Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Eirik Olav Aa
Fredrik Busklein

Evaluating and monitoring reading proficiency - A digital word chain test

Master's thesis in Computer Science
Supervisor: John Krogstie
June 2023

NTNU
Norwegian University of
Science and Technology

Eirik Olav Aa
Fredrik Busklein

Evaluating and monitoring reading proficiency - A digital word chain test

Master's thesis in Computer Science
Supervisor: John Krogstie
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

NTNU
Kunnskap for en bedre verden

Department of Computer Science

Master Thesis

Evaluating and monitoring reading proficiency - A digital word chain test

Authors:
Eirik Olav Aa and Fredrik Busklein

June 2023

Sammendrag

I følge PISA-undersøkelser fra 2015 og 2018 har antallet norske 15-åringer som sliter med leseferdigheter økt fra 15% til 19% i løpet av tre år. Gutter er overrepresentert i denne gruppen, da 26% av guttene sliter med lesing og skriving, sammenlignet med 12% av jentene. En studie av norske 5- og 6-åringer konkluderte med at det allerede er en forskjell i leseferdigheter mellom kjønnene når barna begynner på skolen. Forskning viser også at en viktig metode når man underviser barn i lesing er å vurdere barnas leseferdigheter og gi dem utfordringer på riktig nivå. I dag er en viktig vurderingsmetode i norske barneskoler og ungdomsskoler en såkalt ordkjedetest, en prøve som utføres på papir der elevene får en rekke ordkjeder (et sammensatt ord som består av fire kortere ord) og skal dele opp hver ordkjede i de fire ordene. Hver test består av 90 ordkjeder og blir rettet manuelt. Målet med denne masteroppgaven er å utvikle et digitalt alternativ til den papirbaserte testen, slik at testene blir mindre kostbare å gjennomføre og mer tilgjengelige for lærere og elever. Forskningsmodellen anvendt i dette prosjektet er Design Science Research med en iterativ prosessmodell. Den digitale testen er blitt utviklet gjennom to iterasjoner og har blitt testet med tanke på både sammenlignbarhet med den papirbaserte testen og generell brukervennlighet.

Resultatet av prosjektet er et digitalt system bestående av to applikasjoner. Hovedapplikasjonen er en nettside der elever kan logge inn og ta en digital ordkjedetest som blir vurdert automatisk. Den andre applikasjonen er et lærerpanel, også i form av en nettside, der lærere kan tildele tester og følge med på fremgangen til enkeltelever eller hele klassen. For å måle hvordan resultatene fra den digitale ordkjedetesten sammenligner seg med resultatene fra den papirbaserte testen, ble det gjennomført en sammenlignbarhetstest. Resultatene fra denne testen viste at det var en høy korrelasjon mellom resultatene etter den andre iterasjonen, med noe usikkerhet. I brukervennlighetstesten ga 12 lærere og spesialpedagoger systemet en SUS-score på 94,2, noe som tilsvarer karakteren A+ på en normalisert karakterskala. Selv om resultatene er lovende, må systemet gjennomgå ytterligere evalueringer med ekte elever i klasserom og integreres i de administrative IKT-systemene som brukes i skolen.

Abstract

According to PISA surveys from 2015 and 2018, the number of Norwegian 15-year-olds that struggle with reading literacy has risen from 15% to 19% over the course of three years. Boys are overrepresented in this group, as 26% of the boys struggle with reading and writing, compared to 12% of the girls. A study on Norwegian 5-and-6-year-olds concluded that a gap in reading literacy exists between the genders already when children start school. Research also shows that an important method when teaching children how to read is to assess their literacy level and provide them with challenges of an appropriate level. Today, an important assessment tool in Norwegian elementary- and middle schools is the word chain test, a test conducted on paper in which pupils are given a set of word chains (a compound word consisting of four shorter words) and tasked to split each word chain into the four words. Each test consists of 90 word chains and is graded manually. The aim of this master thesis is to develop a digital alternative to the paper-based test, thus making the tests less costly to conduct and more available to teachers and pupils. The project follows a Design Science Research approach with an iterative process model. The digital test has been developed over two iterations and tested for both comparability to the paper-based test and general usability.

The result of the project is a digital system comprised of two applications. The main application is a website where pupils can log in and take a digital word chain test that is graded automatically. The second application is a teacher's dashboard, where teachers can assign tests and monitor the progress of individual pupils or a whole class. To measure how results from the digital word chain test compare to results from the paper-based word chain test, a comparability test was conducted. The results from this test showed that there was a high correlation between the results after the second iteration, although with some uncertainty. In the usability test, 12 teachers and special education teachers gave the system a SUS-score of 94.2, achieving a grade of A+ on a normalized grading scale. While the results show promise, further evaluations involving real pupils in classroom settings and integration into existing administrative ICT systems utilized in schools are necessary.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The definition of reading literacy has changed over the course of time. Once seen as simply a skill acquired during the first years of school, it is now understood as an ever-expanding set of knowledge, skills, and strategies built upon through interactions with other people in various contexts. In the PISA 2018 Assessment and Analytical Framework, reading literacy is defined as "...understanding, using, evaluating, reflecting on and engaging with texts in order to achieve one's goals, to develop one's knowledge and potential and to participate in society" [1]. Apart from being a requirement for efficiently obtaining knowledge in further education, Sigmundsson et al. also argue that developing reading literacy as a skill helps intellectual, emotional, and social development in children [2].

Reading literacy is an essential skill set on many levels. Despite this, 19% of all Norwegian 15-year-olds struggle with reading, compared to 15% in 2015 according to PISA 2018. When comparing boys and girls, we see that 26% of the boys belong to this group versus 12% of the girls [3]. A study on Norwegian 5-6-year-olds concluded that a gap in reading literacy exists between the genders already when children start school [4]. According to Csikszentmihalyi, an approach to reducing this gap is to assess the literacy level of each child and use this assessment to provide them with challenges of an appropriate level, and follow closely on their progress [5].

