Aalto University
School of Science

Master's Programme in Computer, Communication and Information Sciences

Testing usability metrics with underserved populations

Matthew Hallonbacka

Master's Thesis
2024

Aalto University
School of Science

Author Matthew Hallonbacka

Title Testing usability metrics with underserved populations

Degree programme Computer, Communication and Information Sciences

Major Software and Service Engineering

Supervisor and advisor Sari Kujala, PhD.

Date 18 March 2023 **Number of pages** 55+12 **Language** English

Abstract

Digital services underlie many aspects of everyday life, yet they are often not designed to meet the needs of everyone who may need to use them. Optimizing for usability, especially with historically underserved groups, ensures that as many users as possible can benefit from these services.

Techniques such as participants thinking aloud in laboratory-based usability testing aim to understand the issues affecting the experiences of users, while self-administered questionnaires, like the System Usability Scale (SUS) provide users with a way to articulate and provide quantitative feedback about their experience. It is not yet clear to what extent these methods are suitable for users with intellectual disabilities and those who speak the language of the questionnaire non-natively.

This thesis presents a mixed-methods study in which a redesigned digital health service produced in the Finnish public sector was tested with participants from these groups in a laboratory using the think-aloud technique. It then aims to establish whether three self-administered usability measures capture accurate, comparable measures of these users' experiences and compare these measures to the observations of the tests. Additionally, any specific elements of the service that are especially challenging to participants are identified and used to propose good practices that can be followed when developing services for these groups.

Several specific issues are identified, including participants accepting default options without properly considering them and large, prominent interface elements distracting users. The usability scores studied all showed similar patterns, however answers by the non-native Finnish speaking participants were often inconsistent.

Based on the results, all studied usability measures seem to work appropriately for participants with intellectual disabilities, though their results may not compare with other users directly. The results with non-native Finnish speaking participants are so varied that their use cannot be recommended without further research.

Keywords accessibility , usability , intellectual disability , cognitive accessibility

Preface

I would like to thank Mari, Antti and Yasmin from Kela for the chance to work with a service that so many people value and rely upon.

Special thanks also go to the study participants and those who supported them to take part; your experiences are important and irreplaceable.

And, of course, I want to thank my supervisor, Sari Kujala, for the extensive support and guidance I received from the very start of this project.

Espoo, 18 March 2024

Matthew Hallonbacka

Contents

1 Introduction

In recent years, digital services have become the primary way many people transact with and find services from both the private and public sectors. More than ever before, essential services are available primarily as self-service pathways on the web. Usability, that is, how well these services meet the needs of their users [1], has been the focus of extensive work, in order to ensure that as many users as possible can make use of services which may be essential to daily life, such as banking and health.

A review of published literature concerning usability of eHealth services [2] found that the most common experimental methods were questionnaires, in which participants answer questions about their experience using a service, and usability testing, in which researchers observe users using a service. Several standardized approaches for recording these subjective experiences have been developed, including the System Usability Scale (SUS) [3] and the Usability Metric for User Experience (UMUX) [4]. Recent work has established that most users give comparable scores with the most common metrics, and the scores typically follow similar trends as observations such as the time taken and accuracy [5]–[7].

It is not clear, however, whether these metrics accurately capture the experiences of users with intellectual disabilities, or those who answer the questionnaire in a language other than their native language. Intellectual disabilities may affect up to one tenth of the population [8] and may cause differences in how people understand what they read and hear, interact with other people and critically evaluate information. Every year 30 000-50 000 new immigrants arrive in Finland. Previous research has highlighted the difficulty some immigrants experience in accessing public services, primarily due to language barriers [9]. Very limited literature about the application of self-administered usability measures with these groups has been identified.

The aim of this study is to establish whether three usability measures work reliably with research participants with intellectual disabilities and with non-native speakers of the language of the questionnaire. Therefore, this study will test a redesigned self-service pathway inside a public sector e-health service with four users with intellectual disabilities and four non-native Finnish speakers, and comparing the answers results of self-administered usability measures they answer.

The research questions investigated are as follows:

1. How can a self-service pathway in a patient portal be designed and developed to maximize the usability by users with intellectual disabilities and non-native speakers of the language of the service?

2. Do the results of a usability test with users with intellectual disabilities and non-native speakers of the language of the service, match these users' self-reported usability of the service, using SUS, UMUX and UMUX-Lite?

3. Are the results of SUS, UMUX and UMUX-Lite consistent with one another, when used with users with intellectual disabilities and non-native speakers of the language of the service?

If the self-administered metrics are found to be accurate, they could also be used outside of laboratory-based studies, to understand how unobserved users from these groups experience digital services.

This thesis is structured as follows. Chapter 2 presents a review of relevant previous work on these subjects. Chapter 3 describes the study design and practical arrangements used to collect the data. Chapter 4 reports the results related to each of the research questions in turn. Chapter 5 discussed the results of the study, their wider implications and indicated further work, and Chapter 6 concludes and summarizes the thesis.

2 Background

The following sections will seek to understand the current literature concerning the topics investigated. Section 2.1 looks at the literature concerning patient portals and the groups who are most likely to use them. Section 2.2 examines the concepts of self-reported usability and user experience and how they are measured, and Section 2.3 presents the current approaches to meeting the needs of users with cognitive disabilities in the design of digital services.

2.1 Patient Portals and Their Users

In health systems around the world, patient portals, or patient interfaces to electronic health records, have become commonplace. While specific features vary between health systems, care providers and off-the-shelf software products, the most common features include the ability for patients to read their records, request prescriptions be renewed or refilled, book appointments and communicate with care providers [10], [11]. Records available often include medical notes, test results, prescribed medications and care plans [12]. While no two patient portals are alike, key features, such as the ability to view medical records have become near universal as competing software vendors and health systems seek to match the services offered by one another.

Broadly speaking, patient portal use seems appreciated by patients, though some potential negative effects exist. A systematic review by Goldzweig, Orshansky, Paige, *et al.* [13] found that while patient attitudes to patient portals were generally positive, evidence of a positive outcome on health was inconclusive. More recently, a survey of users carried out by Kujala, Hörhammer, Väyrynen, *et al.* [14] identified positive effects including improved relationships between patients and care providers and increased engagement with the care system, as well as challenges with accessing information, such as the use of unfamiliar or confusing language and the potential anxiety caused by misunderstanding records which are difficult to understand. A large systematic literature review by Carini, Villani, Pezzullo, *et al.* [15] found that while evidence of some benefits has been shown, for example, patients using a patient portal are more likely to manage chronic conditions well, the benefits of patient portal access on most patients "cannot be taken for granted" [15] and patient portal use does not always lead to improvements in health outcomes or health system performance. While patients may feel that they benefit from their care provider offering a patient portal and some positive effects are clear, care must be taken not to assume that all patient effects from portals are positive, as potential negative and neutral effects exist for some groups.

Several studies have noted disparities in the rates at which different cohorts of users engage with patient portals and their healthcare providers more generally. Goldzweig, Orshansky, Paige, *et al.* [13], for example, note that age, sex and socioeconomic class all influenced how likely a patient was to use a patient portal. Younger men of a higher socioeconomic class were the most likely to use a patient portal, as were those with chronic illnesses. A large study, using data from a single institution by Tsai, Bell, Woo, *et al.* [16] showed a bimodal distribution of participant age, with most users aged

30-40 or over 60, which they attributed to the relative use of healthcare by patients in these age groups. While precise distributions vary, one repeating theme appears to be that patients who have higher and more complex healthcare needs make more use of patient portals [13], [17] when they are available.

Research suggests that patients with intellectual disabilities often have greater healthcare needs. Data from Northern Ireland [18] suggest that most patients with intellectual disabilities have multiple chronic physical or mental health conditions. A systematic review by Einfeld, Ellis, and Emerson [19] identified several studies which suggest that young people with intellectual disabilities are significant more likely to experience mental disorder than the general population. As heavier users of healthcare, patients with intellectual disabilities stand to benefit from the use of patient portals when they are offered.

Some work has explored the needs of patients with intellectual disabilities when accessing patient portals. Van Dooren, Lennox, and Stewart [20] interviewed four adults with intellectual disabilities, as well as five support workers, about their needs and experiences in accessing their own health information. A key limitation of this study is it mixes the results concerning the patients who may be able to manage their own care with those who rely on a parent or professional to manage it for them. The study did, however, identify the need for concise, easy-to-understand information directed at the patient themselves. A study based around usability testing by Gonzalez Carceller [21] examined the experiences of four users with intellectual disabilities using a symptom checker, which can sometimes be a feature of a wider patient portal. Here, based on observations, several guidelines were proposed, including avoiding long words and highlighting errors.

Around the world, non-native speakers of the local language or languages, who are often immigrants, have similar healthcare needs as the native-speaking population. Shrestha [9] analyzed survey data from Finland from the early 2010s, which showed that immigrants report less access to a healthcare professional than citizens, with many mentioning language barriers as a difficulty in accessing treatment, though this study limited analysis to three specific populations. A systematic review of studies from five European countries [22] showed that immigrants have lower self-perceived health than their native counterparts in each country, though the authors note that this does not necessarily imply that their health is, in fact lower.

Some studies have examined the use of patient portals by immigrants and members of racial minorities. Chen, Schofield, Hay, *et al.* [23] identified that Hispanic patients in the United States are significantly less likely to be offered, encouraged or able to use patient portals than white patients. Similarly, Yamin, Emani, Williams, *et al.* [24] found that African American and Hispanic patients in one health system were significantly less likely to become users of the available patient portal. The authors suggest that wider health inequalities may be a root cause of this difference.

Some studies have examined how non-native speakers of the language of the service use digital health services. One study with Hispanic users in the United States, by Moore, Bias, Prentice, *et al.* [25], noted that many users report having a family member who acts as an intermediary when accessing information on the web, though this study was published in 2009, so these findings may no longer be applicable.

The study also concluded that easy-to-read language and a streamlined navigational structure aid these users.

Overall, it is clear that patient portals provide an important service and have a positive effect on many users. Patient portals may be especially useful to patients with intellectual disabilities, who are likely to have higher healthcare needs than the general population and the access to information and services offered by such portals are likely to be beneficial to immigrants and other non-native speakers of the local languages, who often experience difficulty in accessing care. Both of these groups may experience barriers to using patient portals, either from the design of the service or due to general health inequalities.

2.2 Self-reported Usability

Usability is a broad concept with several competing definitions. A widely used definition comes from the International Organization for Standardization (ISO) [1], which refers to the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". These three qualities are further defined [1], with effectiveness referring to accuracy and goal completion, efficiency referring to the minimal time and effort cost to users and satisfaction meaning the user's positive response to their needs being met. Since the definitions involve users with distinct needs and differing expectations, usability of the same system may vary from user to user and cohort to cohort. An earlier definition of usability, proposed by Nielsen [26], shares the concepts of efficiency and satisfaction, but adds learnability, memorability and minimization of errors.

User experience (UX) is a closely related, but distinct concept. The ISO [1] describes it as the "perceptions and responses" which users encounter as a result of interacting with, or even anticipating interacting with, a "system, product or service". Recognizing the difficulty in defining the concept Law, Roto, Hassenzahl, *et al.* [27] surveyed 275 researchers and practitioners before settling on a definition of UX as "something individual that emerges from interacting with a product, system, service or an object". UX is fundamentally subjective, depending on the user, their expectations and their needs.

The precise relationship between usability and user experience has been the subject of debate. Hassan and Galal-Edeen [28] suggests that UX includes usability, and additionally includes the user's previous experiences and emotions as a result of their use of the system. Følstad and Rolfsen [29], on the other hand, notes that UX may be an extension of the concept of satisfaction from the ISO definition [1], in which case, UX is an aspect of usability.

Since user experience is highly subjective, and the variety of different possible users make usability challenging to measure, several questionnaires have been developed to quantify and compare user experience, often as a proxy for the usability of the system. Brooke [3] proposed the System Usability Scale (SUS), a questionnaire comprising 10 statements which users agree or disagree with on a five-point Likert scale about their experience using a product or service. An example statement is "I felt very

confident using the system". Some questions are phrased positively, where agreeing with the statement indicates good usability, and some are phrased negatively. SUS has been adopted widely as a measure of usability, but there has been criticism of the approach [30] as a primary measure of usability. Brooke [31] has acknowledged that SUS was intended as a comparative tool, rather than an absolute measure of usability. Issues with the questionnaire notably included some challenges when the original, English-language questionnaire was used with non-native English speakers, and with translating the questionnaire into other languages [32]. Despite these issues, SUS has been shown to have good internal validity, measure both learnability and usability [33] and to show a "significant correlation" with task completion rate [6].

In response to concern about the global applicability of SUS and the limitations of a five-point Likert scale, Finstad [4] proposed the Usability Metric for User Experience (UMUX). UMUX used a seven-point Likert scale, but requires the user to answer only four questions, two positively worded and two negatively worded. Both at the time of publication [4] and in later studies [34], UMUX has been shown to have a good level of correlation with SUS, and therefore give a good measure of subjective usability, as experienced by the user.

UMUX was later further developed with the publication of UMUX-Lite [35]. Here, the correlation of the UMUX questions with SUS was analyzed, and the two positively worded questions were selected to be presented in isolation. A regression equation was produced, which allows UMUX-Lite results to be compared directly with SUS scores.

SUS, UMUX and UMUX-Lite have all been shown to measure usability as experienced, by users, so their scores are highly correlated [4], [35]. It is, however, not yet known whether the results would be highly correlated if the population studied may have additional difficulty understanding the questions or focusing on answering them.

Borsci, Federici, Mele, *et al.* [36] tested SUS, UMUX and UMUX-Lite with blind users. They found that SUS and UMUX-Lite were reliable measures of satisfaction in this cohort of users. To date, no work has been identified which explores the suitability or reliability of SUS, UMUX or UMUX-Lite with users with cognitive disabilities.

2.3 Cognitive Accessibility of Services

Since patient portals are typically web services, research and guidelines concerning the usability and accessibility of web services more broadly are also relevant to the questions explored in this study.

Accessibility refers to the extent to which services can be used by users with disabilities [37]. Guidelines for making web content and services usable by as many users with disabilities as possible have existed in some form since the earliest days of the web [38] and have evolved alongside the platform, leading to the Web Content Accessibility Guidelines (WCAG). The most recent version, version 2.2, which was released in 2023 [39].

WCAG 2.0 [40] was released in 2008 and set out measurable standards to ensure that web sites applications are usable by as many users with disabilities and users

using assistive technology as possible. In 2018, an update to WCAG, version 2.1, was published. WCAG 2.1 had a stated goal of improving the guidance relevant to "users with cognitive or learning disabilities" [41], among other groups. New guidelines included ensuring that actions with the same effect are labelled consistently and providing additional, machine-readable information about the type of information needed in a form, which was intended to support users who rely on auto-completion due to difficulty recalling personal information accurately. The added criteria give some insight to the types of issues that might be experienced by users with intellectual disabilities.

In much of the world, WCAG has been codified into local law. In the United States, the guidelines have been adopted as the basis for Section 508 of the Rehabilitation Act, which requires the federal government to purchase information technology suitable for use by users with disabilities [42]. In the European Union, they are the basis for the standard EN 301 549 [43] which is required of digital services provided by member states' public sectors and parts of their private sectors.

Despite the prevalence of cognitive disability and notwithstanding the additions in WCAG 2.1, few WCAG criteria specifically address the needs of users with cognitive disabilities. Many of those guidelines which do are limited to the AAA-level, which is not a legal requirement anywhere [44] and is considered out-of-reach of most services [39]. For example, providing written information at a "lower secondary reading level" is a requirement only at AAA level [45]. The guidelines, which have become requirements for some services, offer relatively few guidelines to ensure that services meet the needs of users with cognitive disabilities.

A related concept is that of cognitive load, as proposed by Chandler and Sweller [46]. In this theory, the demands of the task influence the working memory required to complete it. A digital service designed, organized and presented logically will lower the working memory required to engage with it [47]. Minimizing the cognitive load of a digital service allows it to meet the needs of those with a lower working memory due to a cognitive disability.

Cognitive load theory also affects non-native speakers of the language of the material. Roussel, Joulia, Tricot, *et al.* [48] showed that adult learners of a foreign language risk overloading their working memory when trying to both acquire a new language and learn new, unrelated material in an academic environment. It is likely that the same effect can be seen when users are seeking to complete any task while practicing a language. Since the service may not be available in their native language, participants may not be able to avoid using the new language, so minimizing the cognitive load required to complete the task could help this cohort of users complete their tasks accurately.

