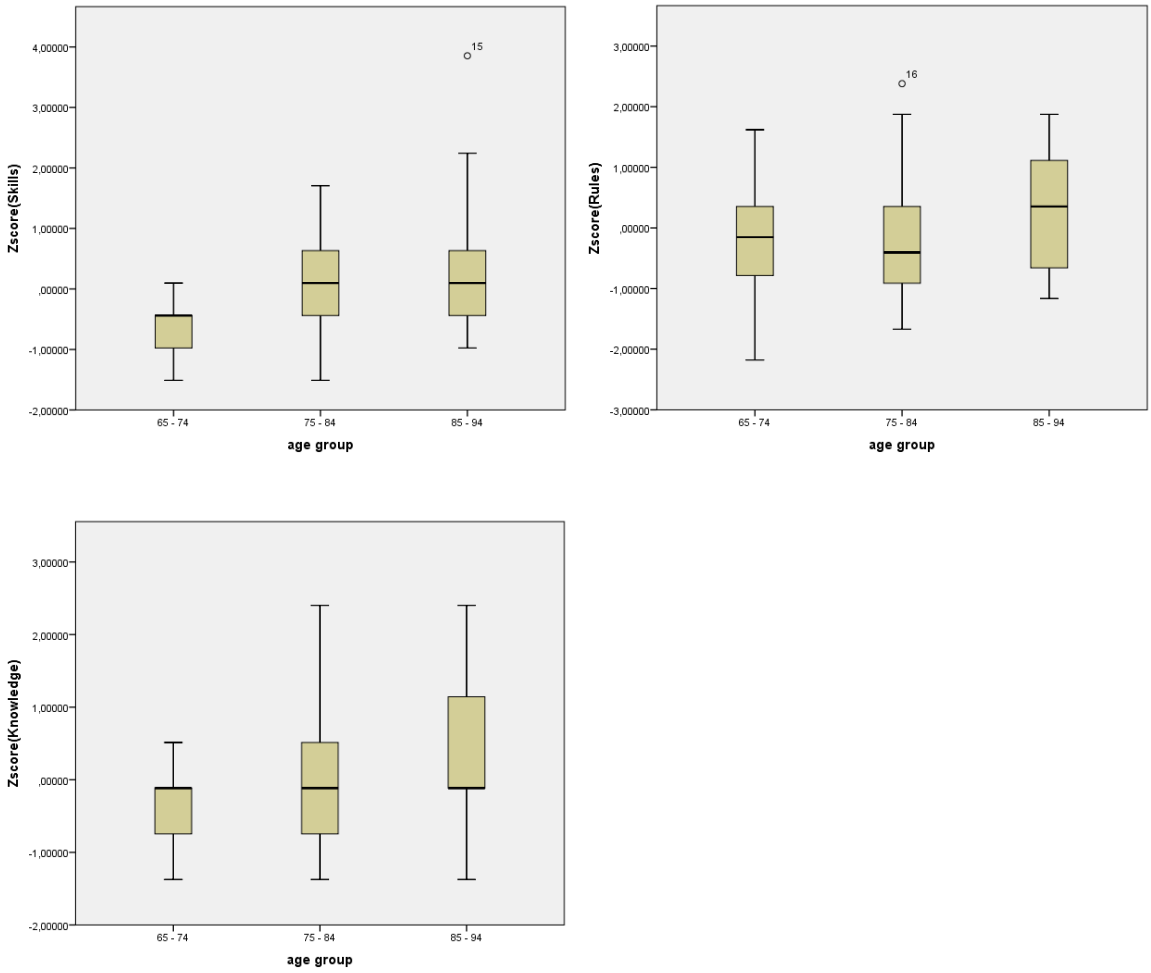
   The prediction is that experience influences the number of problems encountered from different types: Skills ($\mu = 3{,}82$, $\sigma = 1{,}864$, $\sigma^2 = 3{,}474$), Rules ($\mu = 8{,}6$, $\sigma = 3{,}95$, $\sigma^2 = 15{,}6$) and Knowledge ($\mu = 2{,}18$, $\sigma = 1{,}589$, $\sigma^2 = 2{,}525$).

   For each type of problem, we therefore have 2 variables based on previous experience that will serve as predictors, one based on hours of using a computer, laptop, tablet or smart TV on average each month. The second score is based on usage or owning one of these devices, and performing certain acts (e.g. sending an e-mail or surfing the web). Other possible predictors, or covariates of number of problems made are age and gender.

2. Data exploration

Are there outliers?

The data was checked for the Z-scores of each variable:



Figures G13, G14 and G15 (left to right, top to bottom)*: residual boxplots for each type of problem*

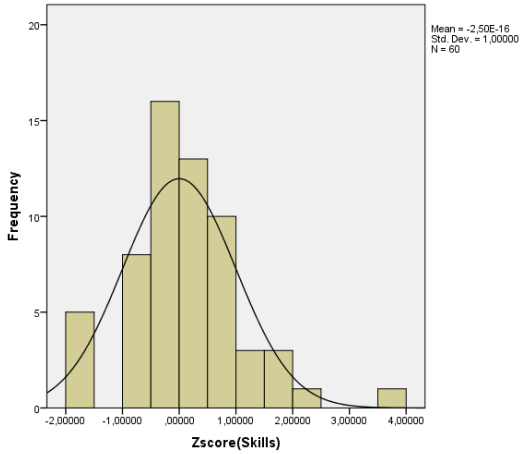
As the boxplots show, both the skill-set and the rule-set of problems have one (different) outlier (G13 and G14). Checking the data and the videos showed that both outliers had no impossible values; The outlier for the rules-set belonged to a participant who overall encountered more problems than average (so, also in the other categories), therefore the outlier was left in the dataset. The outlier in the skill-set belonged to a participant who overall did not encounter a lot more problems than average. As the outlier was relatively high as well, it was chosen to do the analysis for this dataset both with and without the outlier to see whether there were any differences based purely on this outlier.