Since the late 1990s, Norwegian schools have used word-chain tests ("ordkjedetester") to assess reading literacy in Norwegian children from third to tenth grade. In these tests, each pupil is given a list of word chains put together by four words, and their task is to identify the separate words within each word chain by writing a line between them. The goal is to correctly decipher as many word chains as possible within a given time. The test has been normalized twice (1997 and 2007) at schools in Rogaland county to set a benchmark for reading literacy at individual grade levels [6][7]. Currently, the tests are conducted with pen and paper and graded manually by the teacher.

Several Norwegian municipalities have recently started issuing laptops to their pupils [8][9]. Meanwhile, Norwegian schools spend resources on buying test sets from publishers, printing the tests on paper for each pupil, grading each test, and recording test results digitally or by hand. With the introduction of laptops at an early stage in elementary schools, we see the opportunity to capitalize on technological advancements and create a platform that can benefit both teachers and pupils. By digitalizing word chain tests, we hope to reduce the resources spent on assessing the pupils' reading literacy and enhance assessment by providing teachers and pupils with detailed results from each test.

1.2 Research Questions

With this application and master thesis, we aim to answer the three following research questions:

Research Question 1: Is the digitalized word chain test a viable substitute for the paper-based test?

Research Question 2: What are some challenges when creating digital word chain tests?

Research Question 3: Do teachers see the value and show interest in using a digitalized version of the word chain test?

1.3 Structure of the report

Chapter 2 Background: This chapter provides an overview of the central theories on learning how to read, followed by a description of different reading literacy assessment methods, including the word-chain test. In addition, it elaborates on the role of technology in education and the theoretical principles underlying the design of user interfaces and login-based applications for children.

Chapter 3 Research Approach: Describes the research method used in this project. The chapter includes a description of the process model utilized to answer the research questions in this project .

Chapter 4 Methods, tools, and technology: This chapter outlines the development method employed to create the two artifacts. Additionally, it presents the various techniques used to assess the product, along with a brief introduction to all the tools and technologies employed to implement the project.

Chapter 5 Implementation: This chapter describes the design of the teacher dashboard and the test app and the implementation of the apps. The two iterations of development will also be described.

Chapter 6 Results: This chapter presents the results from the tests conducted in this project.

Chapter 7 Evaluation: Evaluation of the results presented in the result section and assessing the extent to which the defined project requirements have been fulfilled.

Chapter 8 Discussion, Conclusion, and Further Work: Contains concluding thoughts on the results and outcome of the project as well as a presentation on potential further work.

Chapter 2

Background

This chapter provides the necessary background theory for the thesis. It covers the process of learning to read, as well as different reading literacy assessment methods used in Norway, including the word chain test. Additionally, the chapter describes the use of technology in education and introduces principles for designing user interfaces and login-based applications for children.

2.1 Definition and Importance of Reading Literacy

The Programme for International Student Assessment (PISA) defined reading literacy as "understanding, using, evaluating, reflecting on, and engaging with texts in order to achieve one's goals, develop one's knowledge and potential, and participate in society" [10]. Reading is not only pivotal for academic success but also essential for most employment opportunities [11]. Moreover, reading is important for a variety of reasons in addition to academic success and employment. Amirova argues that reading has numerous benefits, including cognitive development, language skills, increased empathy, and understanding of others [12]. As a substantial portion of the waking hour is spent reading, those unable to read, such as the visually impaired, non-native language readers, and the illiterate, tend to have a lower socioeconomic status and rely on others for assistance [13]. Being able to read is undoubtedly vital in many aspects of life, and two main components must be mastered to be a competent reader: decoding of words and reading comprehension [6].

Decoding constitutes the more technical aspect of reading, as it involves the application of the principles of written language or coding to detect the intended word [6]. After the pupil has learned to decode enough letters, they can start to comprehend written words [14]. This process involves the linkage of letters to their corresponding sounds and happens effortlessly and automatically for the average reader. Being able to read words is one of the most important steps in the process of becoming an adequate reader, as words are the basic units readers use to create meaning out of text [15].

The reading comprehension part requires more high-level mental work [6]. In the comprehension part of reading, the reader has to connect the content they read to their own experiences and references, draw conclusions and make interpretations. Assessing reading comprehension through listening tests can provide insights into a pupil's understanding of the text. If there is a significant difference between a pupil's reading comprehension and listening comprehension, it may indicate that the decoding skills are impeding their ability to comprehend written text [6]. Therefore learning how to decode complete words is essential for young pupils to be competent readers.

Reading involves much more than having efficient decoding skills, but poor decoding skills will be an obstacle in developing proficient reading skills [16]. There are multiple different strategies that can be used when decoding words, depending on if the word occurs alone or in context. When the words occur alone, the following decoding strategies can be used [6]:

- **Logographic reading**: The reader memorizes the word as a visual memory picture, often with more or less random characteristics. Novice readers can recognize several words by exploiting this strategy even before they have learned the letters.
- **Phonological reading**: The words are decoded by dividing the words into smaller letter segments. The segments are decoded phonetically. Followingly, the sound segments are put together as a complete word.
- **Orthographic reading**: It is possible to decode words immediately. The prerequisite for orthographic reading is that the reader has seen the word several times and thus established a memory image for the word in the long-term memory.

Learning how to read depends on many variables, and it is not possible to define a general course of development. [6]. It is therefore important that each pupil is given sufficient attention and it is made sure that the pupils master the skill of reading words in the first couple of years of education. Struggling pupils need to be provided with sufficient and the right assistance to help them master the task of reading [17]. Dyslexia is one reason why some pupils struggle to learn how to read effectively. One of the symptoms of dyslexia is difficulties with decoding words [6]. Dyslexics rarely reach the orthographic level, and the decoding never becomes automated. As a result, the decoding will pull cognitive resources, leaving less to the reading comprehension process [6].