One approach to creating services which minimize the cognitive load while meeting the needs of as many users as possible is Universal Design, which seeks to design systems which work for as wide an audience as possible "without the need for adaptation or specialized design" [49]. Some principles, such as "equitable use" and "simple and intuitive use" can be directly applied to digital services. Other principles, such as "low physical effort" and "size and space for approach and use" can apply to technology, but not directly to software, so may be less relevant to those developing

self-service pathways for use on the web. Overall, while WCAG provides specific, measurable criteria which can be explicitly passed or failed by a service, the Principles of Universal Design are more difficult to quantify, and provide only a framework for thinking about users' needs.

Another approach is Inclusive Design, which was became popular following the publication of the British Standards Institution's *Guide to managing inclusive design* [50]. This concept has some significant similarities to universal design but focuses on designing mainstream products to meet the needs "of as many people as reasonably possible". A key criticism of this approach is the use of the word "reasonably", which suggests that if a change to include a group of users would be too costly or difficult, it may not be reasonable and can be skipped, excluding that group [51]. Again, this approach offers a good way for designers to consider users' needs, without prescribing specific required features as found in WCAG.

While not usually legally mandated, designing services to be usable by users with intellectual disabilities is beneficial both for the inclusion and independence of these users, and potentially for other users experiencing additional cognitive load, like those accessing services in a language other than their native language. As well as precise guidelines, popular design principles suggest aiming to meet the needs of as many users as possible when building services.

3 Methods

This chapter will explain the methods used in the study and the reasons they were chosen. Section 3.1 presents the high-level research approach. Section 3.2 introduces the service which participants tested. Section 3.3 describes how the participants were recruited. Section 3.4 explains the practical approach to data collection. Sections 3.5, 3.6 and 3.7 describe the content of the test sessions, covering the background questionnaires, usability tests and follow-up questionnaires respectively. Finally, Section 3.8 presents the approach taken to analysis of the data.

3.1 Research Approach

To investigate the research questions, a mixed-methods approach was used. This combined qualitative and quantitative data collection and analysis. Data were collected from video-recorded usability tests and a series of self-administered questionnaires. The process by which each participant took part in the usability test and answered the various questionnaires, is shown in Figure 1.

The background questions were chosen to collect information on the participant since previous studies have shown that participant demographics including age [52] and country of origin [53] can influence participant performance and the perceived value of solutions. The health literacy questionnaire [54] was chosen to give an insight into whether the participants' levels of comfort managing their health influences their perception of the solution.

Field-based studies, such as contextual inquiry and ethnographic interviewing [55] are considered highly insightful, though the use of a health-related service in this study meant that it would be inappropriate to run the study without a version of the software including fictitious information. Technical limitations offering this meant that this study was not suitable for field-based work.

In the laboratory, "think aloud"-style usability testing, in which the participant is instructed to verbalize their thoughts while using the software, has been a preferred method for several decades, and the guidance has remained mostly unchanged in that time [26], [56]. Over this time, it has been widely employed both in academia [57] and in industry [58] to test new software with real users and to gain insights into participants' thought processes in a controlled environment.

Once the data were collected, qualitative analysis focused on identifying the problems highlighted by the usability tests and the open feedback participants offered and seeking the causes of these in the design of the service. Quantitative analysis examined the scores derived from the questionnaire answers, as well as the time taken to complete the tasks and their accuracy during the usability tests. Examining the accuracy and time taken, as well as the satisfaction of the user means that the three qualities mentioned in the ISO definition of usability [1] are measured.

This approach enables both useful information about the observed usability of the service to be collected, while also allowing the comparison of the subjective results of the usability questionnaires with the observations of the tests. It allows questions to be answered which are both useful in the development of the service and provide insight

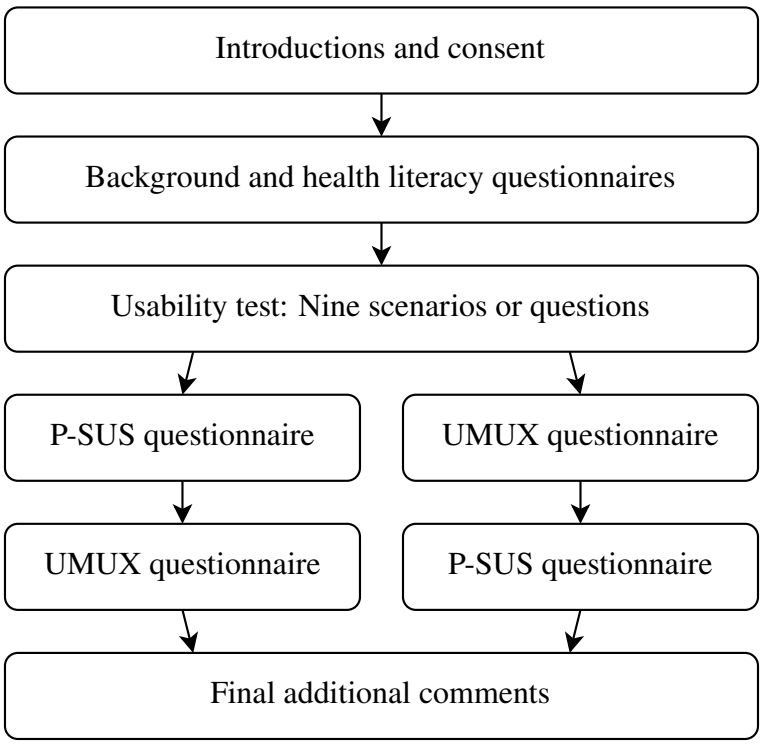

Figure 1: The order of tasks and questionnaires in each test session. Half of participants in each cohort answered each of the usability questionnaires first.

into how the self-administered usability measures studied capture the experiences of these groups.

3.2 My Kanta

In Finland, Kela, the social insurance institution, manages a national, central patient data repository, called Kanta [59] (from the former name, Kansallinen terveysarkisto, National Health Archive, and a noun meaning 'core' or 'root').

All public and most private sector healthcare providers, dental clinics, optometrists, laboratories and social care providers report their patient interactions into Kanta [59], including prescriptions, test results and medical notes.

The service includes a patient portal, called My Kanta (*OmaKanta*), from which patients can see their own records. The service is available in Finnish and Swedish, the two official languages of Finland. My Kanta is widely used, with more than 90% of adults viewing their own information in 2021, at the height of the COVID-19 pandemic [14].

Most aspects of the service only concern information stored in Kanta, so other common patient portal features, like appointment booking and messaging, are not included. These services may be available to patients in Finland, however they are offered by healthcare providers outside of My Kanta.

My Kanta allows patients to request the renewal of prescriptions. Prescription renewal has been described by users as the most useful feature of the service [14].

When requesting a prescription renewal, the patient can choose to one or more items they have been prescribed earlier [60]. They also choose a recipient, for example, the public health center which provides their primary care or their private sector occupational health provider.

The healthcare provider then reviews the request and chooses whether to renew the prescription or not. Patients can choose to receive updates about the request by SMS text message.

3.3 Participant Recruitment

Two cohorts of participants were recruited independently. The first cohort comprised adults with intellectual disabilities, and the second cohort comprised adults who speak Finnish non-natively and have learned Finnish as adults.

The participants with intellectual disabilities were recruited and supported by the Selkeästi meille project, which is run in partnership between the non-profits Kehitysvammatuki 57 and The Finnish Association on Intellectual and Developmental Disabilities (*Kehitysvammaliitto*). Participants with intellectual disabilities were required to be able to use a computer independently, though they were advised that help using understanding the service was available.

Non-native Finnish speakers were recruited by the author and the group's networks. Inclusion criteria for this group were having started learning Finnish as an adult, speaking Finnish below native-level and considering themselves able to use a digital service with the interface language set to Finnish, even if they need to translate some

words. No requirements set a precise level of Finnish required. Participants who spoke Swedish better than they spoke Finnish were excluded, since the service being tested, like all public sector services in Finland, was also available in Swedish.

At the time of recruitment, all participants were informed that they would be testing a digital service produced by the Finnish public sector, but they were not told the name of the service nor the organization which offers it, to avoid them preparing specifically to use My Kanta.

3.4 Test Procedure

Tests were conducted in a dedicated usability testing laboratory at Aalto University. The laboratory was arranged as shown in Figure 2. The tests were carried out on a laptop computer, with the participant seated at a table, in the position labelled 1, and the author, acting as moderator, in position 2. Background information and follow-up questionnaires were administered on a tablet.

A representative of Kela observed the tests from position 3 of Figure 2, behind a one-way mirror, labelled 9. Kela's representative was not involved in the tests or data collection but was present as stipulated in the agreement with Kela.

The participants with intellectual disabilities arrived with a representative from the Selkeästi meille project, and some also brought personal assistants. These people were seated in position 4 of Figure 2, ready to assist the participant, if necessary.

In the laboratory, cameras recorded the participant's face, allowing the analysis of non-verbal cues, and the screen was recorded using an HDMI recorder (at position 5 of Figure 2), to cross-reference the participant video and audio with the actions of the participant inside the software.

In tests 1-4, a single camera, at position 6 in Figure 2 captured the participant, but after test 4, two additional cameras were added at positions 7 and 8, to ensure the failure of a single camera would not result in a loss of research data.

The physical arrangements of the space were chosen to minimize possible distractions and to promote independence and privacy. For example, participants were informed that the Kela representative was present and able to see the contents of the computer display, but questionnaires were answered on the tablet, which was not directly visible to either the moderator or the Kela representative. Small, 6x4cm GoPro cameras were used to capture the participant while being well hidden compared to a traditional video camera.

The test sessions began with introductions and background information. This followed the structure presented by Rubin and Chisnell [56]. Participants were informed of the purpose and structure of the tests and that their participation was voluntary. They then signed a consent form. Participants also received a statement of how their information would be used and their right to withdraw their consent at any time.

Participants were compensated for their time with a gift card, which was handed over at this point, to ensure they did not feel obliged to act in a certain way during the subsequent tasks.

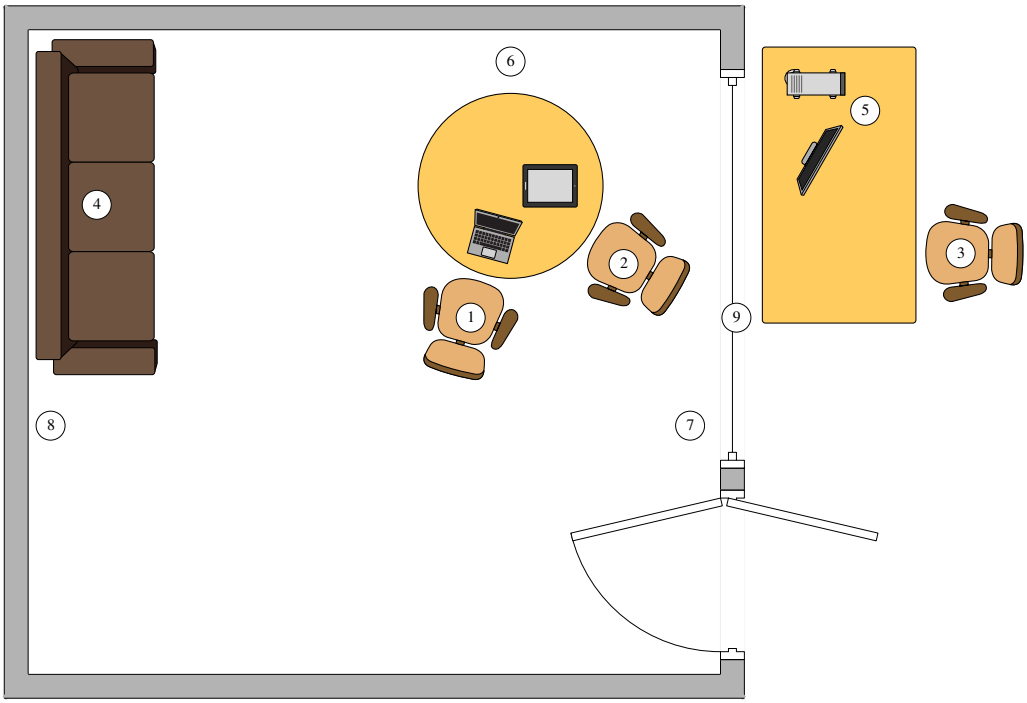

Figure 2: The physical layout of the laboratory for the usability tests.

3.5 Participant Background Information

Participants completed a background questionnaire, which included age, gender, native language, and previous experiences with My Kanta. Participants were asked if they've used any aspect of the service, and whether they have requested the renewal of a prescription before.

After the background questions, participants then answered the 16-question version of the European health literacy questionnaire (HLS-EU-Q16), originally created by Sørensen, Van den Broucke, Pelikan, *et al.* [54], which has a Finnish translation which has been validated for use with older adults [61]. Since health literacy has been shown to be a key determinant of health [62] it is important to understand the health literacy of the participants using a patient portal. The full questionnaire is included in Appendix A.

The questionnaire asks participants to indicate how challenging they find certain tasks, such as finding information about self-care and following the instructions of a doctor or pharmacist. Participants select on a four-point scale from 'very difficult' to 'very easy' and have the option to answer 'I don't know' to any question.

All participants answered the Finnish version of the questionnaire. The questionnaire was originally prepared by Eronen, Paakkari, Portegijs, *et al.* [61], though due to previous participant difficulties with questions concerning the media, the text "(for example, television, newspapers)" ("*(esim. TV, sanomalehdet)*") was added after each occurrence of the word 'media'. If a participant found a question difficult to understand, it would be read aloud, then stated in a different way, then, if the participant still could not understand, they were encouraged to choose 'I don't know'. Participants were free to look up translations of specific words or phrases from their own smartphones, if they preferred.

3.6 Usability Tests

The tasks given to participants belonged to four general types. The list of task types was written based on a structure used by Rubin and Chisnell [56], and is shown in Table 1. It includes a short description, required knowledge or material, success criteria and an expected time to complete the task.

Each task was repeated two or three times, which gave participants the opportunity to familiarize themselves with the service, repeat a journey through the service as they became more familiar with it and self-correct errors they make with practice.

The task types and specific scenarios are presented in Appendix B.

The tasks represented all functionality of the updated part of the service and were designed that some could naturally flow from one to the next. The service's developers advised that the tasks represented the implemented functionality at the time the tasks were planned in October 2023.

Since real users were unlikely to complete so many tasks in a sequence, descriptions would sometimes start with 'Imagine that it is a new day'. To add to the realism and relatability, participants would also be told to imagine having a certain common health condition, though they did not require any existing knowledge of the diagnosis or

Description	Required	Success Criteria	Expected Time
Check the remaining quantity for a specific prescription	Logged into My Kanta, name of the medication	States whether items are available to collect.	1-3 minutes
Find a specific fact about a prescription	Logged into My Kanta, name of the medication, question to answer	Answers the given question correctly.	1-3 minutes
Renew a single prescription	Logged into My Kanta, name of the medication, name, sector and municipality of the patient's health center	Submits a request to the correct health center containing (at least) the correct item.	3-10 minutes
Renew multiple prescriptions	Logged into My Kanta, name of the medications, name, sector and municipality of the patient's health center	Submits a request to the correct health center containing (at least) the correct items.	3-10 minutes

Table 1: The list of task types, written based on a discussion with My Kanta's developers. Here the 'Expected time' was based on the experiences of the author's first time trying the service, and this was used to plan the task list and not as a benchmark for participants.

treatment to understand the task.

During the tasks, participants were encouraged to verbalize their thoughts and 'think aloud' [56].

To avoid the risk of distracting participants, no notes were taken during the tasks. Instead, notes and quantitative measures were taken after the task from the video recording, when it was reviewed several days later.

Since tasks of the same type were repeated, participants were not told whether the task was completed accurately or not. This meant that participants may have thought that their task was completed correctly when it was not.

If a participant became stuck, asked for help or left the service, the participant would be given a small amount of help, such as a hint or a rephrasing of the task, or guided back to the last point where they were on the correct path.

3.7 Usability Questionnaires

Immediately after the final task, participants were asked to complete the Positive SUS (P-SUS) [63] and UMUX [4] usability questionnaires. Both questionnaires use Likert scales. P-SUS includes 10 items answered on a five-point Likert scale, and UMUX includes four statements answered with a seven-point Likert scale. The questions of P-SUS and UMUX, both in the Finnish as presented, and English translations are included in Appendices C and D, respectively.