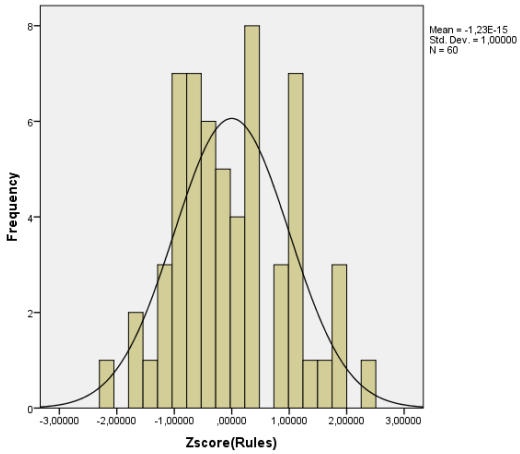
Is there homoscedasticity?

None of the three boxplots show signs of homoscedasticity.

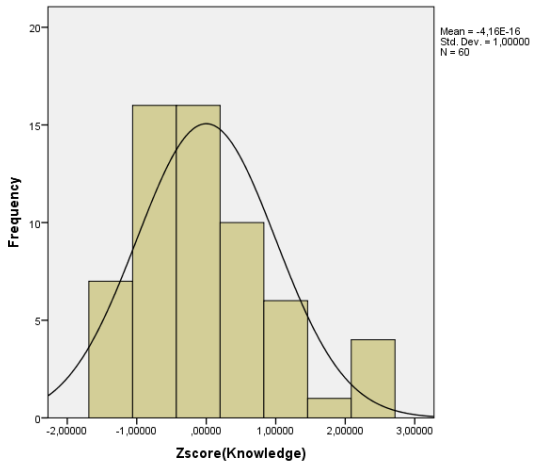
Is there a normal distribution?

Residual histograms were made for all three datasets:





The datasets for both skill problems and knowledge problems show a positive skew (G16 and G18). The rules dataset does not show this positive skew, but rather a slight negative skew (G17). For none of the datasets, normality can be assumed.



Figures G16, G17 and G18 (left to right, top to bottom): *residual histograms for each type of problem*

Are there a lot of zeros in the data?

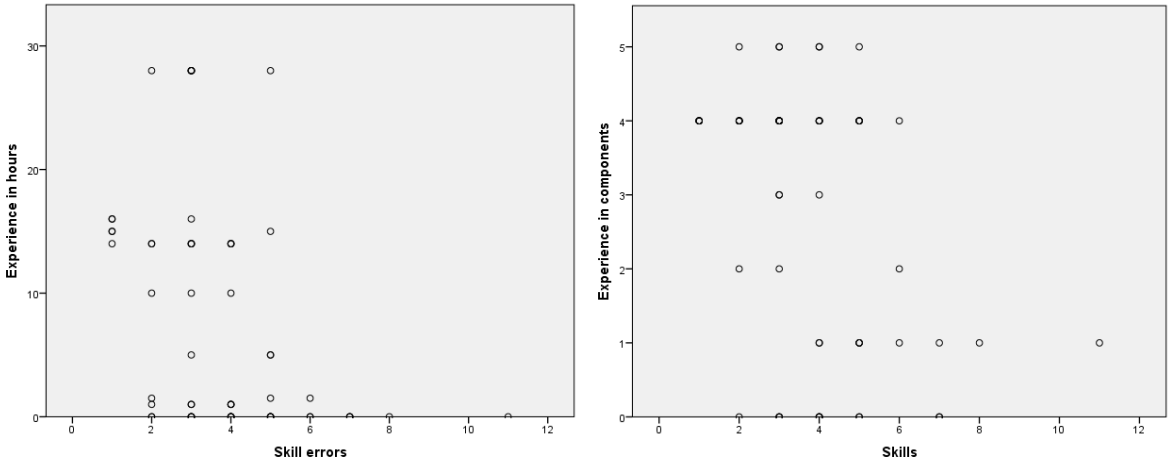
As the same predictors are used here as in the other analysis, it has to be taken into account that in this set as well, there are some zero's detected, and thus this might have to be taken into account when picking a model.

Is there collinearity?

As seen in the assumption checklist for the dataset with all problems, age and experience as measured in hours seem to be correlated. The variable age will therefore be used with caution. However, age might have an influence on skill-based problems in particular, as skill-based problems may also be caused by age-related decline of certain bodily functions (e.g. holding a firm grip on the remote control, or being able to see the screen properly). For the dataset of skill problems, analyses will be performed with age as a predictor as well, to find out whether it improves the fit of the model or not.