2.2 Reading Literacy in Norway

Norway has a strong focus on promoting reading literacy, and the reading literacy of 15-year-olds in Norway is regularly assessed by PISA. According to PISA, Norway scored among the best school systems in the world, averaging higher reading performance than the OECD (Organisation for Economic Co-operation and Development) average. In contrast, the relationship between reading performance and socioeconomic status was weaker than the average. Norway was also among 14 countries where disadvantaged pupils had at least one-in-five chances of attending the same school as a high achiever, those who scored in the top quarter of reading performance in PISA. [10]. This is an implication that Norway is among the best countries in moderating between-school differences. Those statistics imply that Norway is relatively successful in its reading education. However, there is still room for improvement.

One of the issues in Norwegian reading literacy highlighted in the PISA report from 2018 was the gender gap. As aforementioned, 26% of Norwegian 15-year-old boys struggle with reading, while only 12% of the girls of the same age struggle [3]. The average score on the PISA test was 487 points, with the girls outperforming the boys by 30 points. Norway had a higher average score on the reading test with 499 points, but the gap between genders where also higher than the average. The Norwegian 15-year-old girls scored 47 points better on average than the Norwegian boys [18]. The Norwegian girls scored 9.95% better than the Norwegian boys, compared to the world average of 6.36%.

2.3 Reading Literacy Assessment Methods

Assessing reading literacy is a complex process. It involves evaluating various aspects of reading, such as decoding skills, comprehension, and critical analysis. There are several different reading literacy assessment methods used to measure these skills in Norway. Some of them are presented in this section.

2.3.1 Word chain test (Ordkjedetesten)

The word chain test is a simple screening test to assess pupils' decoding skills. It is a suitable instrument used to measure orthographic reading skills, both for dyslexics and other struggling readers. The test was developed in 1987 as a part of the research project *Leseutvikling i Kronoberg*. Since then, the test has been further developed through several research projects, and in 1996 it was standardized for Swedish pupils attending the grades second through ninth. This project is based on the Norwegian version, which is developed with the same main principles used in the Swedish test. The Norwegian version differs slightly from the Swedish one. The number of words in each word chain in the Swedish test varies, but it is predetermined in the Norwegian version. The *Letter Test*, which is a component of the Swedish word chain test, is excluded from the Norwegian version [6].

Description

The test consists of 90 different word chains. The chains are spread out on three pages with 30 word chains on each page, with 10 lines of three chains. In the Norwegian word chain test, each chain comprises four words. The four words in the chain are not separated, as seen in the example word-chains. It is the pupil's task to divide the four individual words by drawing a line between the words in the chains, demonstrated in completed word-chain. Only the word chains with all four words separated correctly will count as completed, e.g., the unapproved word-chains will not give pupils any points. Thus, the maximum number of points a pupil can achieve is 90. The length of the words in the test varies from two to seven letters, and the words can be nouns, verbs, adjectives, prepositions, or numbers. The test does not require any level of reading comprehension as all of the words are common words that usually exist in the vocabulary of even the youngest pupils.

The duration of the test is four minutes. After the elapse of the four minutes, pupils are not permitted to start dividing new word chains or complete word chains that they have begun dividing. To administer the test, the teacher needs a stopwatch or a similar device to keep track of time. The pupils participating in the test use a colored pencil or a pen to divide the words, which aids in the process of correction of the tests. Erasers are not to be used. The test is completed in a relatively short space of time, and time wasted on sharpening pencils or erasing lines is undesired. The estimated time required for providing instructions and completing the test is approximately 15 minutes.

ordpilvedhvem treoverlivse surminstfriku

Example word-chains

ord|pil|ved|hvem

Completed word-chain

ord|pil|vedhvem tre|overliv|se surm|inst|fri|ku

Unapproved word-chains

Instructions to pupil

The test set includes six practice word chains to help the pupils understand the nature of the test. The word chain test manual includes some instructions on how to go through the practice chains and how to conduct the test itself [6]. The instructions are divided into nine steps, listed below:

1. The teacher writes the three first example word chains: "1) musfemrihar, 2) gårhjemishatt og 3) dagkanhusnå" on the blackboard.

2. The test booklet is handed out to the pupils. The teacher specifies that the booklet is not to be opened before the teacher gives them notice.

3. The teacher demonstrates how to complete the first example task and goes through the four words "mus|fem|ri|har" and tells the pupils to draw the same line on their sheet.

4. Subsequently, the pupils are instructed to try for themselves on the next tasks, and they are reminded of the number of words and lines required for each word chain. After some time, the teacher will show the correct solution and demonstrates common mistakes, such as writing the line after "gå" and not "går".

5. The pupil will try again to complete the next word chain, but without any further instruction. The teacher will show the correct solution after completion.

6. All the pupils will have 30 seconds to test themselves on the three word chains on the next line of the booklet. The teacher shows the solution, and the pupils control their answers. The teacher reminds the pupils that the goal of the test is to complete as many word chains as possible in four minutes and that there are three sheets in total with word chains. Then asks if there are any further questions.

7. The teacher informs the pupils to turn the first page. Reminds them that there are three lines to separate words for each word chain and to start in the top left of the test and go from left to right when completing word chains.

   Then say: "Start"

   Start the timer. After exactly four minutes, say: "Stop".

8. Collect the booklets from the pupils.

2.3.2 LUS (LeseUtviklingsSkjema)

LUS is another assessment method used to asses reading literacy. The tool is based on research on how children learn how to read and can be used to determine the progress in the reading development of each pupil. LUS is used to place all readers at a skill level referred to as a developmental stage. These stages are defined based on the research on how we learn to read. The chart is divided into three different phases; the Exploring phase, the expanding phase, and the literate phase. The two initial phases, the exploring and the expanding phase, consist of several stages. In total, the chart consists of 19 progressive stages, where stage 18 has three sub-stages [19]. The teacher does an assessment of the reading level of the pupil and places the pupil at the highest level the pupil has mastered. When the pupil has progressed through all the steps, they reach the literate phase. The steps and phases are visualized in Appendix A.