P-SUS was used in place of SUS as it has been shown to give similar results [63] and has all questions in the positive form, where agreeing with the statement indicates higher usability. Since UMUX includes alternating statements in the positive and negative forms, including P-SUS with only positive questions ensures that participants who agree or disagree with every statement can be distinguished from those who give the 'best' answer to every question.

Half of the participants in each cohort answered P-SUS first, and half answered UMUX first, to avoid any potential bias. Within each questionnaire, the order of the questions was constant, matching the original published order of the questions, which was useful when comparing the results and calculating the scores. The questionnaires were completed on a tablet computer, which the participant was able to hold and use themselves, without their answers being visible to the moderator or the Kela representative.

Participants were allowed to use a translation tool from their own smartphone, if they wished. The P-SUS instructions, as presented on the form, request that participants choose their first reaction to the statement, and not to spend too long thinking, as per the original SUS instructions by Brooke [3]. All questions required an answer, though both Likert scales included a midpoint option, equally distanced from "Agree fully" and "Disagree fully".

3.8 Analysis

The data collected on the online questionnaires were exported to a spreadsheet for analysis.

To understand the experiences of participants using the service, the video and screen recordings were synchronized and reviewed at least twice each. As the videos were reviewed, notes were taken concerning start and end time of each task in the video, whether the task was completed accurately, and any comments made by participants. Any observed usability issues were coded and collated in a spreadsheet, with each issue given:

- A unique identifier
- A description
- The type of task in which the issue occurred
- Whether the issue resulted in failure to complete the task correctly, caused a delay or required intervention.

When reviewing the video recordings, accuracy was recorded as a binary value per participant per task. Accuracy was defined as completing the task correctly, based on the requirements listed in Table 1. Whether the participant required help was also recorded as a binary value per participant per task.

All of the aforementioned quantitative data were then imported into R [64] and combined for further analysis.

The HLS scores were calculated using the method described by Sørensen, Van den Broucke, Pelikan, *et al.* [54], where the mean of the responses to the questions answered is calculated, then the index calculated using the following formula:

$$\text{index} = (\text{mean} - 1) \times \frac{50}{3}$$

According to the author, a score is not calculated if the participant answers "I don't know" to more than three prompts.

The P-SUS scores were calculated using the approach described by Sauro and Lewis [63], this is based on the original SUS questionnaire by Brooke [3]. Instead of handling the odd-numbered and even numbered questions separately, as in the original SUS, all scores each had one subtracted from them. The sum of all scores is then multiplied by 2.5 to give a score out of 100.

The UMUX scores were calculated using the method described by Finstad [4]. First, odd-numbered items have one subtracted, while even numbered items are subtracted from 7. The sum of the four items is then multiplied by $\frac{100}{24}$ to scale the resulting score out of 100.

UMUX-Lite scores were calculated using only questions 1 and 3 of UMUX [35]. Here, the previously mentioned UMUX methodology was used to calculate a score out of 100 (labelled $\text{UMUX}_{1,3}$ below), then the following regression equation is used to create a result which can be compared directly with SUS:

$$\text{UMUX-Lite} = 0.65(\text{UMUX}_{1,3}) + 22.9$$

To identify the potential relationships between the different questionnaire results, the Spearman's rank correlation coefficient, ρ, was calculated between each pair of questionnaires used. ρ is a suitable measure in this case, as the relationship between SUS and UMUX is not expected to be linear.

Pearson's correlation coefficient, r is used to compare the numeric answers to each pair of questions, in order to identify questions which appear to receive correlated answers within each cohort. For each r value, R [64] additionally calculates the p-value using an asymptotic t-approximation, allowing statistically significant values to be identified.

4 Results

This chapter presents the results of the usability tests and related questionnaires, including both information provided by the participants and observations recorded. Section 4.1 explains the demographic structure of the participant cohorts and their self-reported health literacy. Section 4.2 presents quantitative measurements of the usability tests and the issues experienced by participants. Section 4.3 analyzes the results of the usability questionnaires. Finally, section 4.4 compares the results of the self-reported usability metrics with one another.

4.1 Participants

4.1.1 Demographics

Each cohort contained four participants. Both cohorts included men and women, and the age ranges were similar in both cohorts. Participant ages and genders are presented in Table 2.

Table 2: Self-reported participant age and gender information. ID = Intellectual disability, NN = Non-native Finnish speaker.

Participant	Cohort	Age Range	Gender
1	ID	30-35	Male
2	ID	30-35	Female
3	ID	35-40	Male
4	ID	35-40	Female
5	NN	45-50	Male
6	NN	30-35	Male
7	NN	40-45	Female
8	NN	30-35	Female

Participants represented a range of cultural and linguistic backgrounds. All participants with intellectual disabilities were born in Finland and spoke Finnish as their native language. The non-native Finnish speaking participants were born in Nepal (n=2), Bangladesh and Germany. These participants were not asked to quantify their level of Finnish, but all described their own skills as less than native level. Notes taken during the tests suggest that the levels varied from those who were able to speak and read fluently to those who heavily relied on automated translation tools.

In order to identify the possible effect of earlier experience using My Kanta, participants were asked whether they had used it before and whether they has requested the renewal of a prescription earlier. Table 3 shows the answers given, by cohort. Most participants had used My Kanta before. Most participants with intellectual disabilities,

and half of the non-native Finnish speakers, had requested a prescription renewal using My Kanta before.

Table 3: The number of participants with experience with My Kanta, by cohort. ID = Intellectual disability, NN = Non-native Finnish speaker.

	ID	NN
Had used My Kanta	3 (75%)	4 (100%)
Had requested a prescription renewal	3 (75%)	2 (50%)

Other notable observations about the participants were made during the tests. In the cohort of participants with intellectual disabilities, one participant also had a mobility disability. This participant used a wheelchair and reported that they were unable to use a mouse, though they could use a trackpad to use a computer independently and used a tablet to answer the questionnaires without adaptations or adjustments. The mobility disability appeared to have no effect on using the service.

As part of the experiment, participants were paid for their time with a gift card. For legal reasons, participants were required to provide some personal information in connection with this payment, though these were not connected with their responses. These details were collected in an online survey tool without validation of the formats. All four participants with intellectual disabilities provided their identity number, which is a unique personal identifier used in Finland, in an invalid format, while non-native Finnish speakers made no errors. This was corrected after the experiment.

The participants with intellectual disabilities were all experienced testers, having taken part in many previous usability tests through the Selkeästi Meille project. They appeared comfortable and confident with the process and arrangements. Some made comments about the concepts of usability and accessibility during the tests and showed a knowledge which may exceed that of the average user.

The participants who spoke Finnish non-natively had generally high levels of education. All had completed at least secondary education, and two had completed a university-level degree. All had arrived in Finland as adults and learned the language after arriving. One participant said they used the language daily, while others said they use it most days. Some said they had additionally studied Swedish, but none rated their own skills in Swedish as high and all preferred to use Finnish.

4.1.2 Health Literacy

All participants answered at least 13 of the 16 health literacy questions, so all participants received a numerical score. The average scores and standard deviations for each cohort are presented in a table and plotted as a box plot in Figure 3.

Overall, participants with intellectual disabilities rate their own health literacy higher than non-native Finnish speaking participants. The variance in scores is also lower in the cohort with intellectual disabilities, as seen from the standard deviation.

The number of times 'I don't know' was chosen by participants was also compared. On average, participants with intellectual disabilities answered 'I don't know' to 1.25 questions (s.d. 1.5) and the participants who spoke Finnish non-natively also answered 'I don't know' to 1.25 questions (s.d. 1.26).

Three questions were skipped by at least one participant from each cohort. These were:

- How easy/difficult is it for you to judge whether the information on health risks in the media (e.g., TV or newspapers) is reliable?
- How easy/difficult is it for you to decide how you can protect yourself from illness based on information in the media (e.g., newspapers, TV)?
- How easy/difficult is it for you to understand information in the media on how to get healthier (e.g., from TV, newspapers)?

Other questions were skipped by at most one participant.

4.2 Usability Test Results

The following sections present the results of each measurement derived from the recordings of the usability tests. Accuracy and completion time are the quantitative part of the study, while the observed issues are qualitative.

4.2.1 Task Accuracy

No tasks were unsuccessful, meaning the participant would be unable to complete them, even after help was offered. Some tasks, however, were completed inaccurately, according to the planned success criteria listed in Table 1.

The accuracy rates for each task are presented in Table 4, while the results per type of task are visualized in Figure 4. The accuracy can be seen to mostly improve with practice, as participants become familiar with the service and the tasks. All participants who completed a task inaccurately were able to self-correct the cause of their particular inaccuracy on future attempts.

Overall, non-native Finnish speakers were more accurate than participants with intellectual disabilities, and completed all tasks, except task nine, at least as accurately. Many tasks, such as all instances of task type two, and tasks seven and eight, were completed successfully by every participant.

Most inaccuracies in both groups happened when they were asked to renew a prescription (task types 3 and 4). Here, incorrect items and recipients were common among participants in both cohorts.

No failures were caused by participants being unable to continue using the service. Some participants needed guiding back to the correct page, but all participants were

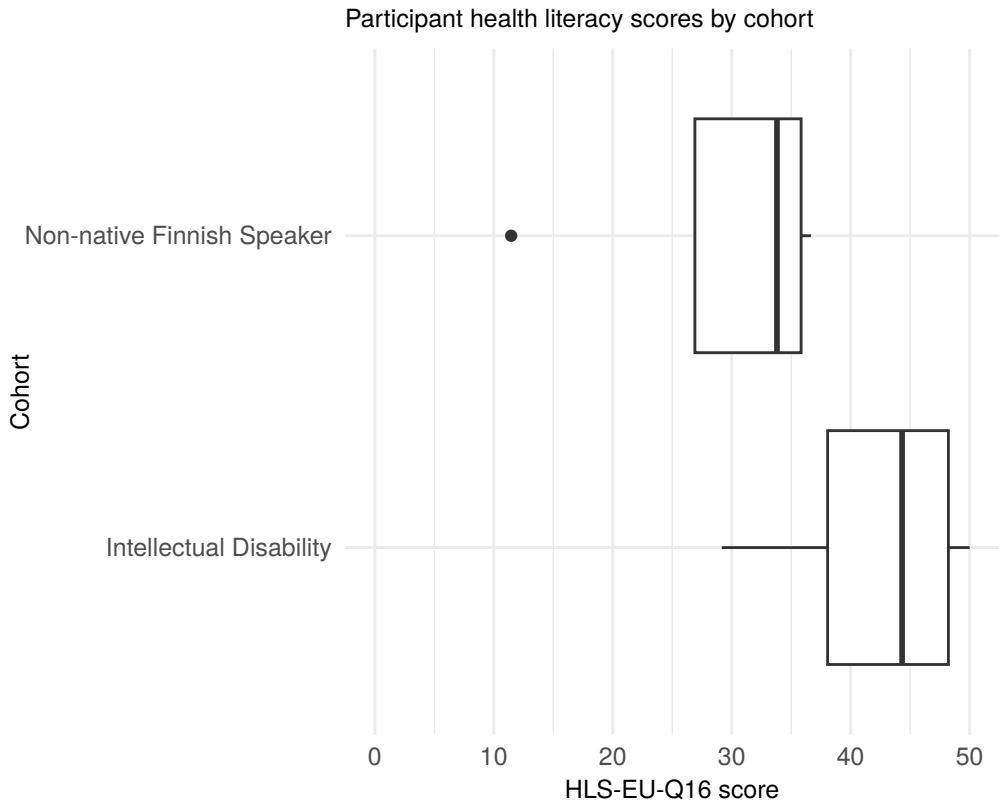

Figure 3: Box plots of participant self-reported health literacy scores using HLS-EU-Q16.

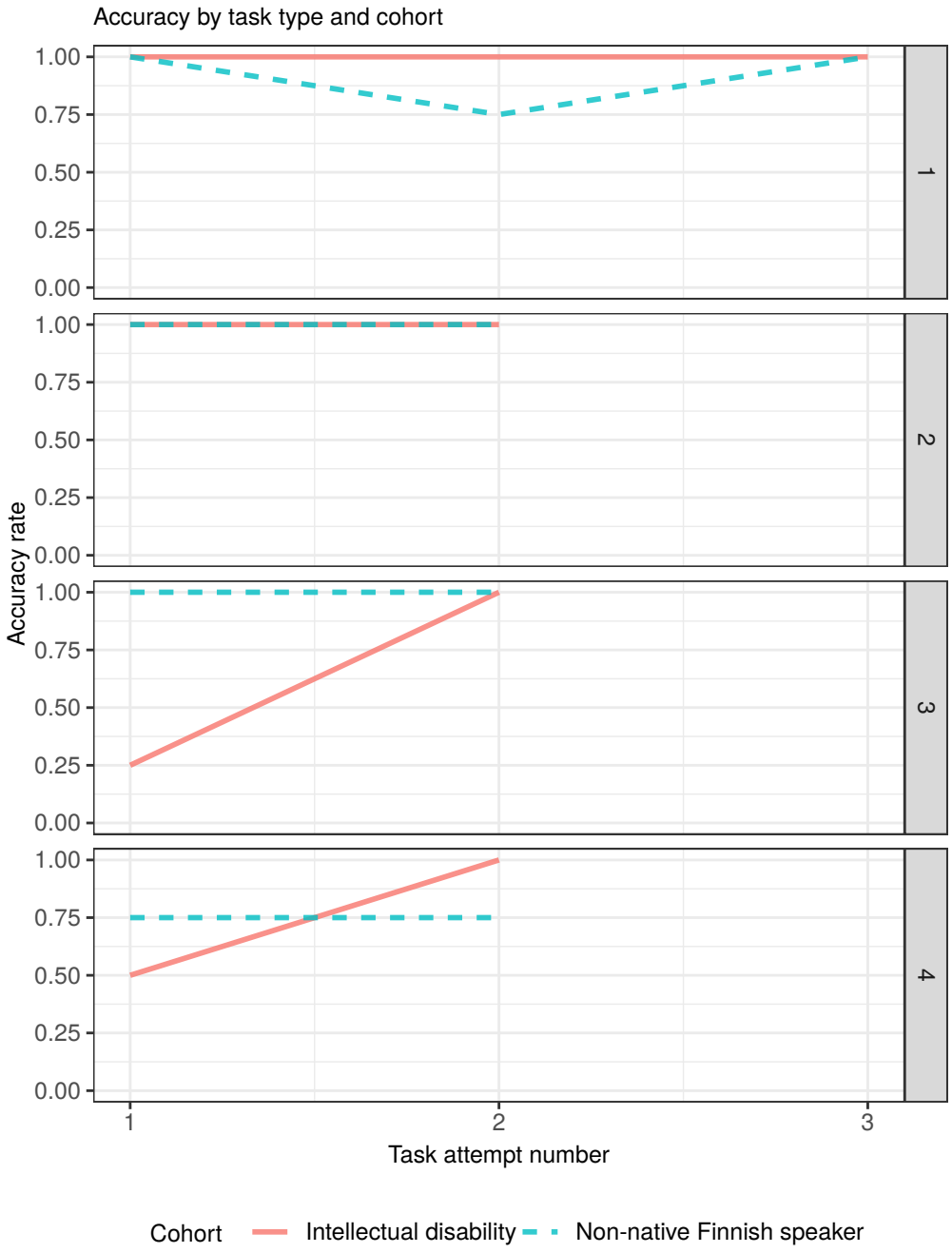

Figure 4: Line graphs per task type, showing the development of the accuracy over subsequent attempts. With more practice and familiarity, participants usually have more success in the tasks.

Table 4: Task accuracy rate, by cohort.

	Task Type	ID (%)	NN (%)
Task 1	1	100	100
Task 2	2	100	100
Task 3	1	75	100
Task 4	3	50	75
Task 5	2	100	100
Task 6	4	25	100
Task 7	3	100	100
Task 8	1	100	100
Task 9	4	100	75

able to reach the end of every task, even if the data the participant provided or the answer they gave did not exactly match the task description.

4.2.2 Task Completion Time

As stated in Section 4.2.1, participants sometimes failed to meet the success criteria for a task. In all cases, they still completed all necessary steps, interacted with the same parts of the service and reached the intended end point of the task. Therefore, the time taken to complete the task was still measured, as the time taken to complete the task accurately is likely to be very close to the time taken.