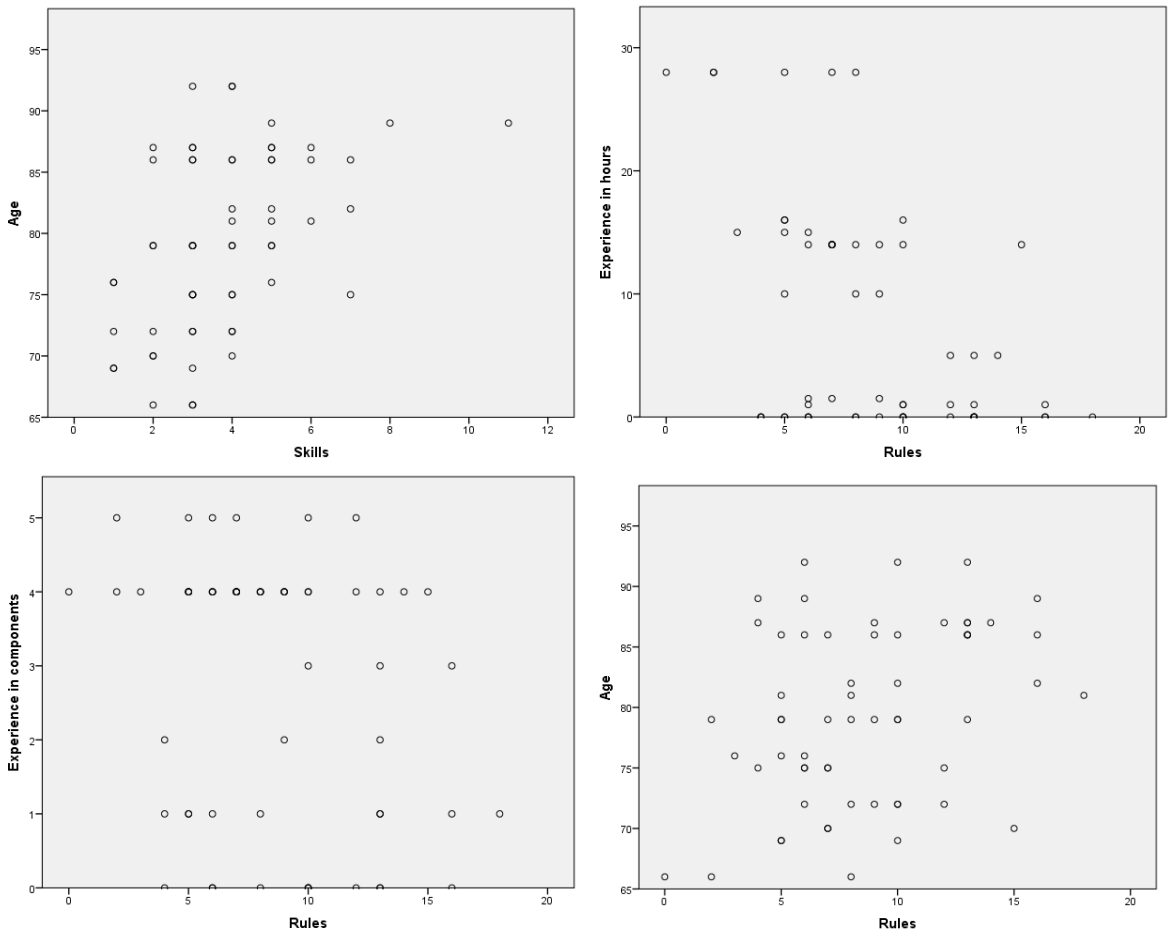
What are the relationships between X and Y?

Scatterplots were made for each dataset and for each Y variable. As an extra control, age was used as a variable as well for each dataset.



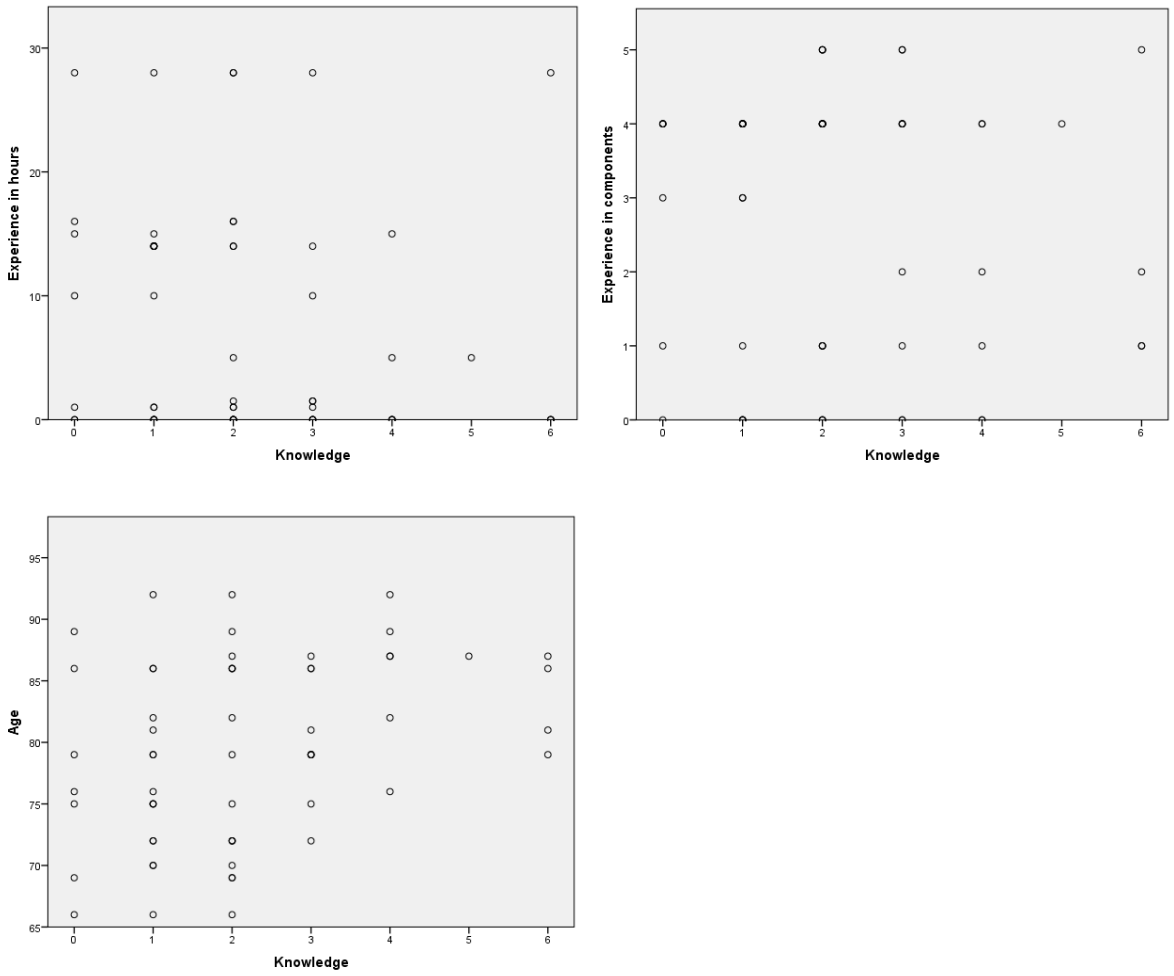
Figures G19 and G20 (left to right): *scatterplots for skills and experience*

The scatterplots for the skills dataset show that there does not seem to be a great relationship between experience in hours and the number of skill problems made (G19). There does seem to be a slight relationship between the experience with different components and the number of problems on the skill level (G20): as the experience level rises, it seems the number of skill problems slightly lessen. As can be seen below (upper left panel), it seems that there might be a relationship between age and number of problems on the skill level. However, the outlier found should be taken into account.