LUS is a common aid used in Norwegian schools, and almost all schools in Oslo use LUS to assess pupils reading literacy. The schools that use LUS all receive training and guidance from instructors over a ten months period [20]. LUS does not specify a specific method on how the reading training should be done but helps the teachers to focus on what the pupil can do rather than what they can not do. The schools themselves possess the pedagogical expertise and better understanding of each pupil. They are able to make a more tailor-made solution for each pupil rather than a common practice for all [19]. Schools report that the use of LUS has made a positive impact, including a better overview of pupils' reading competence [20].

2.3.3 Screening test (Kartleggingsprøve)

The screening test for reading examines if the pupils have obtained the expected level of reading literacy, including the reading of words, sentences, and texts, as well as spelling and reading comprehension. This mapping test is mandatory for pupils in the third grade and optional but recommended for the first grade. The goal of the test is to identify pupils who struggle with reading development at an early age. This makes it possible to take measures early on to prevent

future problems and prevent pupils from dropping out later on in their education [21]. The test is developed by specialists from different parts of the education sector in Norway, including *Utdanningsdirektoratet*. Utdanningsdirektoratet is responsible for ensuring that the tests are developed in accordance with the quality requirements in the framework for mapping tests and that they are sufficiently quality-assured.

The mapping test is an adaptive test. Some of the tasks are common to all, while some of the tasks are adapted to the level of each pupil. This is done to ensure that all of the pupils experience mastery. The test contains the same amount of tasks for each pupil. In the third grade, the test consists of approximately 70 tasks, while for the first graders, there are around 60 tasks. There is no time limit on each task or the test in total, but Utdanningsdirektoratet estimates that the duration of the test is about 40 minutes. They also advise the schools to consider giving the pupils a break during the test. As of spring 2023, the mapping test is completed digitally. The test can be completed on either a tablet, a laptop, or a PC [21]. Some examples of the tasks on the digital mapping test are presented in the following section.

Example tasks

In one of the tasks, the pupils are shown a word on the screen, as displayed in Figure 2.1a. The word is shown for a short period of time and is displayed only once. After reading the word, the pupil must find the word they read among the four options provided. In the example Figure 2.1, the displayed word was *sol*, and *sol* is chosen, shown in Figure 2.1b.

(a) Display of word (b) Choosing a word

Figure 2.1: Example task 1

Another task in the test for the third graders is a task where the pupil reads a short text and tries to link the text to the right image. The pupil is presented with four options of images, as shown in Figure 2.2, and the goal is to click on the image that represents what is described in the text.

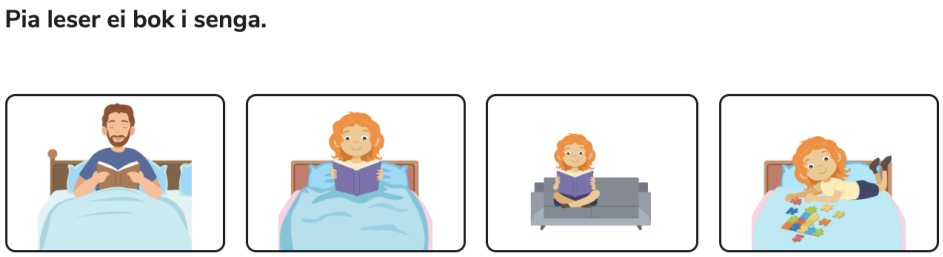


Figure 2.2: Example task 2

After the pupil has completed the assigned tasks, the test will be automatically submitted. The teacher can view the results at *PAS - prøver*. The results display shows pupils that need extra attention. Only the score of the identified struggling pupils will be displayed. For the remaining pupils, it will only show that they have completed the test. If a pupil is identified as need of additional support based on the results of the mapping test, the school will develop an individualized plan for intervention. This could involve additional instruction, practice, or other forms of support [21].

2.3.4 National tests (Nasjonale prøver)

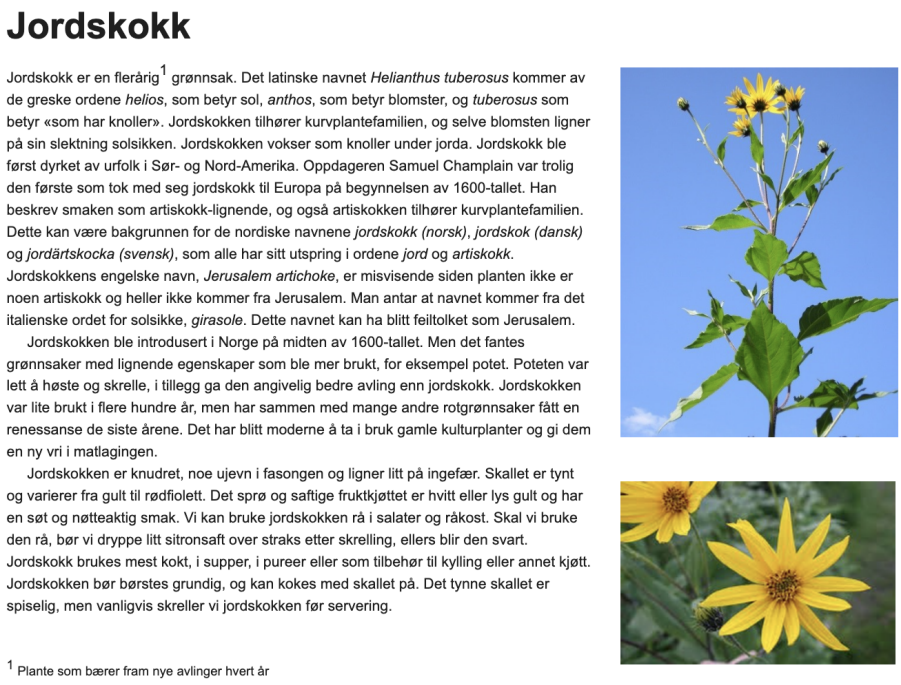
The National Test in reading is also a test from *Utdanningsdirektoratet*. The purpose of this test is to provide the school's insight into their pupil's basic skills in reading. Schools use the results to develop targeted educational strategies for individual pupils and to improve teaching methods and curriculum planning. The teachers can use the results to follow up with pupils and provide adaptive teaching to each pupil. Municipalities and schools can use the results as a basis for quality assessment in Norwegian education, and researchers can use the test results in their research [22].