The average completion times for each task are presented in Figure 5, grouped by task type. Here, the decrease in time taken as participants become more familiar with the service is clear in most cases. Participants were usually able to complete a task more quickly the second time they completed it, and even more quickly the third time.

In most cases, participants with intellectual disabilities were slightly quicker completing the tasks than non-native Finnish speakers. The error bars on the graphs in Figure 5 show the standard deviation of the times taken to complete the tasks. In most cases, the standard deviation is smaller for participants with intellectual disabilities, suggesting that they were more consistent in their performance.

4.2.3 Observed Usability Issues

A total of 24 issues were recorded, of which seven resulted in the task being completed inaccurately. A full list of the issues identified is included in Appendix E. All issues were categorized as either causing the task to be completed inaccurately ('Inaccuracy', in the appendix), causing the participant to need prompting or guiding back to the correct step from which they can continue accurately ('Intervention') or causing a delay without requiring help or causing any inaccuracy ('Delay')

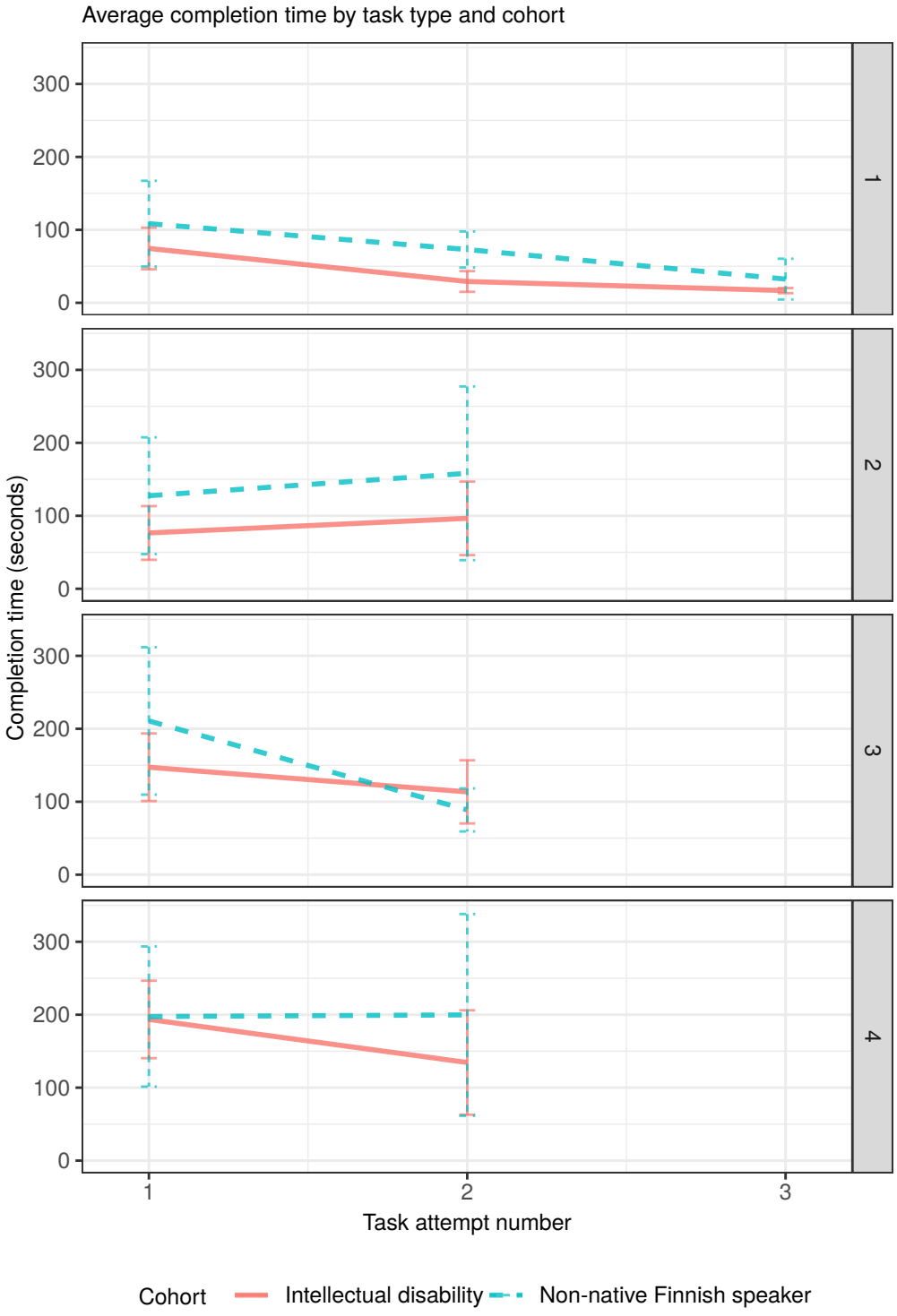

Figure 5: Line graphs per task type, showing the development of the time taken over subsequent attempts. Participants usually get quicker completing the tasks.

Issues which resulted in task being completed inaccurately included the participant answering that a prescription is valid when it is not or sending a renewal request to the wrong recipient. Issues requiring intervention included the participant following a link to another service and not recognizing that some visible text was a link. Issues which caused tasks to take longer than required, but did not result in the task being completed inaccurately, included skipping questions, which the service would not allow, or starting a request over, rather than editing the request they have in progress.

The following issues were the most frequently recorded:

1. When asked to add an item to the prescription request, participants would cancel the request and start over, rather than amending the original request (n = 4, ID = 2, NN = 2, caused a delay)

2. When sending a prescription renewal request, participants would accept the default recipient, rather than choosing the recipient they were asked to (n = 3, ID = 3, NN = 0, caused inaccurate completion of the task)

3. When asked to identify why they were prescribed a medication, participants would miss the information labelled 'reason for use' (*käyttötarkoitus*) at the top of the prescription information page, and would follow the more prominent link to the medicinal product search, labelled 'are you looking for more information about medications?' (*etsitkö lisätietoa lääkkeistä?*) instead (see Figure 6, n = 3, ID = 1, NN = 2, required intervention)

A summary of how many issues were recorded per task type is presented in Table 5. Here, participants with intellectual disabilities can be seen to have more issues, on average, than non-native Finnish speakers.

Table 5: Average and standard deviation of the number of issues recorded per participant per task, grouped by cohort and task type.

Task type	ID	NN
1	0.00 (0.00)	0.17 (0.39)
2	0.38 (0.74)	0.25 (0.46)
3	0.88 (0.83)	0.63 (0.74)
4	1.38 (1.69)	0.88 (0.64)

Many of the identified issues have a clear root in the design of the service. For example, in issue 3, above, the visual prominence given to the link to general information about medications appeared to distract users from information about their own prescription.

For some issues, such as participants not filling in form fields which were marked as required, the cause is more likely to be human error. Commonly used good practices

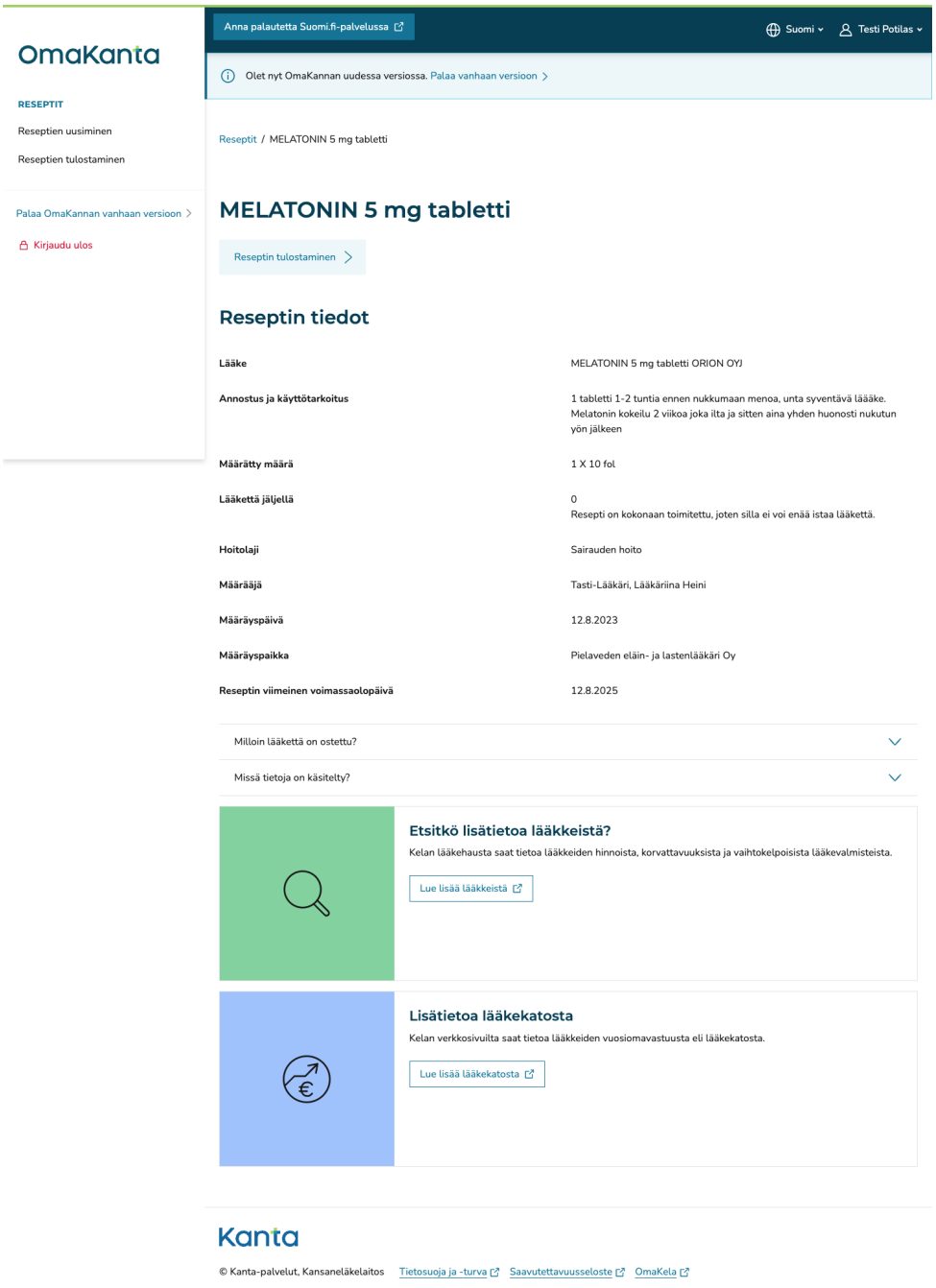

Figure 6: A screenshot of the prescription information page for an example prescription. Two notable issues visible here are the large, bold block labelled 'are you looking for more information about medications?' (*etsitkö lisätietoa lääkkeistä?*), which distracted several participants who were searching for the 'reason for use' (*käyttötarkoitus*), which is near the top of the page. Also visible is the navigation area on the left, where some participants didn't recognize the link labelled 'prescriptions' (*reseptit*) as link due to the different styles from the following links.

for marking required fields have been followed in the design of the service, but this will not cause every user to fill them correctly the first time. In these cases, the software needs a tolerance for error. Solutions, such as validating the values are already present and avoid a missed field causing the task to be failed.

4.2.4 Summary

Overall, all participants were able to reach the end of all tasks, though some participants in both cohorts had made errors in the data they submitted. Usually, participants would be able to correct errors in subsequent attempts of the same type of task.

Participants with intellectual disabilities completed most tasks correctly, especially on the second or third occasion that a type of task was repeated. Non-native Finnish speaking participants made fewer errors which resulted in the failure of a task, but took noticeably longer to complete most tasks.

Many of the issues observed could be addressed in the design of the service, though some may be unavoidable user error. Frequently occurring issues included participants not noticing the option to go back during the process, starting over in order to make a change and participants being distracted by a prominent link to another service.

Overall, participants seemed to become more familiar with the service as they completed tasks. They appeared confident with the type of tasks they were completing and rarely needed help to progress in tasks. However, there was a noticeable difference in the number of issues encountered between the two cohorts, with participants with intellectual disabilities experiencing more issues, yet often not noticing them.

4.3 Self-reported Usability Measures

In the following sections, the results of the P-SUS, UMUX and UMUX-Lite questionnaires will be explored. The overall average scores of each cohort with each questionnaire are presented in Table 6.

Table 6: Participant average P-SUS, UMUX and UMUX-Lite scores and equivalent curved grade using the scale by Sauro and Lewis [5] (only available for P-SUS and UMUX-Lite), by cohort.

	ID	**NN**
Mean P-SUS score (s.d.)	91.3 (5.2) Grade A+	63.1 (23.6) Grade C-
Mean UMUX score (s.d.)	89.6 (20.8)	49.0 (12.4)
Mean UMUX-Lite score (s.d.)	77.2 (5.4) Grade B+	56.9 (11.2) Grade D

4.3.1 P-SUS

All participants with intellectual disabilities rated the service highly, with scores of 85 and above. Participants with intellectual disabilities had a low variance in their scores, with a standard deviation of 5.2. This indicates that their subjective experiences were similar within the group.

On average, non-native Finnish speakers, rated their experience less highly. Their experiences were also far more varied, with scores ranging from 40 to 95.

As reported in Section 4.2, participants with intellectual disabilities were quicker using the service than the non-native Finnish speaking participants, which supports the idea that they found the service more usable. The non-native Finnish speaking participants, however, experienced fewer issues and completed a higher proportion of tasks accurately.

Measuring the internal consistency of all responses, using Cronbach's alpha, results in a value $\rho_\tau = 0.86$, which indicates a high level of internal reliability [65]. Analysis of each cohort in turn yields different results. For non-native Finnish speakers, $\rho_\tau = 0.85$, which also indicates a high level of internal reliability.

For participants with intellectual disabilities, however, $\rho_\tau = -1.07$, which indicates that the assumptions of the test are not met. Further analysis identifies that this is a result of a small sample size, and all participants answering "Strongly Agree" to half of the questions, which were then ignored due to the zero variance. The remaining items had a negative average covariance, leading to the negative value of ρ_τ.

Overall, the P-SUS results suggest that participants with intellectual disabilities found the service more usable than non-native Finnish speakers, despite them being observed to experience more usability issues during the tests. The P-SUS score for participants with intellectual disabilities was very high, while that for non-native Finnish speakers was close to average.

4.3.2 UMUX

The results per cohort of the UMUX questionnaire is presented in Table 6. While the scores are out of 100, they are not directly comparable to the SUS scores and the curved letter grades do not apply to them [5].

Here, a similar trend can be seen in the UMUX results, as appeared in the P-SUS results. Participants with intellectual disabilities rated their experience of using the service as high and non-native Finnish speakers give a low score.

A key difference between the cohorts can be seen in the standard deviations. With the UMUX questionnaire, the variance among participants with intellectual disabilities is much higher than with the P-SUS questionnaire. This suggests that these participants were reporting more varied experiences with the service than they reported using the P-SUS questionnaire.

Internal reliability analysis, using Cronbach's alpha, results in a value $\rho_\tau = 0.75$, which indicates a slightly lower level of internal than the P-SUS questionnaire, but still an acceptable level. Analysis of each cohort in turn gives a value of $\rho_\tau = 0.83$ for participants with intellectual disabilities.

Among the non-native Finnish speakers, however, $\rho_\tau = 0 - .57$, which is caused by a negative covariance among the items. Here, no items were excluded, however the average covariance among items was still negative, leading to the negative value of ρ_τ.

Again, the results here do not full match the observations of the tests, where participants with intellectual disabilities were observed to encounter more issues than non-native Finnish speakers.

4.3.3 UMUX-Lite

The results of the UMUX-Lite calculation are presented in Table 6. Since the UMUX-Lite score is calculated using a regression equation, the results are designed to be comparable to SUS scores, and the results are presented alongside the curved letter grades scale by Sauro and Lewis [5].

The use of the regression equation means that the UMUX-Lite scores are limited to the range 22.9-87.9. This means that the maximum score is lower than the average P-SUS score observed for the participants with intellectual disabilities in this study.