Figures G21, G22, G23, G24 (left to right, top to bottom): scatterplots *for each X variable on the rules type of problems*

The scatterplots from the rules dataset (G21, 22, 23 and 24) show that there is relationship between hours of previous experience and number of problems of the rules kind. There seems to be a possible kind of relationship with experience in components, but less clear. No distinctive relationship pattern can be detected between rule problems and age. Last, the scatterplots for the knowledge dataset (G25, 26, 27) shows a slight possible relationship between number of problems and hours of practice. Furthermore, there does not seem to be any relationship between number of problems and experience in subparts. Last, age seems to have a relationship with number of problems as well, showing that the number of problems increases with age.



Figures G25, G26 and G27 (left to right, top to bottom): Scatterplots *of all X variables on the number of knowledge problems*

Are observations of the response variable independent?

All three datasets consist of the sum of problems over the course of three trials, or a repeated measures. Therefore, there are possible effects of learning or fatigue. These effects need to be accounted for, for example by using autoregressive modeling.

3. Final models used

All datasets contained count data. For the datasets of rules and knowledge problems, it was chosen to use a model with a negative binomial distribution with a log link to handle overdispersion. An autoregressive working correlation was used because of the repeated measures, combined with possible learning and/or fatigue effects.

For both datasets, models with either age or experience in hours were set up together with experience in components to see which would be a better fit. For the rules

dataset, the QICC value for the model using experience in hours as a predictor (table G5) was higher than the value for the model using age (table G6), therefore the former was chosen.

Table G5: GEE *results for total number of rules problems with experience in hours as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	2,265	0,0613	1	0,000	20,190
Exp1	-0,038	0,0301	1	0,211	
Exp2	0,038	0,0364	1	0,298	
Exp1 * Exp2	0,001	0,0057	1	0,891	

Table G6: GEE *results for total number of rules problems with age as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	1,661	1,0256	1	,105	21,914
Age	,008	,0121	1	,518	
Exp2	-,315	,3876	1	,416	
Age * Exp2	,003	,0048	1	,474	

For the knowledge set, the same comparison was made. Here, the model using experience in hours (table G7) as a predictor had the lowest QICC value compared to the model using age (table G8) and was therefore the model discussed in the results section of this study.

Table G7: GEE *results for total number of knowledge problems with experience in hours as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	,851	,1279	1	,000	33.219
Exp1	-,248	,0231	1	,000	
Exp2	,043	,0539	1	,428	
Exp1 * Exp2	,051	,0039	1	,000	

Table G8: GEE *results for total number of knowledge problems with age as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	-2,646	1,8129	1	,144	33.934
Age	,041	,0214	1	,055	
Exp2	-,055	,5159	1	,915	
Age * Exp2	,001	,0063	1	,840	

The skill problems dataset had underdispersion, a rare condition in which the variance is smaller than the mean. Because of this, it is normally not adviced to use a normal Poisson distribution as this would lead to overestimating the standard error, increasing the chance of finding that predictors are seen as non-significant while they actually are. One option is to use a Maxwell-Conway (COM) Poisson which adds a parameter that can be used to account for both under- and overdispersion (Shmueli et al., 2005). Still, because our data was taken over multiple trials and had possibilities of learning and fatigue effects (which could not be taken into account properly when using a COM-Poisson) and because the underdispersion was very small, it was chosen to use a Poisson distribution with log linear nonetheless. This Poisson distributed model had a autoregressive working correlation matrix to take into account the above effects. Output from comparing two models was the following:

Table G9: GEE *results for total number of skill problems with experience in hours as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	1,598	,1087	1	,000	45,763
Exp1	-,120	,0343	1	,000	
Exp2	-,050	,0315	1	,110	
Exp1 * Exp2	,024	,0065	1	,000	

Table G10: GEE *results for total number of skill problems with age as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	,406	1,6538	1	,806	44,949
Age	,014	,0208	1	,502	
Exp2	-,374	,4443	1	,400	
Age * Exp2	,004	,0056	1	,494	

Again, two models were compared. Even though the first model (table G9), using experience in hours as a predictor, has a significant intercept, main effect and significant interaction effect, the fit from the model using age as a predictor (table G10) with no significant effects whatsoever was a better fit for the data (lower QICC value). It was therefore chosen to use the second model instead.

GEE 4: ASQ scores

1. **Situation and variables**

   The prediction is that the ASQ scores will heighten over time, showing that users are satisfied with the system. ASQ is expected to be higher when time on task is lower. Furthermore, effects are expected from age and gender. We have the variable average ASQ ($\mu = 5{,}05$, $\sigma = 1{,}775$, $\sigma^2 = 3{,}151$), taken as an average of the three ASQ questions to serve as a dependent variable and time on task, age, gender and previous experience as independent variables.

2. **Data exploration**

   Are there outliers?

   The data was first checked for impossible values, but no such values were found and therefore the entire set could be used. Residual scores -boxplots were used for visualization.
   To have a proper variable for the X-axis, a new variable for 'age' was created, dividing it in three classes. This created the boxplot as seen below (figure G28). Two outliers were detected. Looking back into the data showed that these values were not impossible, hence they were not removed from the dataset. The dataset showed no outliers.

   Is there homoscedasticity?

   As is visible in the boxplots, there were no signs of homoscedasticity

   Is there a normal distribution?

   A residual histogram was made for the average ASQ scores (figure G29). This histogram showed that there was no normal distribution. The distribution has a negative skew and seems to resemble a gamma distribution.