To make it possible to track development and still change the tasks of the test from year to year, some of the tasks, *Ankeroppgave* (anchor task), are repeated each year. Each pupil receives one or two anchor tasks as part of their test. By using the same anchor tasks each year, the tests can be linked together and compared. By replacing approximately 20 percent of the anchor tasks each year, the entire anchor is renewed every fifth year. The anchor tasks are not made public [22].

Pupils in the fifth, eighth, and ninth grades complete the national reading test. The test is the same for the eighth and nine graders to better facilitate the schools in comparing the results of the pupils and tracking their progress. The test incorporates additional demanding tasks designed to engage and test even the most academically advanced students. The test is digital and consists of both multiple-choice questions and writing tasks.

Example tasks

The pupils are presented with a text to read in advance of the task, as visualized in Figure 2.3. After reading the text, the pupil can proceed to the tasks. There are usually seven texts in total in one test. The number of tasks associated with each text varies but typically ranging from four to seven.



Jordskokk

Jordskokk er en flerårig[1] grønnsak. Det latinske navnet *Helianthus tuberosus* kommer av de greske ordene *helios*, som betyr sol, *anthos*, som betyr blomster, og *tuberosus* som betyr «som har knoller». Jordskokken tilhører kurvplantefamilien, og selve blomsten ligner på sin slektning solsikken. Jordskokken vokser som knoller under jorda. Jordskokk ble først dyrket av urfolk i Sør- og Nord-Amerika. Oppdageren Samuel Champlain var trolig den første som tok med seg jordskokk til Europa på begynnelsen av 1600-tallet. Han beskrev smaken som artiskokk-lignende, og også artiskokken tilhører kurvplantefamilien. Dette kan være bakgrunnen for de nordiske navnene *jordskokk (norsk)*, *jordskok (dansk)* og *jordärtskocka (svensk)*, som alle har sitt utspring i ordene *jord* og *artiskokk*. Jordskokkens engelske navn, *Jerusalem artichoke*, er misvisende siden planten ikke er noen artiskokk og heller ikke kommer fra Jerusalem. Man antar at navnet kommer fra det italienske ordet for solsikke, *girasole*. Dette navnet kan ha blitt feiltolket som Jerusalem.

Jordskokken ble introdusert i Norge på midten av 1600-tallet. Men det fantes grønnsaker med lignende egenskaper som ble mer brukt, for eksempel potet. Poteten var lett å høste og skrelle, i tillegg ga den angivelig bedre avling enn jordskokk. Jordskokken var lite brukt i flere hundre år, men har sammen med mange andre rotgrønnsaker fått en renessanse de siste årene. Det har blitt moderne å ta i bruk gamle kulturplanter og gi dem en ny vri i matlagingen.

Jordskokken er knudret, noe ujevn i fasongen og ligner litt på ingefær. Skallet er tynt og varierer fra gult til rødfiolett. Det sprø og saftige fruktkjøttet er hvitt eller lys gult og har en søt og nøtteaktig smak. Vi kan bruke jordskokken rå i salater og råkost. Skal vi bruke den rå, bør vi dryppe litt sitronsaft over straks etter skrelling, ellers blir den svart. Jordskokk brukes mest kokt, i supper, i pureer eller som tilbehør til kylling eller annet kjøtt. Jordskokken bør børstes grundig, og kan kokes med skallet på. Det tynne skallet er spiselig, men vanligvis skreller vi jordskokken før servering.

[1] Plante som bærer fram nye avlinger hvert år

Figure 2.3: Task

As aforementioned, there are different types of tasks related to each text. Some tasks are multiple-choice tasks and provide pupils with four options when answering a question, as seen in Figure 2.4a. Other tasks require that the pupil formulate their own answer (Figure 2.4b).

(a) Multiple-choice (b) Writing

Figure 2.4: Example task

The mentioned reading assessment methods all aim to detect struggling pupils and intervene early to ensure pupils receive the proper help to avoid falling behind their peers. As the world is getting increasingly digitalized, so are these tests and the rest of the education. Both the national test and the screening test have been digitalized in recent years.

2.4 The Role of Technology in Education

The 21st century is often regarded as the era of technology, and it affects more or less every aspect of life. Education is no exception to this. The presence of educational technology is growing in the classroom, and the new generation of kids is ready to work with these new technologies. Today's children are more used to modern technical equipment from an early age, and new technology should not be a problem for them. Still, it is important that the primary focus is the educational value. The technology should enhance the learning experience and provide resources, opportunities,

and tools that otherwise would be unavailable to them. It is vital that the nature of learning should drive the use of technology, not the other way around [23]. L. Stošić lists five areas of software programs that have the potential to strongly influence the learning experience for children [24].

- The educational value of the program
- Its ability to engage children
- Ease of use
- Interactivity between program and child
- The possibility that the program monitors the progress of the child

As per the previously cited PISA survey [10], access to new technology increases at a remarkable rate, with pupils spending about 3 hours online on weekdays and 3.5 hours online on weekend days. While improved access to technology provides new opportunities, it also raises the standard of literacy proficiency [10]. Thus, it is essential to exploit the opportunities technology provides. To be able to do so, it is vital that teachers get the opportunity to develop their digital competencies in line with the rapid technological advancements in society. The systematic use of digital technology in classrooms requires that schools facilitate such use, that teachers are encouraged to utilize the technology, that they have a positive attitude towards digital technology, and that they receive adequate training in using such technology for teaching [25].

2.4.1 In Norway

A significant majority of pupils in Norway own personal digital devices. 8 out of 10 pupils in grades 1 through 4, nine out of ten in grades 5 through 7, and nearly all pupils (98%) in grades 8 through 10 possess a digital device of some sort. In addition to this, Norwegian schools issue digital devices to their pupils [8][9]. iPad is the most common device at the younger level, while PC or Chromebook is more common among the older pupils [26].