Internal validity analysis, using Cronbach's alpha was not undertaken, as a two-item scale does not meet the requirements to use the test. Despite this, the UMUX-Lite scores broadly follow the trend seen in the SUS scores. Participants with intellectual disabilities rated their experience very highly, while non-native Finnish speakers rated theirs as slightly below average. Once again, the standard deviation of the scores for participants with intellectual disabilities is lower, indicating a more consistent experience.

4.3.4 Summary

Three different usability metrics were used to record participants' subjective experiences using the service. All three metrics showed a difference in the experiences of participants with intellectual disabilities and non-native Finnish speaking participants. In every case, participants with intellectual disabilities rated their experience more highly than non-native Finnish speakers.

All scores were very high for participants with intellectual disabilities, showing that they found the service extremely usable. This was despite every being observed to make more errors and require more help than the non-native Finnish speaking participants.

The variance of scores was not consistent. Using P-SUS and UMUX-Lite, participants with intellectual disabilities reported a more consistent experience, while using UMUX, they reported a more varied experience then the cohort of non-native Finnish speakers.

4.4 Comparison of Usability Measures

Previous sections have considered the observable experiences of participants and their subjective experiences, as reported using questionnaires. This section will examine

the relationships between the different questionnaire results and whether their results are consistent with one another for these cohorts of users.

4.4.1 Correlations

Table 7 shows the results of the correlation analysis. Since UMUX-Lite is calculated from a subset of the UMUX responses, the correlation between UMUX and UMUX-Lite is expected to be high and is not explored further. The cohort-specific correlations show a stark difference between the relationship between P-SUS and UMUX in the two cohorts.

Table 7: Matrix of Spearman's rank correlation coefficients, ρ, between the different usability measures. * = $p < 0.05$, ** = $p < 0.01$.

	P-SUS	UMUX	UMUX-Lite
P-SUS	1.00		
UMUX	All: 0.43 ID: 0.78 NN: -0.63	1.00	
UMUX-Lite	All: 0.77* ID: 0.78 NN: 0.32	All: 0.88** ID: 1.00** NN: 0.50	1.00

Participants with intellectual disabilities show a fairly strong positive correlation between P-SUS and UMUX, as well as between P-SUS and UMUX-Lite. This suggests that these participants were consistent in their responses to both questionnaires and understood what they were being asked. Non-native Finnish speakers, however, show a moderate negative correlation between P-SUS and UMUX, and a weak positive correlation between P-SUS and UMUX-Lite. No correlations, other than the inevitable high correlation between UMUX and UMUX-Lite, were statistically significant within either cohort.

Overall, the correlations show that the results of the different questionnaires are generally consistent with one another when answered by participants with intellectual disabilities, however non-native Finnish speaking participants seem to answer inconsistently.

4.4.2 Question Consistency

The results of Pearson's correlation coefficient (r) analysis of the pairs of questions are shown in Table 8. The negatively worded questions have had their scores reversed, so that a positive correlation always indicates similarity of response.

Table 8: Matrix of Pearson's correlation coefficients, r, between the different usability measures. * = $p < 0.05$, ** = $p < 0.01$.

	UMUX Prompt			
P-SUS Prompt	Q1	Q2	Q3	Q4
Q1	ID: Undefined NN: 0.73	ID: Undefined NN: -0.05	ID: Undefined NN: 0.57	ID: Undefined NN: -0.38
Q2	ID: Undefined NN: -0.54	ID: Undefined NN: -0.91	ID: Undefined NN: 0.69	ID: Undefined NN: -0.15
Q3	ID: Undefined NN: 0.48	ID: Undefined NN: -0.62	ID: Undefined NN: 0.58	ID: Undefined NN: 0.00
Q4	ID: Undefined NN: -0.56	ID: -0.58 NN: -0.96	ID: -0.58 NN: 0.49	ID: -0.58 NN: 0.09
Q5	ID: Undefined NN: 0.23	ID: 0.58 NN: -0.83	ID: 0.58 NN: 0.59	ID: 0.58 NN: 0.14
Q6	ID: Undefined NN: -0.08	ID: -0.44 NN: -0.05	ID: -0.44 NN: 0.91	ID: -0.44 NN: -0.94
Q7	ID: Undefined NN: 0.17	ID: 1.00** NN: -0.99	ID: 1.00** NN: 0.41	ID: 1.00** NN: 0.30
Q8	ID: Undefined NN: 0.48	ID: 1.00** NN: -0.78	ID: 1.00** NN: 0.87	ID: 1.00** NN: -0.43
Q9	ID: Undefined NN: 0.83	ID: Undefined NN: -0.54	ID: Undefined NN: 0.67	ID: Undefined NN: -0.49
Q10	ID: Undefined NN: 0.83	ID: Undefined NN: -0.54	ID: Undefined NN: -0.18	ID: Undefined NN: 0.41

Since the sample size is small, some questions received the same answer from all participants in one cohort. In this case, the variance is zero, and computing the correlation coefficient would require dividing by zero. These are labelled as 'undefined', in the table. Few results were statistically significant, in part due to the small sample size.

One striking correlation is between P-SUS prompt 7 and UMUX prompt 2. P-SUS prompt 7 states "I imagine that most people would learn to use My Kanta very quickly", while UMUX prompt 2 states "Using My Kanta is a frustrating experience". Here, all participants agreed or strongly agreed with prompt 7 of P-SUS. Three of the four participants with intellectual disabilities strongly disagreed with prompt 2 of UMUX, while half of the non-native Finnish speaking participants agreed with prompt 2 of UMUX. This means there's a very strong positive correlation between the two prompts for participants with intellectual disabilities, yet a very strong negative correlation for non-native Finnish speakers.

While not visible in the table, prompt 3 of P-SUS and prompt 3 of UMUX have a very similar wording to one another. P-SUS prompt 3 states "My Kanta is easy to use", while UMUX prompt 3 states "I think that My Kanta is easy to use". Here, all participants with intellectual disabilities strongly agreed (5/5) with prompt 3 of P-SUS and agreed with prompt 3 of UMUX (three participants strongly agreed, one agreed). These are very similar answers, however it is not possible to calculate a correlation coefficient for this cohort for the reason discussed above.

The answers of the non-native Finnish speaking participants varied more. Three participants had similar answers for both prompts, either agreeing strongly with both or neither agreeing nor disagreeing with both. One participant, however, strongly disagreed with prompt 3 of P-SUS, but agreed with prompt 3 of UMUX.

Overall, there analysis of the questions shows that participants with intellectual disabilities were generally consistent in their responses to questions, both between participants and between pairs of questions. Non-native Finnish speakers, however, were less consistent in their responses, and had a few clearly outlying responses.

4.4.3 Summary

The results of the P-SUS questionnaire are correlated with the results of the UMUX and UMUX-Lite questionnaires, when they are answered by participants with intellectual disabilities. There is no clear relationship between the results when they are answered by non-native Finnish speakers.

Due to the small sample size, however, it's not possible to draw any conclusions from the specific answers recorded, as a single participant not understanding the question fully could skew an entire cohort's results.

5 Discussion

This section will examine the results of the study in the context of the research questions and what was already known about the topic.

5.1 The Research Questions

The following sections will consider the original research questions, as outlined in Section 1, what the results say about them and the causes of any unexpected results.

5.1.1 Designing Inclusive Services

The first research question asked how services can be designed to meet the needs of users with intellectual disabilities and non-native speakers of the language of the service. The observations of the usability tests showed that a variety of different issues were encountered, some by members of both cohorts and some only by members of a single cohort.

Members of both cohorts were also equally likely to cancel their request before submission when asked to add a second item to the request. This created a significant amount of repeat work, when it was possible to edit the items using a small 'back' button and continue. Services can design to ensure that the option to go back and edit details is presented with adequate prominence and visibility, and as consistently as possible to ensure users recognize the possibility.

Participants with intellectual disabilities were frequently accepting default values without noticing that they were incorrect. In one task, three of the four participants in this cohort did this when renewing a prescription, while no participants in the non-native cohort attempted to submit their request containing incorrect information. Here, services could avoid pre-filling default information entirely to increase the chance that the correct data are provided, or ask users to confirm their selection explicitly in a second step.

Non-native Finnish speaking participants were more frequently distracted by the prominent 'are you looking for more information about medications?' banner when viewing information about their prescriptions. Participants were often surprised or confused when this took them away from My Kanta and into another service. All participants who left My Kanta needed help to return to the service from this point. This may have been avoided if the banner had been smaller, less prominently displayed or not present in the interface at all. This could be generalized as a recommendation that services avoid displaying unrelated information inside the main content area of the page.

The recommendations, generalized for application to any service, can be summarized as follows:

- Don't show any unnecessary information in a list of items, only those which are essential to identifying the item in question. Provide further information on a single-item page when the user clicks on an item.

- Within an area of a page or view, style all items which serve the same purpose in the same way. For example, all links in the same part of the page should have similar styling.
- When a record in the main content area, don't display links to other services which are not about that record. The main content area should concern only the record being viewed.
- Do not provide a pre-selected default option, or use, for example a confirmation step, to require the user to accept the default option, if choosing the wrong value would have a serious negative consequence.
- Validate all input as precisely as possible and provide clear explanations of detected problems as validation messages.
- In a multi-step process, provide "back" and "next" options with similar sizes and visually close together.

Some of the issues identified, for example, the confusion caused by the inconsistent appearance of some interface elements is similar to the issue which criterion 3.2.4 of WCAG [66] seeks to avoid and the recommendation for validation closely resembles criterion 3.3.3 [67]. Others, such as the inaccuracy caused by default values being incorrectly accepted, have not been identified in previous guidelines.

5.1.2 Self-reported Usability

The second research question investigated the extent to which the self-reported usability of the service reflected the observed usability. This should give an indication whether subjective usability, as participants experience, and their performance, as measured in usability tests, are likely to be correlated.

Overall, the results of P-SUS, UMUX and UMUX-Lite reflected the speed, but not the accuracy of the participants in the usability tests. Participants with intellectual disabilities completed the tasks quickly and gave very high P-SUS scores.

For comparison of the P-SUS scores, Kortum and Bangor [68] published a list of SUS scores of 14 common, everyday products, based on survey data, in which only Google Search scored above 90 (n = 948), while the cohort of participants with intellectual disabilities gave My Kanta an average of 91.3.

While participants with intellectual disabilities reported extremely high experiences of usability using all questionnaires, and were generally quicker completing tasks, they also experienced more issues than non-native Finnish speakers. The non-native Finnish speaking participants reported a more varied experience and average usability, yet experienced fewer observed issues.

The results are unexpected and suggest that the participants with intellectual disabilities were unaware of, or unbothered by the issues they encountered. It is also possible that their assessment of My Kanta was influenced by their experiences with other services, and that the relative usability of My Kanta was much higher than other services the participants had used or tested.

Another possible cause relates to the participants' expectations. Higher initial expectations have been shown to lead to higher subjective usability [69]. Here, most participants are familiar and comfortable with My Kanta, which may have given them especially high expectations. As a result of these, some participants may have felt that the service was more usable despite encountering more issues.

The presence of Kela staff observing the sessions may also have influenced the participants' responses. While all participants were asked to be honest and told that the Kela staff present were not involved in the development of the service, it's possible that participants were afraid of offending those who had worked on the service. Branaghan, O'Brian, Hildebrand, *et al.* [70] note that the mere presence of observers adds stress for participants, which may influence their behavior and performance. In this study, up to four observers were present for some tests. Likewise, Schrier [71] also suggests that participants should be unaware of observers' presence behind a one-way mirror and should interact only with the moderator. Rubin and Chisnell [56], on the other hand, call it "common courtesy" that participant by introduced to everyone present. For ethical reasons, full disclosure was given in this study, as informed consent was required, so the extent of the observation was explained to participants before the session started.

It's also possible that the participants with intellectual disabilities were embarrassed by their difficulties, and did not want to admit to them, so reported a very positive experience to compensate. Schrier [71] noted this risk of embarrassment, but suggested that recruiting "strangers, rather than friends" would reduce this risk. Here, participants were strangers, though they were recruited through a project focused on testing digital services. On the one hand, this may have caused them to be especially critical and precise with their examination, the way an expert tester may, though, on the other hand, this appearance of expertise and experience assessing digital services may cause them to be afraid of appearing as though they don't understand something.

From these results, it seems that self-reported usability measures show different patterns to observation-based testing. The results of the questionnaires may be too subjective to be compared between different cohorts. Since the responses of participants with intellectual disabilities were generally quite consistent, it may be appropriate to use these questionnaires to compare different services or versions of the same service within this cohort, however these results should be treated with caution due to the small sample size.

Without more data, the use of the Finnish language P-SUS and UMUX questionnaires with non-native Finnish speakers cannot be recommended. While the sample is small, these results indicate that the scope for misunderstanding is high and may negatively affect the results.

5.1.3 Comparison of Usability Measures

The third research question explored the relationship between the scores from the different usability measures. While the questionnaires were designed to measure 'usability', the questionnaires mostly use different prompts and different scales. While the results are not directly comparable, they are well-correlated when answered by

participants with intellectual disabilities, but not when answered by non-native Finnish speakers.

The results concerning participants with intellectual disabilities showed a good overall correlation between all three questionnaires, suggesting that the shorter UMUX and UMUX-Lite questionnaires can be used in research in place of SUS, when studying the experiences of users with intellectual disabilities. The result was not statistically significant, which was likely due to the small sample size.

These results are consistent with the original UMUX study by Finstad [4] as well as the UMUX-Lite publication by Lewis, Utesch, and Maher [35], which show a clear, positive correlation between SUS with UMUX and UMUX-Lite. While there are too few data points for the results of this study to be statistically significant, the overall trend matches that which has been demonstrated with high reliability in bigger studies.

These results suggest that the questionnaires may not be suitable for use with non-native Finnish speakers. The reason for this is unclear, though the very small sample size means one participant's misunderstanding of a question can skew the entire cohort's results. The overall result does not match any identified earlier studies. The results might suggest that these participants may not have understood the questions fully, or they may have been confused by the negative wording of some of the questions.

Analysis of answers gives some insights. For example, the correlation between answers to prompt 7 of P-SUS and prompt 2 of UMUX, as presented in Section 4.4.2. Here, all participants, in both cohorts agreed or strongly agreed that they would expect new users to learn to use My Kanta quickly. Participants with intellectual disabilities mostly strongly disagreed with the statement that My Kanta was frustrating to use, while most non-native Finnish speakers agreed with this statement. It is unclear whether the non-native Finnish speaking participants had difficulty understanding one of these statements, or genuinely found the software both frustrating to use (perhaps due to the interface language) and imagined most users would learn to use it quickly.

It is possible that at least one non-native Finnish speaking participant had great difficulties understanding the questions, though did not ask for help. The observation, mentioned in Section 4.4.2, that one participant strongly disagreed with the statement that "My Kanta is easy to use" while also agreeing with the statement "I think that My Kanta is easy to use", suggests that this participant may not have understood the statements. Since one participant misunderstanding the statement could have a drastic effect on the correlation, all results from this cohort should be treated with caution.

Since most answers were not obvious mistakes, it's not possible to isolate specific questions or questionnaires which performed unexpectedly. It's possible that with more participants, a study could be powered to detect under-performing questions with this cohort of participants. Until further research can be undertaken, none of the Finnish-language self-administered usability questionnaires can be recommended for use with non-native Finnish speakers.

5.2 Evaluation of the Study

The experimental design mirrored similar earlier studies [21], [36] where participants test the same service, then answer quantitative usability questionnaires. This gives

insights both into the service and into the performance of the questionnaires.

Participants' health literacy scores showed an unexpected pattern. For example, the scores of the participants with intellectual disabilities showed that all participants were comfortable managing their own health, with three of the four participants scoring over 40, including one scoring the maximum of 50. In contrast, in the study of older Finns by Eronen, Paakkari, Portegijs, *et al.* [61], mean score was 35.9.

The non-native Finnish speaking participants, however, got much lower scores on average. The highest score was 36.6, while the lowest was 11.5. The mean was 28.9 (s.d. 11.8), which is clearly lower than the participants from the study by Eronen, Paakkari, Portegijs, *et al.* [61].

The health literacy questionnaire results may also reflect the role that language plays in understanding one's own health and accessing healthcare. The participants with intellectual disabilities can access the entire healthcare journey in their native language, and professionals are able to adapt their language to the needs of the patient. While services in English are widely available in Finland, some information, services and tasks are only available in Finnish and Swedish. Even where patients can access services in English, no participants in this study spoke English as their native language, so they would still be working through a second language.