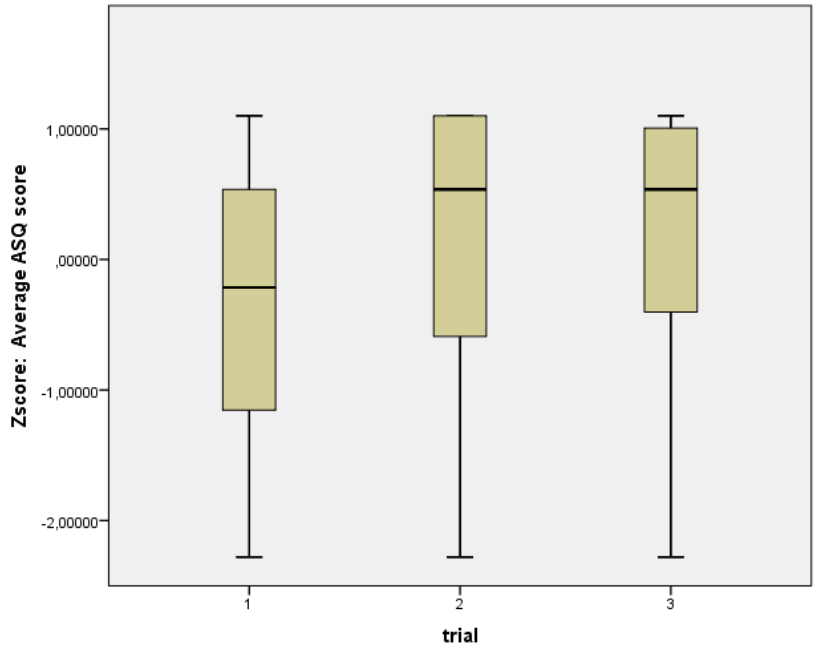


Figure G28: *residual boxplots for the average ASQ scores*

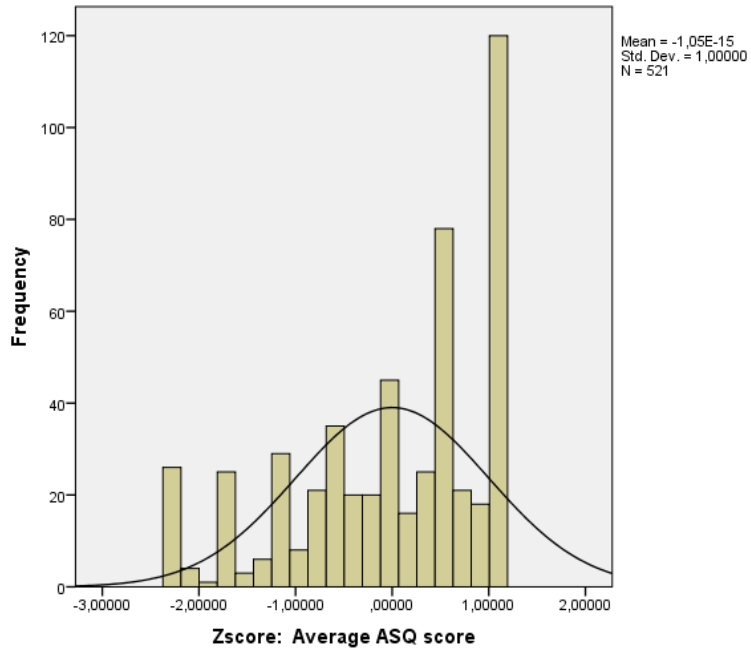


Figure G29: *residual histogram for average ASQ scores*

Are there a lot of zeros in the data?

This has to be taken into account for all X variables. Some participants left the ASQ questions blank when they had not performed a task. These values were treated as missing values, as the lowest value possible within the ASQ rating was a 1. For the scores of experience, there are

some zero's. This makes using a zero-inflated model plausible. Age and gender logically do not have any zero's.

Is there collinearity?

Scatter plots showed a possible collinearity effect for age and experience measured in hours (figure G30). They seem to correlate with each other. Models with either one variable will have to be compared in order to find out which variable fits the data best. No collinearity was found between either age and the score of experience based on hours nor on parts.

What are the relationships between X and Y?

Scatterplots showed that the average ASQ score and time on task seem to be correlated to one another (figure G31). There also seemed to be a slight correlation between previous the average ASQ scoires and experience in hours (figure G32) and with age (figure G33). There do not seem to be any directly visible correlation between the average ASQ score and previous experience in components, age or gender (figures g34 and G35).