The Cooperation-project *GrunnDig - Digitalisering I grunnopplæring: kunnskaper, trender og framtidig forskningsbehov* examined the use of tablets in education and their impact on teaching methods and pupil learning. The study concluded that, while the introduction of iPads did not result in notable changes to teaching methods, the overall impact was still more positive than negative. This may be connected to the positive effect the tablets had on pupil motivation [25]. The pupils are used to using electronic devices and enjoy using them. Even though it may hamper their focus, the total effect on learning is positive.

2.5 Designing User Interfaces for Children

When designing an assessment tool for children, one must consider that young users may lack the cognitive abilities we assume of adults. Having their reading skills evaluated can be stressful enough, so the user interface should put as little cognitive load on the pupils as possible. In addition, if the pupils have difficulties using the assessment application, the teacher will have to spend extra time helping them, which can stall the whole class. The following paragraphs will describe our findings on creating a learning environment with a low cognitive load by balancing multiple aspects of assessment- and learning environment design.

In *Elements of Effective e-Learning Design*, to keep pupils motivated, Brown and Voltz suggest creating a scenario, a context in which the tasks take place and have meaning [27]. Assuming that young children will have difficulties following text instructions and engaging in a textual scenario, the application can benefit from the extended use of multimedia such as images, icons, and music. Although not directly conflicting, this must be done while keeping the findings from *User Interface Design for E-Learning Software* in mind, which argues that the optimal environment for learning

in electronic applications is well organized and eliminates unnecessary distractions, like music and animated figures [28]. The findings in these two articles and the assumptions of a higher need for audiovisual aids suggest that application designers must find a way to use such aids without cluttering the learning environment.

Another aspect of designing interfaces for children is the balance of flat hierarchical structure and low cluttering. While Faghih et al. emphasize the need for a learning environment with as few distractions as possible [28], *Interface design for children's searching and browsing* [29] has a different angle on the cognitive load on children linked to the hierarchical structure of information. While browsing and searching for information in a user interface, adults can easily handle and take advantage of utilities like search bars, filtering, categories, and custom queries. Such utilities allow the designers to minimize clutter and reduce the cognitive load the interface poses on the user by organizing content hierarchically. However, young children lack the cognitive capacity to utilize such features, demanding a different way to search and browse information. A study conducted on elementary-school-aged children by researchers at the University of Maryland in 2005 indicates that a flat, non-hierarchical interface was easier and faster to use in most cases, especially for younger children [29]. However, this also suggests a more cluttered interface, as the application must display more information at the time. This conflicts with the findings of Faghih et al., which state that only relevant information should be displayed [28].

Although not directly conflicting, the findings in these articles discuss aspects of user interface design that must be balanced against each other when designing for children. Considering their assumed lack of reading abilities, children can benefit from using design elements that may negatively affect an older user group. The findings by Hutchinson et al. also show a significant gap in understanding between first -and fifth-graders regarding how to browse and search for information using a graphical user interface [29]. To summarize, there seems to be no correct conclusion to the design problem but rather a set of design decisions that must be considered and weighed against each other in the context of the target user group.

2.6 Login-Based Applications for Children

The application will need an authentication system to ensure that the results recorded during the assessment belong to the correct pupils. Usually, a typical username and password-based login system would be sufficient, but this may not be the optimal solution for young pupils. In their paper *Designing Textual Password Systems for Children*, J. Read and B. Cassidy examine password usage and habits of young children and propose design requirements for password systems aimed at this target group. Two of their main proposals relevant to this application are 1) keeping passwords short and 2) keeping them simple by avoiding the requirement of using both letters and numbers [30]. These proposals conflict with the usual security-oriented password requirements usually found in applications but can, in turn, significantly increase usability. When designing the authentication system for the application, it is important to consider this trade-off between security and usability.

Chapter 3

Research Approach

This chapter describes the research approach applied to this project, Design Science Research. It presents the chosen research method used to answer the research questions. It outlines the six integral steps in the process and the seven guidelines applied to follow the process model.

3.1 Research Method

The goal of the project is to design and implement a system that can assist young pupils in the process of learning how to read words and to make it easier for teachers to gain a better understanding of the reading level of their pupils and the class as a whole. A suitable research approach for the project is selected to achieve this, namely, design science research (DSR). DSR is a problem-solving approach that aims to create an artifact that helps humans solve problems to a greater extent than before by providing intellectual and computational tools [31]. Peffers et al. present a process model, which is a synthesis of prior research done in this field. The process consists of six activities in a nominal sequence. It is an iterative process, after completing the evaluation step, researchers can decide if they want to start a new iteration from step 3 to try to improve the artifact [32]. The steps are illustrated in Figure 3.1 and explained below.

1. **Problem Identification and Motivation:** Define the specific research problem and justify the value of a solution. It is necessary to explain why the solution is important and the benefits it could provide.
2. **Objectives of a Solution:** Infer the objectives of a solution from the problem definition. This step aims to set clear and achievable goals for what the solution should accomplish.
3. **Design and Development:** Create the artifactual solution. Design involves specifying the artifact's features and function, while development involves creating an actual, working artifact.
4. **Demonstration:** Demonstrate the efficacy of the artifact to solve the problem. Important to show that the artifact solves one or more instances of the problem.
5. **Evaluation:** Observe and measure how well the artifact supports a solution to the problem.
6. **Communication:** Communicate the problem and its importance, the artifact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences, such as practicing professionals, when appropriate.

The process is structured in nominally sequential order, but it is possible for researchers to start at four different entry points and move outwards. The entry point is chosen based on the type of research. The natural entry point for this project would be the Design & Development-centered

approach. The two initial steps of problem identification and motivation and objectives of a solution were completed in the specialization project found in Appendix Section H. The focus of this project is creating and evaluating the artifacts. The four different possible approaches are presented below [32].