Some questions in the health literacy questionnaire were skipped by participants in both cohorts. These questions were the only questions to mention information from the media. This was expected, as these results match those of Eronen, Paakkari, Portegijs, *et al.* [61], who found that the older adults studied most frequently described these questions as difficult and chose "I don't know" in the questionnaire. Based on the small number of participants in this study, it seems that adding the explanatory text, "(for example, television, newspapers)" didn't aid the understanding of these statements. Perhaps participants know the meaning of the word media but are unclear about how they receive health information from it.

Finally, while the health literacy questionnaire [54] has been widely used and validated, it is important to note that it measures only participants' perceptions of their health literacy. In other measures of subject-specific literacy, the participant perception may not always align with a more objective measure. For example, when investigating computer literacy, Merritt, Smith, and Di Renzo [72] noted that there was a statistically significant difference between participants' reported computer literacy and objective computer literacy measured using an interactive tool. Crocker, Feng, and Duncan [73] performed a literature review seeking to identify a similar, performance-based measure of eHealth literacy, and found 29 different approaches from publications. These included quizzing participants on their knowledge and observing them carrying out an internet-based research task. No studies have been identified which propose an objective measure of health literacy, which is not specifically eHealth literacy.

The results concerning the first research question result in the list of recommendations shown in Section 5.1.1. While these recommendations originate from My Kanta, the service used in this study, they all relate to the platform (in this case, the web), rather than to aspects unique to the service. Some of the recommendations match or extend elements of the Web Content Accessibility Guidelines [39]. This strongly suggests that they reflect widely recognized issues and support common best practices,

and that the results concerning this research question are applicable to a wide range of services.

The results concerning the subsequent research questions required quantitative analysis by cohort. In these cases, it's necessary to consider the cohorts in turn.

The relationship between the self-reported usability scores and the observed factors, like completion time and number of issues encountered did not completely follow the trends seen in earlier work [6] with the participants with intellectual disabilities. These participants rated the usability extremely highly, while completing tasks quickly, yet also encountering clear difficulties. Participants who spoke Finnish non-natively, on the other hand, showed results which did seem to follow the trends of earlier work, with average usability scores and some issues encountered.

The relationship between the different usability measures showed that the participants with intellectual disabilities answered the questionnaires consistently between the questionnaires, and all measures seem to work similarly with members of this group. This supports the results of earlier work [4], [34]. These results, however, were not statistically significant, due to the small sample size. The participants who spoke Finnish non-natively, on the other hand, did not show the same trend, which is a pattern not identified in earlier literature.

5.3 Limitations

While the participants with intellectual disabilities provided an insight into how they would use the service, it is unclear to what extent this can be generalized due to the way they were recruited. These participants seemed extremely confident, both using the service and in their self-assessed health literacy. The participants sometimes acted more like professional testers, sometimes making comments outside of the scope of the task, such as "what if I was using a screen reader?". This may reflect the experience that these participants bring, having worked with a project which also involves accessibility more broadly.

The two cohorts studied were distinct and both had unique issues, experiences and patterns of use. Each cohort comprised four participants. A review by Hwang and Salvendy [74] has suggested that in a "think aloud" style usability test, at least nine participants are required to discover 80% of issues. If the two cohorts studied here are expected to experience different issues with the service, neither cohort is close to this threshold.

Likewise, some qualitative questions were asked which it was not possible to answer with statistical significance. Limitations were encountered when calculating reliability and correlations, which may have been avoided, or at least minimized, with more participants.

The experiences of any minority group can vary significantly between members of that group, and it was clear that the participants with intellectual disabilities involved in this study were somewhat similar. There was a little more diversity among the non-native Finnish speaking participants, however some aspects, like level of education and comfort with technology were also similar.

On the other hand, within the non-native Finnish speaking cohort, there were clear differences in some important aspects, including their command of Finnish. It needs to be considered that treating non-native Finnish speakers as a monolith oversimplifies the situation. Some non-native Finnish speakers can have close to native-level skills, while others can rely on automatic translation, and it is unreasonable to expect significant similarities between the experiences of these groups.

Specifically concerning the non-native Finnish speaking participants, it is unclear exactly how much they could understand, since quantitative information about their language skills was not collected. Language skills are difficult to quantify and can vary depending on the scenario and how comfortable a participant is with the subject matter. The most reliable way to quantify language skills, with a standardized test, is not practical to include in a study such as this. Relying exclusively on language examinations participants have taken earlier selects in favor of those who have needed to complete such an exam for work or immigration purposes, and the level indicated may not reflect their level at the time of the experiment.

Additionally, the quality of the "thinking aloud" by non-native Finnish speaking participants, compared to the participants with intellectual disabilities was clearly lower, likely due to the sessions being run mostly in Finnish. To capture their thoughts as accurately as possible, participants should be free to verbalize their thoughts in the language that comes most naturally to them, though this would place an additional burden on the analysis of the results.

Finally, one important limitation came from the design of the experiment. Participants were not told whether they had completed a task correctly, to avoid instructing them in how to complete a task immediately before they tried the task a second or third time. This was valuable, as it allowed meaningful measurement of whether they made fewer errors or got quicker on subsequent attempts without coaching. Conversely, it meant that they were not aware of some of their mistakes when they answered the questionnaires. Some answered as though using the service was very easy, with one stating aloud how easy the tasks had been, having made several errors completing the tasks.

When used for real, the service would not give immediate feedback on some issues. Mistakenly thinking that an item is prescribed when it is not will result in a frustrating trip to the pharmacy and sending a renewal request to the wrong health care unit will result in it being rejected several days later. There is currently no accepted best practice to simulate the passing of time in a study like this, eventually give the participant realistic feedback and allow them to try again, as they would outside of an experimental setting days later.

The limited time with each participant meant that tasks and questions were presented in quick succession, which does not reflect the way that users may access self-service pathways and patient portals independently. It is possible with more time to explore and get comfortable with a service, participants would rate the usability differently.

5.4 Further work

Repeating this type of study with more participants in the non-native cohort, and perhaps with participants who meet a certain level of language skill, could yield more useful results. It would be useful to understand whether, for example, users who use a certain level of Finnish in their everyday life give more reliable answers to the self-administered usability measures.

Similarly, 'intellectual disabilities' are a broad, diverse category, so limiting this study to four participants with similar intellectual disabilities has left a breadth of potential experiences unexplored. Labelling and categorizing disabilities is difficult. Relying on medical diagnoses fails to account for the natural variation that can occur among populations with the same condition. The International Classification of Functioning, Disability, and Health [75] seeks to describe the functional effects of disabilities on people, and may be a good way to categorize participants, however application of this system would likely require the involvement of a health professional.

In some cases, it's unclear whether participants who speak Finnish non-natively did not understand some questions, or there is an aspect of usability which is not being recognized which causes their unexpected performance here. One possible approach to establish this would be to follow the experiment in this study with a review interview, in the participant's native language, where their answers are discussed. They can then explain, for example, which aspect of the service was frustrating. This way, if the problem concerns their understanding of the questions, specific issues can be identified.

One key limitation concerned the lack of feedback and, in turn, the lack of realism in the participants' experience. As noted by several earlier authors [25], [56], [71], there are limitations caused by the laboratory environment. As technology has developed, it may now be feasible to run similar usability tests more like a diary study. Participants could be loaned a smartphone upon joining the study, which they can use in their own home, on which audio and video could also be captured (the ethics and consent elements would need to be carefully considered and informed). Then, for example, after submitting a prescription renewal request, they would later receive a response telling them whether the request would have been approved.

Using a smartphone means the participant is comfortable in their own environment, has access to the type of support they usually would, and a delay like "3-5 working days", as it often takes for a prescription renewal request to be handled, could be followed realistically. The longer delay between repeats of the same task would give better insights into how well participants learn to use the service than them repeating tasks almost immediately after completing them.

Finally, it would be useful to understand the overall trend of P-SUS scores from test participants with intellectual disabilities, like those studied here. To understand this distribution, a larger-scale study, perhaps similar to that by Kortum and Bangor [68], in which everyday products are assessed with a large number of users, but focused on users with intellectual disabilities, would be valuable. It is possible that the distribution is bimodal, or it may be normally distributed with a wider spread than research on larger cohorts of users suggests. The same technique could be applied to other usability

measures, UMUX and UMUX-Lite in this study, however such comprehensive studies with typical users with these questionnaires have not been identified.

6 Conclusions

This study set out to establish how digital services can meet the needs of users with intellectual disabilities and non-native speakers of the service language, to examine whether these users' observed experiences matched the experiences captured by self-reported usability questionnaires, and to see whether the scores from these questionnaires were consistent among these groups of users. It aimed to provide the first foundations of an evidence base for future studies involving these groups of participants and a recommendation for which methods of measuring usability may suit the needs of these users.

To achieve this, a series of usability tests were carried out on a new digital service. The performance of participants was recorded and observed throughout, and the performance and issues experienced were analyzed both quantitatively and qualitatively. Additionally, all participants answered two usability questionnaires, allowing three usability scoring tools to be used to quantify the experiences of these users. The scores and the relationship between them were analyzed to identify possible correlations and trends. There is already a good evidence base for the use of laboratory-based usability tests with many user groups, and for the application of self-administered usability questionnaires with typical users.

Since this was a small study, with four participants in each cohort, the findings should be approached with caution. The participants with intellectual disabilities produced consistent results on most measures, but only represent a small subset of the wide variety of intellectual disabilities and their effects, while the non-native Finnish speakers had such different levels of Finnish that this may contribute to the variety seen in their results.

These results identify some potential improvements to the design of the studied service, which may improve accessibility for users with intellectual disabilities and usability for non-native Finnish-speaking users. Improvements like these, which concern the how users understand the service and how to use it, are likely to significantly improve the experience of these user groups, while having a potential positive effect on all users.

The potential improvements can be generalized into guidelines which may be applied to other services. The derived guidelines are explained further in Section 5.1.1).

Additionally, within a single service, these results suggest that P-SUS, UMUX and UMUX-Lite may be suitable for self-administration by users with intellectual disabilities, however their scores did not correlate with users more generally, so comparing usability metric performance of these users with those from the wider population is not recommended.

Without further data, the use of the Finnish versions of P-SUS, UMUX and UMUX-Lite with participants known to speak Finnish non-natively cannot be recommended, due to the inconsistent results between the three questionnaires observed in this study. Further work could clarify the results concerning non-native Finnish speakers by focusing on a single level of Finnish language skills at a time. Such work may establish the relative performance of these questionnaires with participants of different skill

levels, and with which participants they should work predictably.

Additionally, an expanded study involving participants with intellectual disabilities could cover a broader range of functional effects and may build a better picture of how different types of intellectual disability may affect both the user of the service and the usability metrics.

References

[1] *Ergonomics of human-system interaction*, ISO 9241-210, 2019.

[2] I. Maramba, A. Chatterjee, and C. Newman, "Methods of usability testing in the development of eHealth applications: A scoping review", *International Journal of Medical Informatics*, vol. 126, pp. 95–104, Jun. 2019. DOI: 10.1016/j.ijmedinf.2019.03.018.

[3] J. Brooke, "SUS: A 'quick and dirty' usability scale", in *Usability evaluation in industry*, London: Taylor & Francis, 1996, pp. 189–194.

[4] K. Finstad, "The Usability Metric for User Experience", *Interacting with Computers*, vol. 22, no. 5, pp. 323–327, May 2010. DOI: 10.1016/j.intcom.2010.04.004.

[5] J. Sauro and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*. Massachusetts: Morgan Kaufmann, 2016.

[6] P. Kortum and S. C. Peres, "The Relationship Between System Effectiveness and Subjective Usability Scores Using the System Usability Scale", *International Journal of Human-Computer Interaction*, vol. 30, no. 7, pp. 575–584, Jul. 2014. DOI: 10.1080/10447318.2014.904177.

[7] J. R. Lewis, "Critical Review of 'The Usability Metric for User Experience'", *Interacting with Computers*, vol. 25, no. 4, pp. 320–324, Jul. 2013. DOI: 10.1093/iwc/iwt013.

[8] World Health Organization, *Disability Fact Sheet*, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/disability-and-health.

[9] A. Shrestha, "Immigrants' access and utilization of health care services in Finland: Maamu study", M.S. thesis, University of Eastern Finland, Kuopio, 2017.

[10] S. Emont, *Measuring the impact of patient portals*, 2011. [Online]. Available: https://www.mnhospitals.org/Portals/0/Documents/patientsafety/Perinatal/PDF%20MeasuringImpactPatientPortals.pdf.

[11] M. G. Antonio, O. Petrovskaya, and F. Lau, "The State of Evidence in Patient Portals: Umbrella Review", *Journal of Medical Internet Research*, vol. 22, no. 11, Nov. 2020. DOI: 10.2196/23851.

[12] T. Irizarry, A. DeVito Dabbs, and C. R. Curran, "Patient Portals and Patient Engagement: A State of the Science Review", *Journal of Medical Internet Research*, vol. 17, no. 6, Jun. 2015. DOI: 10.2196/jmir.4255.

[13] C. L. Goldzweig, G. Orshansky, N. M. Paige, *et al.*, "Electronic patient portals: Evidence on health outcomes, satisfaction, efficiency, and attitudes: A systematic review", *Annals of internal medicine*, vol. 159, no. 10, pp. 677–687, Nov. 2013. DOI: 10.7326/0003-4819-159-10-201311190-00006.

[14] S. Kujala, I. Hörhammer, A. Väyrynen, *et al.*, "Patients' Experiences of Web-Based Access to Electronic Health Records in Finland: Cross-sectional Survey", *J Med Internet Res*, vol. 24, no. 6, Jun. 2022. DOI: 10.2196/37438.

[15] E. Carini, L. Villani, A. M. Pezzullo, *et al.*, "The Impact of Digital Patient Portals on Health Outcomes, System Efficiency, and Patient Attitudes: Updated Systematic Literature Review", *Journal of Medical Internet Research*, vol. 23, no. 9, Sep. 2021. DOI: 10.2196/26189.

[16] R. Tsai, E. Bell, H. Woo, K. Baldwin, and M. Pfeffer, "How Patients Use a Patient Portal: An Institutional Case Study of Demographics and Usage Patterns", *Applied Clinical Informatics*, vol. 10, no. 01, pp. 96–102, Jan. 2019. DOI: 10.1055/s-0038-1677528.

[17] J. D. Ralston, D. Carrell, R. Reid, M. Anderson, M. Moran, and J. Hereford, "Patient Web Services Integrated with a Shared Medical Record: Patient Use and Satisfaction", *Journal of the American Medical Informatics Association*, vol. 14, no. 6, pp. 798–806, Nov. 2007. DOI: 10.1197/jamia.M2302.

[18] O. McBride, P. Heslop, G. Glover, *et al.*, "Prevalence estimation of intellectual disability using national administrative and household survey data: The importance of survey question specificity", *International Journal of Population Data Science*, vol. 6, no. 1, Jan. 2021. DOI: 10.23889/ijpds.v6i1.1342.

[19] S. L. Einfeld, L. A. Ellis, and E. Emerson, "Comorbidity of intellectual disability and mental disorder in children and adolescents: A systematic review", *Journal of Intellectual & Developmental Disability*, vol. 36, no. 2, pp. 137–143, Jun. 2011. DOI: 10.1080/13668250.2011.572548.

[20] K. Van Dooren, N. Lennox, and M. Stewart, "Improving access to electronic health records for people with intellectual disability: A qualitative study", *Australian Journal of Primary Health*, vol. 19, no. 4, p. 336, Jul. 2013. DOI: 10.1071/PY13042.

[21] F. Gonzalez Carceller, "Improving the usability of online symptom checkers to avoid the digital exclusion of vulnerable user groups", M.S. thesis, Aalto University, Espoo, 2021.

[22] S. S. Nielsen and A. Krasnik, "Poorer self-perceived health among migrants and ethnic minorities versus the majority population in Europe: A systematic review", *International Journal of Public Health*, vol. 55, no. 5, pp. 357–371, Oct. 2010. DOI: 10.1007/s00038-010-0145-4.

[23] X. Chen, E. Schofield, J. L. Hay, E. A. Waters, M. T. Kiviniemi, and H. Orom, "Race/Ethnicity, Nativity Status, and Patient Portal Access and Use", *Journal of Health Care for the Poor and Underserved*, vol. 32, no. 2, pp. 700–711, 2021. DOI: 10.1353/hpu.2021.0099.