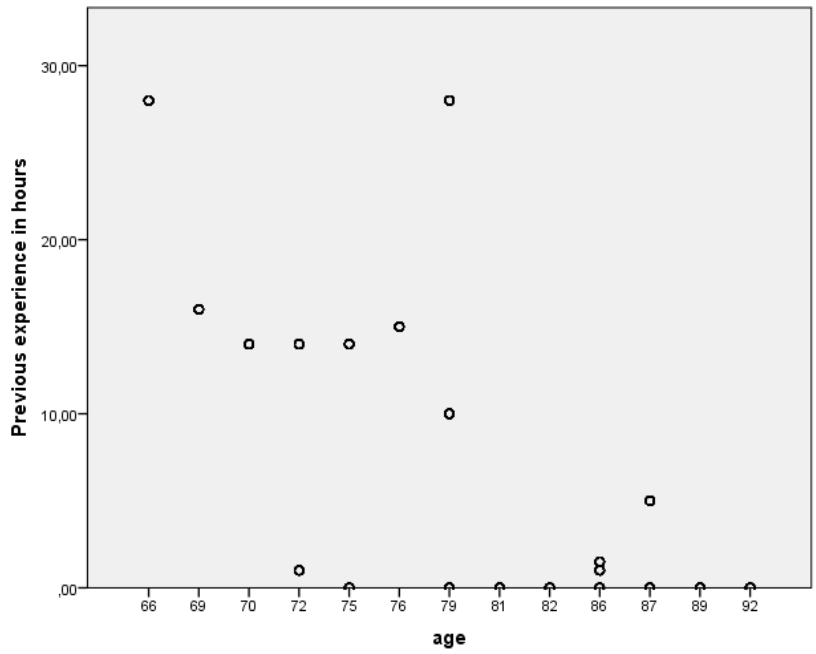
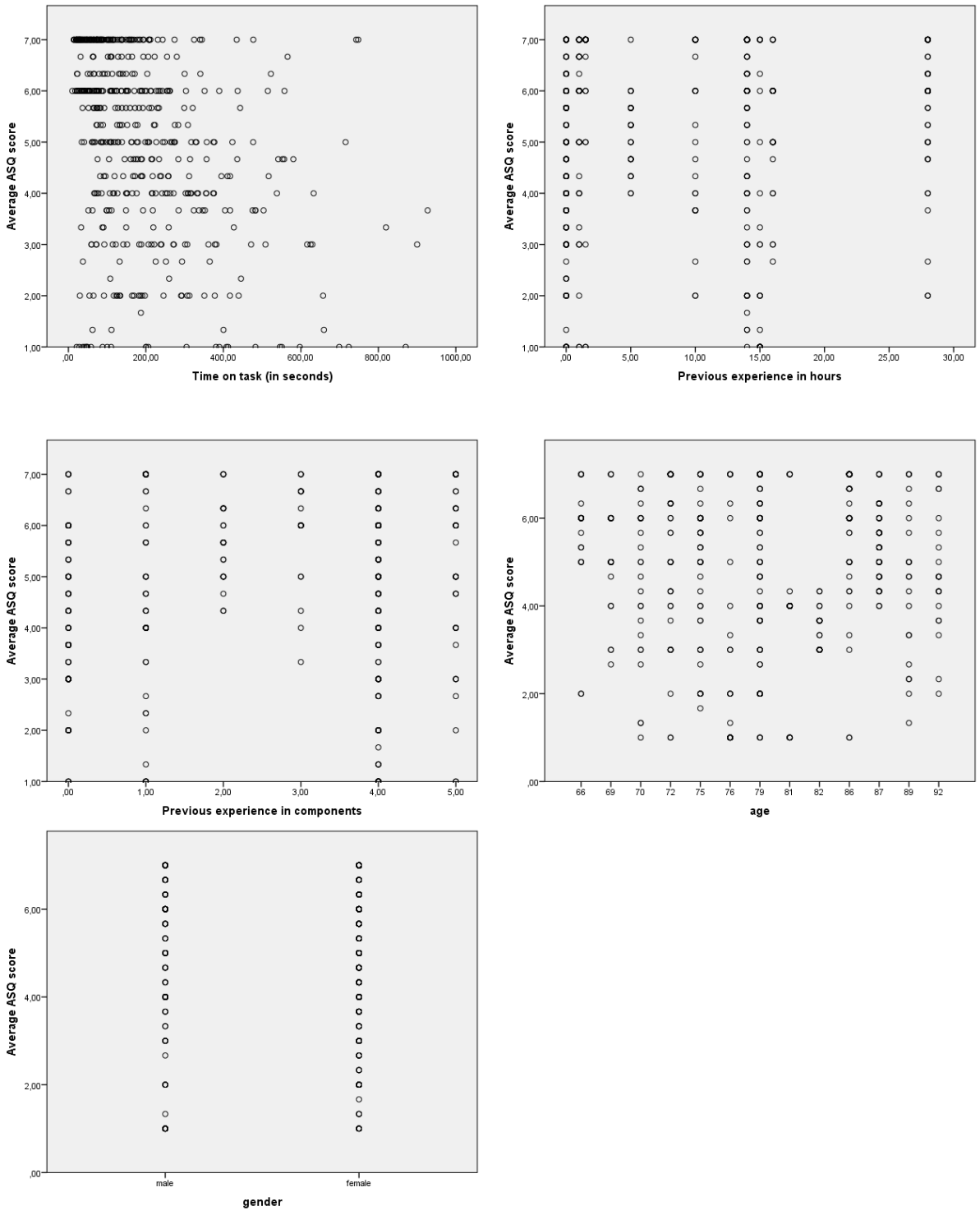


Figure G30: *scatterplot for previous experience in hours and age*



Figures G31, G32, G33, G34 and G35 (left to right, top to bottom) ; *Scatterplots for all X variables on average ASQ scores*

Are observations of the response variable independent?

The average ASQ scores consist of the scores taken from the three ASQ questions, which have been answered by the same 20 participants after each of the nine tasks, in 3 trials. There are possibly effects of learning or fatigue that may influence how participants rated the ASQ. Such effects can be accounted for by using autoregressive modeling.

3. Final model used

Generalized Estimating Equations was used. As the data could only have values from positive to infinity, and the residual distribution had a negative skew, it was chosen to use a gamma distribution with a log link. To take into account the effects of repeated measures, including possible learning or fatigue effects, an autoregressive working correlation matrix was used. As there seemed to be a correlation between age and previous experience in hours, models were compared using either one of the two factors. As the model using age (table G11) as a predictor was a better fit, it was chosen to use age in the model instead of previous experience in hours (table G12).

Table G11: GEE *results for average ASQ scores with previous experience in hours as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	1,534	,1116	1	,000	100,464
Trial 1	0[a]	.	.	.	
Trial 2	,161	,0382	1	,000	
Trial 3	,156	,0534	1	,003	
Gender (M)	0[a]	.	.		
Gender (F)	,039	,1228	1	,751	
Time on Task	-,001	,0002	1	,000	
Previous exp 1	-,004	,0052	1	,496	
Previous exp 2	,035	,0321	1	,281	

a. Set to zero because this parameter is redundant. This is the parameter to which the others are compared.

Table G12: GEE *results for average ASQ scores with age as a predictor*

Parameter	B	Standard error	Degrees of freedom	Significance	QICC value
Intercept	0,693	,4637	1	,135	98,758
Trial 1	0[a]	.	.	.	

Trial 2	,165	,0092	1	,000
Trial 3	,162	,0059	1	,002
Gender (M)	0[a]	.	.	
Gender (F)	,027	,1228	1	,807
Time on Task	-,001	,0002	1	,000
Age	,010	,0052	1	,047
Previous exp 2	,041	,0321	1	,142

a. Set to zero because this parameter is redundant. This is the parameter to which the others are compared.