1. **Problem-centered approach:** The basis of the nominal sequence, starting from the first activity. Researchers typically begin at this point if the research idea results from an observation of the problem or from suggested future research in a paper from a prior project.
2. **Object-centered solution:** Researchers start at activity two when the research is initiated by a need within an industry or research domain that can be met through the development of an artifact.
3. **Design & Development centered approach**: The Design and Development centered approach is employed by researchers when there is an existing artifact that has not been specifically designed to solve a problem in a domain in which it will be used. These types of artifacts might come from other domains, have been used to solve a different problem, or might have appeared as an analogical idea. The Design & Development-centered approach to start with activity 3.
4. **Observing a solution:** Starting with activity 4, this approach involves retroactively applying rigor to a practical solution that has already been implemented. Researchers work backward to create a design science solution based on the observed success of the practical solution.

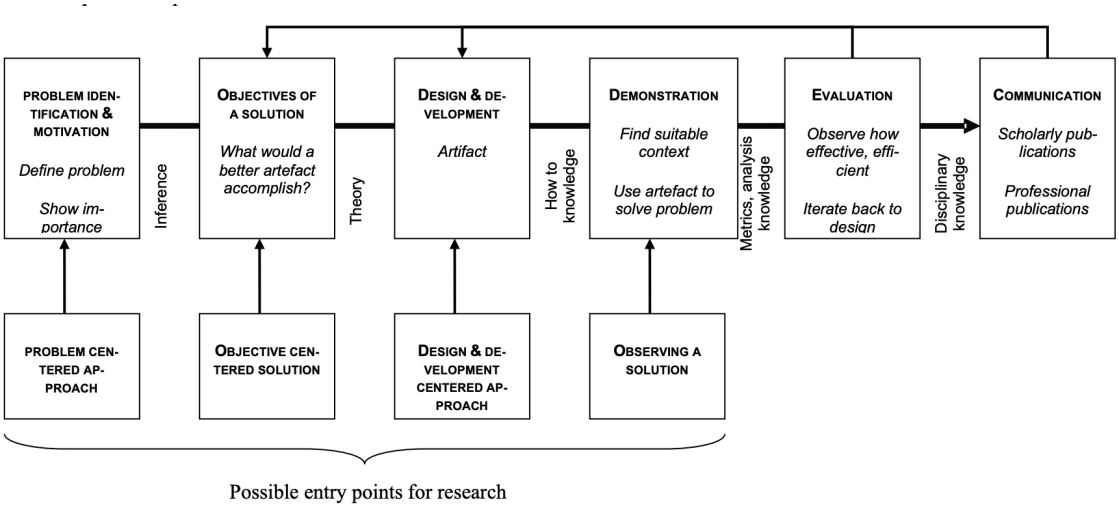


Figure 3.1: Design science research process model

To follow the process model, Hevner et al. established seven guidelines, described in Table 3.1, to assist researchers, reviewers, editors, and readers in understanding the requirements for effective design-science research. They advise that researchers use their judgment to determine when, where, and how to apply the guidelines in each project [31]. A description of how the guidelines were applied in this project is listed below.

1. **Guideline 1:** The artifact created in this project is the digitalized word chain test and the teacher dashboard.
2. **Guideline 2:** The solution will assist teachers in evaluating their pupil's reading literacy and give them more insight into the reading level of the class.

3. **Guideline 3:** The artifact will be tested in multiple iterations. One of the tests is a comparability test, comparing the digital test to the analog test. Usability testing will also be conducted. The results of the comparability tests and usability tests will be presented in Chapter 5. The tests will be used as a basis for the evaluation of the artifacts in Chapter 6.
4. **Guideline 4:** The contribution of this research is an artifact for digital assessment of pupils' decoding skills, an artifact assisting teachers in the work of teaching pupils how to read, and the evaluation results from comparability testing and usability testing.
5. **Guideline 5:** The project utilizes a variety of tools, methods, and tests in both the construction and evaluation of the design artifact presented in Chapter 4.
6. **Guideline 6:** The artifacts developed in this project is iteration based. Each iteration is evaluated and taken into consideration in the next iteration to improve the quality of the artifacts.
7. **Guideline 7:** This report aims to present the project in an efficient manner, both to technology- and management-oriented audiences.

Guideline	**Description**
Guideline 1: Design as an Artifact	Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation
Guideline 2: Problem Relevance	The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
Guideline 3: Design Evaluation	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
Guideline 4: Research Contributions	Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
Guideline 5: Research Rigor	Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
Guideline 6: Design as a search process	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
Guideline 7: Communication of Research	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

Table 3.1: Design-science guidelines

Chapter 4

Methods, tools, and technology

The following chapter describes the development method used in this project and a description of the tests used to evaluate the artifact. The final section describes all the tools and technologies used in the development process.

4.1 Developement method

Being a small team of two programmers working on three distinct code repositories, we opted for a hybrid between trunk-based development and merge-request based development. Trunk-based development is founded on the concept of small teams collaborating on a single development branch. In contrast, merge-request-based development emphasizes the creation of separate branches for each developer, which are then merged into the main branch via code reviews and merge-requests [33]. Our development method adopted the habit of creating separate branches for each developer from merge-request-based development and embraced the philosophy of trunk-based development by keeping the separate branches feature-based and as short-lived as possible. Merge requests and code reviews were also conducted to ensure that both developers had sufficient insight into all parts of the system. The Github repositories and their links are listed in Appendix B.

The development was conducted over two iterations. The first iteration focused on implementing the application as it was sketched and prototyped in the specialization project. As the specialization project mainly contained functional and non-functional requirements, in addition to design sketches, the first iteration also included more extensive planning such as data modeling, technology stack research, architectural planning, and setting up development environments. After the first iteration, we conducted a moderated comparability test to see how comparable our digital test was to the paper-based test. Results and feedback during this test laid the foundation for further development of the tester's frontend in the second iteration, where we focused on fixing the issues we discovered. For the teacher's dashboard and backend, the second iteration focused mainly on making the project deployable to external servers.