[24] C. K. Yamin, S. Emani, D. H. Williams, *et al.*, "The Digital Divide in Adoption and Use of a Personal Health Record", *Archives of Internal Medicine*, vol. 171, no. 6, Mar. 2011. DOI: 10.1001/archinternmed.2011.34.

[25] M. Moore, R. G. Bias, K. Prentice, R. Fletcher, and T. Vaughn, "Web usability testing with a Hispanic medically underserved population", *Journal of the Medical Library Association: JMLA*, vol. 97, no. 2, p. 114, Apr. 2009. DOI: 10.3163/1536-5050.97.2.008.

[26] J. Nielsen, *Usability engineering*. Massachusetts: Morgan Kaufmann, 1994.

[27] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. Vermeeren, and J. Kort, "Understanding, scoping and defining user experience: A survey approach", in *CHI '09: CHI Conference on Human Factors in Computing Systems*, Boston MA, Apr. 2009, pp. 719–728. DOI: 10.1145/1518701.1518813.

[28] H. M. Hassan and G. H. Galal-Edeen, "From usability to user experience", in *2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Okinawa, Japan, Nov. 2017, pp. 216–222. DOI: 10.1109/ICIIBMS.2017.8279761.

[29] A. Følstad and R. K. Rolfsen, "Measuring the effect of User Experience design changes in e-Commerce web sites: A case on customer guidance", in *2nd COST294-MAUSE International Open Workshop*, Oslo, Norway, Oct. 2006.

[30] M. Broekhuis, L. Van Velsen, and H. Hermens, "Assessing usability of eHealth technology: A comparison of usability benchmarking instruments", *International Journal of Medical Informatics*, vol. 128, pp. 24–31, Aug. 2019. DOI: 10.1016/j.ijmedinf.2019.05.001.

[31] J. Brooke, "SUS: A retrospective", *Journal of usability studies*, vol. 8, no. 2, pp. 29–40, Feb. 2013.

[32] K. Finstad, "The system usability scale and non-native English speakers", *Journal of usability studies*, vol. 1, no. 4, pp. 185–188, Aug. 2006.

[33] J. R. Lewis and J. Sauro, "The Factor Structure of the System Usability Scale", in *First International Conference, HCD 2009*, Berlin, Germany, Aug. 2009, pp. 94–103. DOI: 10.1007/978-3-642-02806-9_12.

[34] J. R. Lewis, "Measuring perceived usability: The CSUQ, SUS, and UMUX", *International Journal of Human–Computer Interaction*, vol. 34, no. 12, pp. 1148–1156, Jan. 2018. DOI: 10.1080/10447318.2017.1418805.

[35] J. R. Lewis, B. S. Utesch, and D. E. Maher, "UMUX-LITE: When There's No Time for the SUS", in *SIGCHI Conference on Human Factors in Computing Systems*, event-place: Paris, France, New York, NY, Apr. 2013, pp. 2099–2102. DOI: 10.1145/2470654.2481287.

[36] S. Borsci, S. Federici, M. L. Mele, and M. Conti, "Short scales of satisfaction assessment: A proxy to involve disabled users in the usability testing of websites", in *17th International Conference on Human-Computer Interaction*, Aug. 2015, pp. 35–42. DOI: 10.1007/978-3-319-21006-3_4.

[37] S. L. Henry, *Introduction to Web Accessibility*, 2024. [Online]. Available: https://www.w3.org/WAI/fundamentals/accessibility-intro/.

[38] L. F. Laux, P. R. McNally, M. G. Paciello, and G. C. Vanderheiden, "Designing the World Wide Web for people with disabilities: A user centered design approach", in *The second annual ACM conference on assistive technologies - Assets '96*, Vancouver, Canada, Apr. 1996, pp. 94–101. DOI: 10.1145/228347.228363.

[39] World Wide Web Consortium, *Web content accessibility guidelines (WCAG) 2.2*, 2023.

[40] World Wide Web Consortium, *Web content accessibility guidelines (WCAG) 2.0*, 2008.

[41] World Wide Web Consortium, *Web content accessibility guidelines (WCAG) 2.1*, 2018.

[42] *Rehabilitation Act, Section 508*, (29 U.S.C. § 798), Amended 1998, 1973.

[43] *The harmonized European Standard for ICT Accessibility*, EN 301 549, 2021.

[44] Web Accessibility Initiative, *Web Accessibility Laws & Policies*, 2023. [Online]. Available: https://www.w3.org/WAI/policies/.

[45] Web Accessibility Initiative, *Understanding Success Criterion 3.1.5: Reading Level*, 2023. [Online]. Available: https://www.w3.org/WAI/WCAG22/Understanding/reading-level (visited on 02/02/2024).

[46] P. Chandler and J. Sweller, "Cognitive Load Theory and the Format of Instruction", *Cognition and Instruction*, vol. 8, no. 4, pp. 293–332, Dec. 1991. DOI: 10.1207/s1532690xci0804_2.

[47] P. D. Blanck, *EQuality: the struggle for web accessibility by persons with cognitive disabilities*. New York: Cambridge University Press, 2016.

[48] S. Roussel, D. Joulia, A. Tricot, and J. Sweller, "Learning subject content through a foreign language should not ignore human cognitive architecture: A cognitive load theory approach", *Learning and Instruction*, vol. 52, pp. 69–79, Dec. 2017. DOI: 10.1016/j.learninstruc.2017.04.007.

[49] R. L. Mace, "What is universal design?", The Center for Universal Design at North Carolina State University, North Carolina, Tech. Rep., 1997.

[50] S. Keates, "Developing BS7000 Part 6 – Guide to Managing Inclusive Design", in *8th ERCIM Workshop on User Interfaces for All*, C. Stary and C. Stephanidis, Eds., Berlin, Germany, Jun. 2004, pp. 332–339. DOI: 10.1007/978-3-540-30111-0_29.

[51] H. Persson, H. Åhman, A. A. Yngling, and J. Gulliksen, "Universal design, inclusive design, accessible design, design for all: Different concepts—one goal? On the concept of accessibility—historical, methodological and philosophical aspects", *Universal Access in the Information Society*, vol. 14, no. 4, pp. 505–526, Nov. 2015. DOI: 10.1007/s10209-014-0358-z.

[52] A. Sonderegger, S. Schmutz, and J. Sauer, "The influence of age in usability testing", *Applied Ergonomics*, vol. 52, pp. 291–300, Jan. 2016. DOI: 10.1016/j.apergo.2015.06.012.

[53] A. Sonderegger and J. Sauer, "The influence of socio-cultural background and product value in usability testing", *Applied Ergonomics*, vol. 44, no. 3, pp. 341–349, May 2013. DOI: 10.1016/j.apergo.2012.09.004.

[54] K. Sørensen, S. Van den Broucke, J. M. Pelikan, *et al.*, "Measuring health literacy in populations: Illuminating the design and development process of the European Health Literacy Survey Questionnaire (HLS-EU-Q)", *BMC Public Health*, vol. 13, no. 1, Oct. 2013. DOI: 10.1186/1471-2458-13-948.

[55] L. Kantner, D. H. Sova, and S. Rosenbaum, "Alternative methods for field usability research", in *SIGDOC03: ACM 21st Annual International Conference on Documentation*, San Francisco, CA, Oct. 2003, pp. 68–72. DOI: 10.1145/944868.944883.

[56] J. Rubin and D. Chisnell, *Handbook of usability testing: How to plan, design, and conduct effective tests*. New Jersey: John Wiley & Sons, 2008.

[57] M. Fan, Q. Zhao, and V. Tibdewal, "Older Adults' Think-Aloud Verbalizations and Speech Features for Identifying User Experience Problems", in *CHI '21: CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, May 2021, pp. 1–13. DOI: 10.1145/3411764.3445680.

[58] S. Denning, D. Hoiem, M. Simpson, and K. Sullivan, "The Value of Thinking-Aloud Protocols in Industry: A Case Study at Microsoft Corporation", in *The Human Factors Society Annual Meeting*, vol. 34, Oct. 1990, pp. 1285–1289. DOI: 10.1177/154193129003401723.

[59] Kanta-palvelut, *Tietoa Kannasta*, in Finnish, 2023. [Online]. Available: https://www.kanta.fi/ammattilaiset/tietoa-kannasta.

[60] Kanta-palvelut, *Ohjeita reseptin uusimispyynnön lähettämiseen*, in Finnish, 2023. [Online]. Available: https://www.kanta.fi/ohjeita-uusimispyynnon-lahettamiseen.

[61] J. Eronen, L. Paakkari, E. Portegijs, and T. Rantanen, "Assessment of health literacy among older Finns", *Aging Clinical and Experimental Research*, vol. 31, no. 4, pp. 549–556, Apr. 2019. DOI: https://doi.org/10.1007/s40520-018-1104-9.

[62] I. Kickbusch, *Health Literacy, the solid facts*. Geneva: World Health Organization, 2013.

[63] J. Sauro and J. R. Lewis, "When designing usability questionnaires, does it hurt to be positive?", in *The SIGCHI conference on human factors in computing systems*, Vancouver, Canada, 2011, pp. 2215–2224. DOI: 10.1145/1978942.1979266.

[64] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2023.

[65] L. J. Cronbach, "Coefficient alpha and the internal structure of tests", *psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951. DOI: 10.1007/BF02310555.

[66] Web Accessibility Initiative, *Understanding Success Criterion 3.2.4: Consistent Identification*, 2023. [Online]. Available: https://www.w3.org/WAI/WCAG21/Understanding/consistent-identification.html.

[67] Web Accessibility Initiative, *Understanding Success Criterion 3.3.3: Error Suggestion*, 2023. [Online]. Available: https://www.w3.org/WAI/WCAG21/Understanding/error-suggestion.

[68] P. T. Kortum and A. Bangor, "Usability ratings for everyday products measured with the system usability scale", *International Journal of Human-Computer Interaction*, vol. 29, no. 2, pp. 67–76, Jan. 2013. DOI: 10.1080/10447318.2012.681221.

[69] S. Kujala, R. Mugge, and T. Miron-Shatz, "The role of expectations in service evaluation: A longitudinal study of a proximity mobile payment service", *International Journal of Human-Computer Studies*, vol. 98, pp. 51–61, Feb. 2017. DOI: 10.1016/j.ijhcs.2016.09.011.

[70] R. J. Branaghan, J. S. O'Brian, E. A. Hildebrand, *et al.*, "Usability evaluation", in *Humanizing Healthcare–Human Factors for Medical Device Design*, New York: Springer, 2021, pp. 69–96.

[71] J. R. Schrier, "Reducing Stress Associated with Participating in a Usability Test", in *The Human Factors Society Annual Meeting*, vol. 36, Oct. 1992, pp. 1210–1214. DOI: 10.1177/154193129203601606.

[72] K. Merritt, K. D. Smith, and J. C. Di Renzo, "An Investigation of self-reported computer literacy: Is it reliable?", *Issues In Information Systems*, vol. 6, no. 1, pp. 289–295, 2005. DOI: 10.48009/1_iis_2005_289-295.

[73] B. Crocker, O. Feng, and L. R. Duncan, "Performance-Based Measurement of eHealth Literacy: Systematic Scoping Review", *Journal of Medical Internet Research*, vol. 25, Jun. 2023. DOI: 10.2196/44602.

[74] W. Hwang and G. Salvendy, "Number of people required for usability evaluation: The 10±2 rule", *Communications of the ACM*, vol. 53, no. 5, pp. 130–133, May 2010. DOI: 10.1145/1735223.1735255.

[75] World Health Organization, *International Classification of Functioning, Disability and Health*, Geneva, 2001.

A HLS-EU-Q16 Questions

The following are the questions from the HLS-EU-Q16 questionnaire, in Finnish as presented to participants in the study, along with English translations for reference.

The questionnaire began with the question “Kuinka helpoksi kuvailisit seuraavia asioita?” (“How easy would you describe the following things?”). The options were “hyvin vaikeaa” (“very difficult”), “melko vaikeaa” (“quite difficult”), “melko helppoa” (“quite easy”), “hyvin helppoa” (“very easy”) and “en tieda” (“I don’t know”).

1. löytää tietoa sinua huolestuttavien sairauksien hoidosta (find information about symptoms of illnesses that concern you)
2. saada selville, mistä saat ammattiapua, kun olet sairas (find out where to get professional help when you are ill)
3. ymmärtää, mitä lääkärisi sanoo sinulle (understand what a doctor says to you)
4. ymmärtää lääkärisi tai apteekkarin ohjeistusta sinulle määrätyn lääkkeen käytöstä (understand your doctor’s or pharmacist’s instruction on how to take a prescribed medicine)
5. arvioida, milloin saattaa olla tarpeen kysyä myös toisen lääkärin mielipidettä (judge if you may need to get a second opinion from another doctor)
6. käyttää lääkärin antamaa tietoa tehdessäsi sairauteesi liittyviä päätöksiä (use information your doctor gives to you to make decisions about your illness)
7. noudattaa lääkärisi tai apteekkarin ohjeita (act on advice from your doctor or pharmacist)
8. löytää tietoa siitä, kuinka mielenterveysongelmia kuten stressiä tai masennusta hallitaan (find information on how to handle mental health problems such as stress or depression)
9. ymmärtää varoituksia esimerkiksi tupakoinnin, vähäisen liikunnan ja liiallisen alkoholinkäytön terveyshaitoista (understand information about unhealthy habits such as smoking, low physical activity or drinking too much alcohol)
10. ymmärtää, miksi sinun tarvitsee osallistua terveysseulontoihin (understand information about recommended health screenings or examinations)
11. arvioida, onko median (esim. TV, sanomalehdet) välittämä tieto terveysriskeistä luotettavaa (judge if the information on health risks in the media (e.g. television, newspapers) is reliable)
12. päättää median (esim. TV, sanomalehdet) välittämän tiedon pohjalta, kuinka voit suojautua sairaudelta (decide how you can protect yourself from illness using information from the media (e.g. television, newspapers))

13. ottaa selvää toiminnasta, joka on hyväksi henkiselle hyvinvoinnillesi (find information about activities that are good for your mental health and well-being)

14. ymmärtää perheenjäsenten tai ystävien antamia terveyteen liittyviä neuvoja (understand advice concerning your health from family or friends)

15. ymmärtää median (esim. TV, sanomalehdet) välittämää tietoa siitä, kuinka tullaan terveemmäksi (understand information in the media (e.g. television, newspapers) on how to improve your health)

16. arvioida, mikä päivittäinen toimintasi on yhteydessä terveyteesi (judge which everyday habits affect your health)

B Task Scenarios

There were four types of task, each repeated two or three times. The tasks were:

1. Check the amount remaining on a specific prescription.
2. Check the instructions (dose or reason for use) of a specific prescription.
3. Send a renewal request for a specific item.
4. Send a renewal request containing multiple items.

They were presented in an order so that they may follow naturally from one to the next, and tasks of the same type were not repeated immediately.

Task 1, attempt 1

I want you to imagine that you have Asthma, and you've been using Ventoline for a few weeks and you notice that it is running low. You want to get more, so that you don't run out. You can't remember whether the doctor prescribed one or two packages.

When you saw the doctor, he/she told you that you can see information about your prescriptions in My Kanta.

Check whether you're able to collect any more from the pharmacy, or whether you will need a new prescription.

Task 2, attempt 1

You've been using Melatonin for just one week, and you're already running low. This seems too quick, so you wonder "have I used it correctly?". You've taken one tablet every day.

Can you find the doctor's instructions somewhere, and can you tell me whether you were using it correctly?

Task 1, attempt 2

Next, let's check Sirdalud, a muscle relaxant. Can you find some information about your prescription and tell me whether there is any ready to be collected from the pharmacy?

Task 3, attempt 1, starting from the prescription information page

Now that you know that there are no more boxes remaining, you know you'll need to request more. You remember when you collected the last box, the pharmacist told you that you can send a request to renew the prescription in My Kanta.

The name of your health center is written here. Send a request to your health center to renew this prescription.

For the purpose of this test, let's pretend that this is your mobile phone number. [Show a safe number on a piece of paper]

Task 2, attempt 2

Ok, let's imagine a different situation. You've recently had several different health issues and seen different doctors. You're trying to make sense of all the different instructions, but you were a little overwhelmed when you originally went to the appointments.