Appendix H – SPSS output

GEE 1: Time on task

```
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=trial ZToT MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: trial=col(source(s), name("trial"), unit.category())
  DATA: ZToT=col(source(s), name("ZToT"))
  DATA: id=col(source(s), name("$CASENUM"), unit.category())
  GUIDE: axis(dim(1), label("trial"))
  GUIDE: axis(dim(2), label("Zscore:  Time on task (in seconds)"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: schema(position(bin.quantile.letter(trial*ZToT)), label(id))
END GPL.

GRAPH
  /HISTOGRAM(NORMAL)=ZToT.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp1 age
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp1=col(source(s), name("prev_exp1"))
  DATA: age=col(source(s), name("age"), unit.category())
  GUIDE: axis(dim(1), label("Previous experience in hours"))
  GUIDE: axis(dim(2), label("age"))
  ELEMENT: point(position(prev_exp1*age))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp2 age
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp2=col(source(s), name("prev_exp2"))
  DATA: age=col(source(s), name("age"), unit.category())
  GUIDE: axis(dim(1), label("Previous experience in components"))
  GUIDE: axis(dim(2), label("age"))
  ELEMENT: point(position(prev_exp2*age))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp1 gender
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
```

```
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp1=col(source(s), name("prev_exp1"))
  DATA: gender=col(source(s), name("gender"), unit.category())
  GUIDE: axis(dim(1), label("Previous experience in hours"))
  GUIDE: axis(dim(2), label("gender"))
  SCALE: cat(dim(2), include("1", "2"))
  ELEMENT: point(position(prev_exp1*gender))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp2 gender
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp2=col(source(s), name("prev_exp2"))
  DATA: gender=col(source(s), name("gender"), unit.category())
  GUIDE: axis(dim(1), label("Previous experience in components"))
  GUIDE: axis(dim(2), label("gender"))
  SCALE: cat(dim(2), include("1", "2"))
  ELEMENT: point(position(prev_exp2*gender))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp2 prev_exp1
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp2=col(source(s), name("prev_exp2"))
  DATA: prev_exp1=col(source(s), name("prev_exp1"))
  GUIDE: axis(dim(1), label("Previous experience in components"))
  GUIDE: axis(dim(2), label("Previous experience in hours"))
  ELEMENT: point(position(prev_exp2*prev_exp1))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=trial ToT MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: trial=col(source(s), name("trial"), unit.category())
  DATA: ToT=col(source(s), name("ToT"))
  GUIDE: axis(dim(1), label("trial"))
  GUIDE: axis(dim(2), label("Time on task (in seconds)"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: point(position(trial*ToT))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp1 ToT
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
```

```
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp1=col(source(s), name("prev_exp1"))
  DATA: ToT=col(source(s), name("ToT"))
  GUIDE: axis(dim(1), label("Previous experience in hours"))
  GUIDE: axis(dim(2), label("Time on task (in seconds)"))
  ELEMENT: point(position(prev_exp1*ToT))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp2 ToT
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp2=col(source(s), name("prev_exp2"))
  DATA: ToT=col(source(s), name("ToT"))
  GUIDE: axis(dim(1), label("Previous experience in components"))
  GUIDE: axis(dim(2), label("Time on task (in seconds)"))
  ELEMENT: point(position(prev_exp2*ToT))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=age ToT MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: age=col(source(s), name("age"), unit.category())
  DATA: ToT=col(source(s), name("ToT"))
  GUIDE: axis(dim(1), label("age"))
  GUIDE: axis(dim(2), label("Time on task (in seconds)"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: point(position(age*ToT))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=gender ToT MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: gender=col(source(s), name("gender"), unit.category())
  DATA: ToT=col(source(s), name("ToT"))
  GUIDE: axis(dim(1), label("gender"))
  GUIDE: axis(dim(2), label("Time on task (in seconds)"))
  SCALE: cat(dim(1), include("1", "2"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: point(position(gender*ToT))
END GPL.


* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=trial ZToT MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
```

```
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: trial=col(source(s), name("trial"), unit.category())
  DATA: ZToT=col(source(s), name("ZToT"))
  DATA: id=col(source(s), name("$CASENUM"), unit.category())
  GUIDE: axis(dim(1), label("trial"))
  GUIDE: axis(dim(2), label("Zscore:  Time on task (in seconds)"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: schema(position(bin.quantile.letter(trial*ZToT)), label(id))
END GPL.

GRAPH
  /HISTOGRAM(NORMAL)=ZToT.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp1 age
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp1=col(source(s), name("prev_exp1"))
  DATA: age=col(source(s), name("age"), unit.category())
  GUIDE: axis(dim(1), label("Previous experience in hours"))
  GUIDE: axis(dim(2), label("age"))
  ELEMENT: point(position(prev_exp1*age))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp2 age
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp2=col(source(s), name("prev_exp2"))
  DATA: age=col(source(s), name("age"), unit.category())
  GUIDE: axis(dim(1), label("Previous experience in components"))
  GUIDE: axis(dim(2), label("age"))
  ELEMENT: point(position(prev_exp2*age))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp1 gender
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp1=col(source(s), name("prev_exp1"))
  DATA: gender=col(source(s), name("gender"), unit.category())
  GUIDE: axis(dim(1), label("Previous experience in hours"))
  GUIDE: axis(dim(2), label("gender"))
  SCALE: cat(dim(2), include("1", "2"))
  ELEMENT: point(position(prev_exp1*gender))
END GPL.

* Chart Builder.
GGRAPH
```

```
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp2 gender
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp2=col(source(s), name("prev_exp2"))
  DATA: gender=col(source(s), name("gender"), unit.category())
  GUIDE: axis(dim(1), label("Previous experience in components"))
  GUIDE: axis(dim(2), label("gender"))
  SCALE: cat(dim(2), include("1", "2"))
  ELEMENT: point(position(prev_exp2*gender))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp2 prev_exp1
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp2=col(source(s), name("prev_exp2"))
  DATA: prev_exp1=col(source(s), name("prev_exp1"))
  GUIDE: axis(dim(1), label("Previous experience in components"))
  GUIDE: axis(dim(2), label("Previous experience in hours"))
  ELEMENT: point(position(prev_exp2*prev_exp1))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=trial ToT MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: trial=col(source(s), name("trial"), unit.category())
  DATA: ToT=col(source(s), name("ToT"))
  GUIDE: axis(dim(1), label("trial"))
  GUIDE: axis(dim(2), label("Time on task (in seconds)"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: point(position(trial*ToT))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp1 ToT
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp1=col(source(s), name("prev_exp1"))
  DATA: ToT=col(source(s), name("ToT"))
  GUIDE: axis(dim(1), label("Previous experience in hours"))
  GUIDE: axis(dim(2), label("Time on task (in seconds)"))
  ELEMENT: point(position(prev_exp1*ToT))
END GPL.

* Chart Builder.
GGRAPH
```

```
  /GRAPHDATASET NAME="graphdataset" VARIABLES=prev_exp2 ToT
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: prev_exp2=col(source(s), name("prev_exp2"))
  DATA: ToT=col(source(s), name("ToT"))
  GUIDE: axis(dim(1), label("Previous experience in components"))
  GUIDE: axis(dim(2), label("Time on task (in seconds)"))
  ELEMENT: point(position(prev_exp2*ToT))
END GPL.


* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=age ToT MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: age=col(source(s), name("age"), unit.category())
  DATA: ToT=col(source(s), name("ToT"))
  GUIDE: axis(dim(1), label("age"))
  GUIDE: axis(dim(2), label("Time on task (in seconds)"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: point(position(age*ToT))
END GPL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=gender ToT MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: gender=col(source(s), name("gender"), unit.category())
  DATA: ToT=col(source(s), name("ToT"))
  GUIDE: axis(dim(1), label("gender"))
  GUIDE: axis(dim(2), label("Time on task (in seconds)"))
  SCALE: cat(dim(1), include("1", "2"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: point(position(gender*ToT))
END GPL.

* Generalized Estimating Equations.
GENLIN ToT BY trial gender Task (ORDER=DESCENDING) WITH age prev_exp1
prev_exp2
  /MODEL trial gender age prev_exp2 INTERCEPT=YES
 DISTRIBUTION=GAMMA LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=MLE MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.

* Generalized Estimating Equations.
```

```
GENLIN ToT BY trial gender Task (ORDER=DESCENDING) WITH age prev_exp1
prev_exp2
  /MODEL trial gender prev_exp2 prev_exp1 INTERCEPT=YES
 DISTRIBUTION=GAMMA LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=MLE MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.


* Generalized Estimating Equations (for interaction effects).
GENLIN ToT BY trial gender Task (ORDER=DESCENDING) WITH age prev_exp1
prev_exp2
  /MODEL trial gender age prev_exp1 trial*prev_exp1 prev_exp2
trial*prev_exp2 INTERCEPT=YES
 DISTRIBUTION=GAMMA LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=MLE MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

GEE 2: total number of problems

```
* Generalized Estimating Equations.
GENLIN Total_Errors BY Gender (ORDER=ASCENDING) WITH Age Exp1 Exp2
  /MODEL Age Exp2 Age*Exp2 INTERCEPT=YES
 DISTRIBUTION=NEGBIN(1) LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.

* Generalized Estimating Equations.
GENLIN Total_Errors BY Gender (ORDER=ASCENDING) WITH Age Exp1 Exp2
  /MODEL Exp1 Exp2 Exp1*Exp2 INTERCEPT=YES
 DISTRIBUTION=NEGBIN(1) LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

GEE 3: Number of problems per error type

```
* Generalized Estimating Equations.
GENLIN Rules BY Gender (ORDER=ASCENDING) WITH Age Exp1 Exp2
```

```
  /MODEL Exp1 Exp2 Exp1*Exp2 INTERCEPT=YES
 DISTRIBUTION=NEGBIN(1) LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.

* Generalized Estimating Equations.
GENLIN Rules BY Gender (ORDER=ASCENDING) WITH Age Exp1 Exp2
  /MODEL Age Exp2 Age*Exp2 INTERCEPT=YES
 DISTRIBUTION=NEGBIN(1) LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.

* Generalized Estimating Equations.
GENLIN Knowledge BY Gender (ORDER=ASCENDING) WITH Age Exp1 Exp2
  /MODEL Exp1 Exp2 Exp1*Exp2 INTERCEPT=YES
 DISTRIBUTION=NEGBIN(1) LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.

* Generalized Estimating Equations.
GENLIN Knowledge BY Gender (ORDER=ASCENDING) WITH Age Exp1 Exp2
  /MODEL Age Exp2 Age*Exp2 INTERCEPT=YES
 DISTRIBUTION=NEGBIN(1) LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.

* Generalized Estimating Equations.
GENLIN Skills BY Gender (ORDER=ASCENDING) WITH Age Exp1 Exp2
  /MODEL Age Exp2 Age*Exp2 INTERCEPT=YES
 DISTRIBUTION=POISSON LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.

* Generalized Estimating Equations.
```

```
GENLIN Skills BY Gender (ORDER=ASCENDING) WITH Age Exp1 Exp2
  /MODEL Exp1 Exp2 Exp1*Exp2 INTERCEPT=YES
 DISTRIBUTION=POISSON LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

GEE 4: ASQ scores

```
* Generalized Estimating Equations.
GENLIN ASQaverage BY trial Task gender (ORDER=DESCENDING) WITH age ToT
prev_exp1 prev_exp2
  /MODEL trial ToT gender prev_exp1 prev_exp2 INTERCEPT=YES
 DISTRIBUTION=GAMMA LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=MLE MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.

* Generalized Estimating Equations.
GENLIN ASQaverage BY trial Task gender (ORDER=DESCENDING) WITH age ToT
prev_exp1 prev_exp2
  /MODEL trial ToT gender age prev_exp2 INTERCEPT=YES
 DISTRIBUTION=GAMMA LINK=LOG
  /CRITERIA METHOD=FISHER(1) SCALE=MLE MAXITERATIONS=100 MAXSTEPHALVING=5
PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95
LIKELIHOOD=FULL
  /REPEATED SUBJECT=participant SORT=YES CORRTYPE=AR(1) ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

Appendix I – Remote control