First, you've got a packet of Dasselta, but you can't remember which issues these were prescribed for. Can you find this information from My Kanta?

Task 4, attempt 1, from prescription list

This time, I'd like you to imagine that you're going to visit family for a month soon. You don't want to run out of medication while you're away, so you need to renew the prescriptions for two different items. Can you send one renewal request to your health center which includes both Panadol and Ventoline?

Task 3, attempt 2, from prescriptions list

Now it's a different day and you only need to renew a single item. Can you send a renewal request for Panacod?

Task 1, attempt 3

Next, you're running low on Duact. Can you check whether you've got any more available to collect from the pharmacy?

Task 4, starting from the item information page

And now you want to send a renewal request for the Duact, using the same details as before.

[After they start the task] Now you remember that you will also need some Burana at the same time. Can you add that to this request?

C P-SUS Questionnaire

The questionnaire was presented to participants exclusively in Finnish, the English translations are presented here for reference.

Before answering the Positive System Usability Scale (P-SUS) questionnaire, the following text was shown: “Valitse yksi vaihtoehto, joka kuvaa sinun kokemustasi OmaKannasta tänään. Valitse ensimmäinen reaktiosi - sinun ei tarvitse harkita pitkään”, which translates as “Choose one option which describes your experience with My Kanta today. Choose your first reaction - you don’t need to reflect for a long time”.

Column 1 was labelled “Täysin eri mieltä” (“Completely disagree”) and column 5 “Täysin samaa mieltä” (“Completely agree”), other columns were not labelled.

	1	2	3	4	5
Käyttäisin mielelläni OmaKantaa usein (I think that I would use My Kanta frequently)					
Koin OmaKannan olevan yksinkertainen (I found My Kanta to be simple)					
OmaKantaa oli mielestäni helppo käyttää (I thought My Kanta was easy to use)					
Osaisin käyttää OmaKantaa ilman teknisen henkilön opastusta (I think that I could use My Kanta without the support of a technical person)					
Mielestäni OmaKannan eri osat toimivat keskenään hyvin yhteen (I found the various functions in My Kanta were well integrated)					
Mielestäni OmaKannan eri osat toimivat samalla tavalla (I thought there was a lot of consistency in My Kanta)					
Kuvittelen, että useimmat oppisivat OmaKannan käytön erittäin nopeasti (I would imagine that most people would learn to use My Kanta very quickly)					
Mielestäni OmaKannan käyttö oli erittäin intuitiivista (= oli erittäin helppo arvata, miten se toimii) (I found My Kanta very intuitive)					
Tunsin itseni hyvin varmaksi, kun käytin OmaKantaa (I felt very confident using My Kanta)					
Osaisin käyttää OmaKantaa ilman, että minun täytyy opetella mitään uusia asioita (I could use My Kanta without having to learn anything new)					

D UMUX Questionnaire

The questionnaire was presented to participants exclusively in Finnish, the English translations are presented here for reference.

Before the UMUX questionnaire, participants were shown a prompt which read “Arvioi kokemuksesi perusteella OmaKantaa” (“based on your experience, evaluate OmaKanta”).

Column 1 was labelled “Täysin eri mieltä” (“Completely disagree”) and column 7 “Täysin samaa mieltä” (“Completely agree”), other columns were not labelled.

	1	2	3	4	5	6	7
OmaKannan ominaisuudet vastaavat tarpeitani (My Kanta’s capabilities met my requirements)							
OmaKannan käyttäminen on turhauttava kokemus (Using My Kanta is a frustrating experience)							
OmaKanta on helppokäyttöinen (My Kanta is easy to use)							
Joudun käyttämään liian paljon aikaa asioiden korjaamiseen kun käytän OmaKantaa (I have to spend too much time correcting things with My Kanta)							

E Identified Usability Issues

Task Type 1

In these tasks, using different phrasing, participants were asked whether they had a valid prescription with some of the medication available to purchase. In order to succeed, they must have an amount remaining greater than zero.

Mistook expiry date for availability Inaccuracy

Description: When asked to identify whether a prescribed medication could be collected from a pharmacy, the participant noticed only the expiry date, which is far in the future, and not the amount remaining. As a result, they believed that medication was available to collect when the prescription had been fully delivered and subsequently failed the task.

Recommendation: Consider whether the expiry date needs to be shown, especially in the prescription list, when the prescription is fully delivered.

Occurrences per task: 1: ID: 0, NN: 1 3: ID: 0, NN: 1 8: ID: 0, NN: 0

Task Type 2

In these tasks, the participant was asked a question about their fictitious use of medications, which they could find the answer to from the prescriptions section. Task two asked them to check the usage instructions (listed as dosage, *annostus*, in the interface) for a prescription. Task five asked them to identify the ailment for which a prescription was prescribed (listed as reason for use, *käyttötarkoitus* in the interface).

Followed 'old version' link Delay

Description: Starting from the confirmation that the renewal request is sent, the participant clicked the link labelled "Back to the old version of My Kanta" (*Palaa takaisin OmaKannan vanhaan versioon*) instead of the "Prescriptions" (*Reseptit*) link, when they were asked to find information on a current prescription. They had previously been told that all activities were to be completed in the new version. They may have focused on the "Go back to" (*Palaa takaisin*) part of the link text when deciding to follow it. The participant reached an error page and went back without intervention, so this issue did not cause a failure of the task.

Recommendation: Consider the prominence of this option, compared to the navigation items, and whether the intended navigation item is, in fact, being recognized as a link. This is, however, less important, as this link won't be part of the final version once the old version is removed.

Occurrences per task: 2: ID: 0, NN: 0 5: ID: 0, NN: 1

Did not recognize the "Prescriptions" link Intervention

Description: Starting from the confirmation that the renewal request is sent, the participant was unable to recognize that "Prescriptions" (*Reseptit*) was a link, as it is styled distinctly from the active "Prescription renewal" (*Reseptien uusiminen*) link. The participant tried many possible links, and was unable to proceed without intervention, so the task was subsequently recorded as failed
Recommendation: Interface elements which perform the same function, such as links in an navigation area, should be styled consistently so they can be recognized having the same function. This recommendation is similar to success criterion 3.2.4 of WCAG 2.1 and later [66].
Occurrences per task: 2: ID: 0, NN: 0 5: ID: 0, NN: 1

Opened the medicinal product search Intervention
Description: On the prescription information page, the participant was scanning for the reason for use. They did not notice "Reason for use" (*käyttötarkoitus*) in the information near the top of the page, but followed the "are you looking for more information about medications?" (*etsitkö lisätietoa lääkkeistä?*) link at the bottom of the page, which was large and brightly colored. As a result, they arrived on Kela's "medicinal product search" (*lääkehaku*), which primarily provides price information. Some participants then searched the product name in this service. As there were no links back to the service, participants needed guiding back to the original service by closing the new tab.
Recommendation: Consider the prominence and necessity of the link to this service. A less prominent, or more descriptive link, may avoid users following it when the information they are seeking is already visible on the page.
Occurrences per task: 2: ID: 0, NN: 0 5: ID: 1, NN: 2

Task Type 3

In task 3 participants were asked to send a renewal request for a single item to a named fictitious health center.

Accepted the default health center (not corrected) Inaccuracy
Description: When renewing a prescription, the interface showed two options for selecting a health center, a one default health center (based on user profile information) and "other recipients" (*Muut vastaanottajat*). The health center participants were instructed to use was never shown as the default option, so participants were always required to select 'other recipients'. These participants selected the default and did not correct it themselves. Notably, while the name of the default and intended recipient were different, neither looked like a realistic health center name, so it is unclear whether users would make this problem when using the service for real.
Recommendation: Collect additional data on how frequently real requests are sent to the incorrect recipient. If it happens frequently, consider removing the default option, and requiring participants to always select an option from the list.
Occurrences per task: 4: ID: 3, NN: 0 7: ID: 1, NN: 0

Accepted the default health center (self-corrected) Delay

Description: As above, though the participant noticed the error at the confirmation stage and corrected the choice. This delayed completion of the task but would not result in the request being rejected.
Recommendation: None.
Occurrences per task: 4: ID: 1, NN: 2 7: ID: 0, NN: 0

Skipped a mandatory question (text message) Delay
Description: The participant did not answer the question in which they were asked whether they wanted a text message notification about their request. This caused submission to fail with an explanation that a mandatory question was missed, but the participant was able to self-correct this.
Recommendation: None. It was clear that once participant read and understood the validation message, they did not make the error again.
Occurrences per task: 4: ID: 1, NN: 0 7: ID: 0, NN: 0

Skipped a mandatory question (health center) Delay
Description: As above, but concerning the choice of health center.
Recommendation: As above, none.
Occurrences per task: 4: ID: 0, NN: 1 7: ID: 0, NN: 0

Opened the feedback form Delay
Description: Every page of the service included a large blue banner at the top, which has a prominent button labelled "Give feedback in the Suomi.fi service" (*Anna palautetta Suomi.fi-palvelussa*). The participant clicked the prominently displayed button without appearing to read the content. From the feedback form they were able to close the tab and continue unaided.
Recommendation: Consider whether it makes sense to give the request for feedback such prominence on every page. Since this issue was self-corrected and not repeated, it can be considered low priority.
Occurrences per task: 4: ID: 1, NN: 0 7: ID: 0, NN: 0

Needed prompting to scroll down Intervention
Description: After selecting an item to add to the request, the participant expected to be taken to the next stage, when they needed to scroll to the bottom of the page to find a "next" (*Seuraava*) button. The participant stated they didn't know how to continue. Once they were told to scroll to the bottom of the page, they were able to continue unaided.
Recommendation: Consider testing a 'sticky' or floating interface element, though it should be noted that this may cause additional difficulties for some users. Alternatively, consider using checkboxes for selecting one or more prescriptions for renewal, rather than buttons, which users may expect to submit the form.
Occurrences per task: 4: ID: 0, NN: 1 7: ID: 0, NN: 0

Task Type 4

In this task, participants were asked to submit a prescription renewal request to the health center named which contained two items. In task six, participants were told both items are the start of the instructions. In task nine, participants were told one

item at in the initial instructions, then, once they reached the confirmation step, they were told to imagine that they suddenly remembered that they needed another item.

Accepted the default health center (not corrected) Inaccuracy

Description: When renewing a prescription, the interface showed two options for selecting a health center, a one default health center (based on user profile information) and "other recipients" (*Muut vastaanottajat*). The health center participants were instructed to use was never shown as the default option, so participants were always required to select 'other recipients'. These participants selected the default and did not correct it themselves. Notably, while the name of the default and intended recipient were different, neither looked like a realistic health center name, so it is unclear whether users would make this problem when using the service for real.

Recommendation: Collect additional data on how frequently real requests are sent to the incorrect recipient. If it happens frequently, consider removing the default option, and requiring participants to always select an option from the list.

Occurrences per task: 6: ID: 1, NN: 0 9: ID: 0, NN: 0

Accepted the default health center (self-corrected) Delay

Description: As above, though the participant noticed the error at the confirmation stage and corrected the choice. This delayed completion of the task, but would not result in the request being rejected.

Recommendation: None.

Occurrences per task: 6: ID: 1, NN: 0 9: ID: 0, NN: 0

Skipped a mandatory question (text message) Delay

Description: The participant did not answer the question in which they were asked whether they wanted a text message notification about their request. This caused submission to fail with an explanation that a mandatory question was missed, but the participant was able to self-correct this.

Recommendation: None. It was clear that once participant read and understood the validation message, they did not make the error again.

Occurrences per task: 6: ID: 1, NN: 0 9: ID: 0, NN: 0

Incorrect item in request (in place of correct item) Inaccuracy

Description: The participant was instructed to select two items, one of which was Panadol. In place of Panadol, they chose Burana and continued the task otherwise successfully.

Recommendation: None. This likely reflects the marketing of the medications rather than anything present in the service.

Occurrences per task: 6: ID: 0, NN: 0 9: ID: 0, NN: 1

Default item left in request (self-corrected) Delay

Description: The participant followed the "prescription renewal" (*reseptien uusiminen*) link from an item information page. The items were pre-selected, possibly due to an earlier attempt, so the participant was delivered directly to step two, "Select renewal request recipient" (*Valitse uusimispyynnön vastaanottaja*). The participant continued with the process, then, at the confirmation stage (*tarkista ja lähetä*), noticed and removed the incorrect item, returning them to the first step.

Recommendation: Identify the cause of the default item in the request. Consider showing the participant step one, including the pre-selected items, when the enter the process.
Occurrences per task: 6: ID: 1, NN: 0 9: ID: 0, NN: 0

Missing item from request (not corrected) Inaccuracy
Description: The participant started the process from the item information page for one of the prescriptions they were asked to renew, they were taken to step one, "select prescriptions to renew" (*valitse uusittavat reseptit*) with no items pre-selected. The instructions were to add two items, and this was repeated several times, however only one item was added. It's possible that the participant thought that following the link from the other item's page would start that request with that item pre-filled.
Recommendation: Consider removing the link to renew prescriptions from the prescription information page if the information cannot reliably be pre-filled.
Occurrences per task: 6: ID: 1, NN: 1 9: ID: 0, NN: 0

Item incorrectly removed from request Inaccuracy
Description: The participant removed an item from their request
Recommendation: None.
Occurrences per task: 6: ID: 0, NN: 1 9: ID: 0, NN: 0

Incorrect item in request (self-corrected) Delay
Description: The participant selected an item with a similar name or the same name and a different strength. They noticed this unprompted at the confirmation step and corrected it.
Recommendation: The default order of prescriptions in the first step of the renewal request process is most recent first, if this were alphabetical, users may see similar names and alternative strengths at the same time and avoid mistakes like this.
Occurrences per task: 6: ID: 1, NN: 0 9: ID: 0, NN: 1

Missed pagination Intervention
Description: The participant was searching for an item in the list of items that could be added to the request, which were split over two pages. Eventually they announced that the item wasn't present, at which point it was suggested he try the next page.
Recommendation: Consider both the number of items shown, the vertical height of items and the size and prominence of the pagination controls.
Occurrences per task: 6: ID: 1, NN: 0 9: ID: 0, NN: 0

Incorrect value in municipality field Intervention
Description: When searching for their health center, the participant typed the name of the health center they had been told to use into the field labelled "municipality" (*kunta*). It appeared that the participant had not read the field label. When they moved on to the next question, the value disappeared, as no municipality was selected, and this visible confused the participant. After the participant was asked to read out the visible field label, the situation became clear, and they continued unaided.
Recommendation: Consider whether the value clearing when focus moves away is the expected behavior.

Occurrences per task: 6: ID: 1, NN: 0 9: ID: 0, NN: 0

Incorrect value in service unit field Intervention

Description: When searching for their health center, the participant typed the name of the health center they had been told to use into the correct field without first selecting a category (public or private), so no values were present in the combobox. The participant seemed unaware that there should have been values present and was surprised when what they typed was not accepted and was automatically cleared. The participant submitted the form with the field empty, then corrected the error when the validation message instructed them of the problem.

Recommendation: Consider whether the value clearing when focus moves away is the expected behavior or whether the field could be hidden until the prerequisite question is answered.

Occurrences per task: 6: ID: 1, NN: 1 9: ID: 0, NN: 0

Participant cancels and starts over to add a second item Delay

Description: When asked to add a second item to the request, or when noticing that an item was incorrect, participants would cancel the request and start over, rather than going back and editing the request.

Recommendation: At the confirmation stage, consider an explicit "edit" button in addition to the "remove selection" (*poista valinta*) button.

Occurrences per task: 6: ID: 0, NN: 1 9: ID: 2, NN: 2

Removes item from request incorrectly Inaccuracy

Description: When asked to add a second item to the request the participant said they wanted to go back, but then clicked the "Remove selection" (*poista valinta*) button on the already added item. They then added the second item, leaving the first item out of the request they sent.

Recommendation: Consider an 'edit' button close to the "remove selection" (*poista valinta*) button, as that appears to be the functionality the participant was expecting here.

Occurrences per task: 6: ID: 0, NN: 0 9: ID: 0, NN: 1

Attempts to click the step heading Delay

Description: When trying to add an item to the request, the participant tried to click the heading of a previous step, which was now visible with no content. The participant seemed to expect clicking it to return them to the named step.

Recommendation: Consider whether a heading with no content makes sense, or whether it should function as a link to that step, like the participant expected.

Occurrences per task: 6: ID: 0, NN: 0 9: ID: 0, NN: 1