4.2 Usability Testing

To evaluate the system and the value it can bring to Norwegian teachers and pupils, a multi-part usability test was conducted at the end of the second iteration, with four questionnaires bundled in a single survey. The goal of the survey was, first and foremost, to test the functional requirements defined for the application. Secondly, we wanted to assess the application using the System Usability Scale (SUS) to get a more general evaluation of the system, something that will be discussed in depth later in this section. The survey also collected additional data about the participants and their backgrounds to explore whether these metrics affected the user experience or not.

Usability testing can be done in many forms. The metrics collected can be qualitative or quantitative, the test environment can be remote or in a test lab, and the tests can be moderated or unmoderated [34]. In order to reach as many participants as possible, the project opted for **unmoderated remote usability testing** [35]. This was conducted in a multi-part survey containing the following parts/tests:

1. Usability test
2. A System Usability Scale (SUS) questionnaire
3. A sampling of the participant's age and experience as a teacher
4. Questions regarding the participant's final thoughts and perceived value of the application

Before conducting usability tests with actual participants, it is beneficial to perform a preliminary pilot test in the context of unmoderated tests. This pilot test serves as a valuable step to identify and address any potential issues that may arise after the test is distributed to participants [35]. In our case, prior to involving real users, we conducted a pilot test using fellow students. This pilot test helped us uncover problems with question-wording that led to misinterpretations, as well as identify minor application bugs.

The survey, along with the participants' answers, can be found in Appendix G. Since the target group for the application is Norwegian teachers, all questions were in Norwegian.

4.2.1 Usability Test

The first and main part of the survey was a usability test that focused on key functional requirements and user tasks (scenarios) within the application. The test comprised multiple tasks the user had to complete, along with an evaluation of each task. The evaluation collected quantitative data in the form of a rating on a scale of 1-to-5 of how easy the task was to complete. Remote unmoderated usability tests in the form of a survey lack the ability to monitor how the user interacts with the application. To compensate for this, the survey also collected qualitative data in an optional field for additional feedback.

4.2.2 System Usability Scale

Product-specific usability tests help designers evaluate their applications based on functional requirements and pre-defined tasks. Still, they often require extensive testing and can not evaluate the system's perceived usefulness on a standardized scale. One way to quickly collect a user's subjective rating of a product's usability is by using System Usability Scale (SUS) created by John Brooke in 1996, which has been proven as an important tool for usability testers since its release [36].

A SUS-evaluation takes the form of a ten-item questionnaire, where for each question, the participant answers on a scale from 1, "Strongly disagree" to 5, "Strongly agree". To prevent response biases where the participant does not fully think about their answer, the questions alternate between positive wording for even-numbered questions and negative wording for odd-numbered questions. The output of the scale is a SUS-score, which is a composite measure of the overall usability of the system being studied [37]. The SUS-score can be computed using the following procedure:

1. For questions 1,3,5,7, and 9, subtract 1 from all the answers before summarizing them
2. For questions 2,4,6,8, and 10, subtract the score from 5 for all answers before summarizing them
3. Multiply the sum of scores by 2.5 to produce the final SUS-score

The SUS-score will be a number between 0 and 100 but must not be considered a percentage of attainment. Using data from 446 studies and over 5000 individual SUS-responses, Jeff Sauro and James R. Lewis studied the SUS-scores and how they could be normalized. The study found that the average SUS-score was 68 and that a score of 50 would fall within the bottom 14% [38]. To better evaluate a system using the SUS-score, Sauro and Lewis used the data to create a curved grading scale for SUS-scores in 2011, named the Sauro-Lewis CGS [39]. This scale will be used to evaluate the system from the SUS-scores, and is shown in Table 4.1

SUS-score	**Grade**	**Percentile range**
84.1-100	A+	96-100
80.8-84	A	90-95
78.9-80.7	A-	85-89
77.2-78.8	B+	80-84
74.1-77.1	B	70-79
72.6-74	B-	65-69
71.1-72.5	C+	60-64
65-71	C	41-59
62.7-64.9	C-	35-40
51.7-62.6	D	15-34
0-51.7	F	0-14

Table 4.1: Sauro-Lewis Curved Grading Scale

4.3 Comparability Testing

To test the performance and ensure that the digital and analog tests had the same testing basis, comparative tests were conducted. Verifying that users achieve similar results on both types of tests provides reason to believe that the digital test is able to assess the decoding abilities of the user in the same manner as the analog test.

The first comparative test was conducted after the first iteration of implementation. This test involved testing a group of users on both the analog and digital tests. The participants in the test were recruited by convenience sampling, a type of non-random sampling where members of the target population meet certain practical criteria, such as easy accessibility, availability, or willingness to participate [40]. Convenience sampling is easy and inexpensive and, with limited resources and time, a fitting sampling method for this project. The sampling is not likely to be representative of the whole population, but it can provide a sense of difference in performance on the two types of tests by comparing scores.

Each participant completed both types of tests. The participants had a two days gap between each of their tests. This was done to prevent them from performing better on the second test because they had the word chains fresh in memory. The test group was divided into two subgroups where one subgroup completed the paper test first and then the computer-based test, while the other subgroup did the opposite by taking the computer-based test first. This helps to level out any possible advantages of having one test before the other when comparing the results of the tests. To level out any gender differences, both groups had an even distribution of female and male participants.

The test was a moderated test. Each participant was placed in an empty room with only a

moderator present. They were provided with instructions on how to carry out the test. Prior to the paper-based test, the instructions were carried out accordingly to the manual (Section 2.3.1. Equivalent instructions were provided before the digital test, including more technical instructions, e.g., how to write or delete a divider ("|"). The instructions for the analog test are described in Section 2.3.1, and the instructions for the digital test are described in Section E. Every tester carried out the digital test on the same machine. During the digital test, the moderator observed the struggles and difficulties the users experienced with the test. After completing the test, the user was able to submit feedback to the moderator.

The second comparative test was conducted two weeks subsequent to the first test, following the completion of the second implementation cycle. The participants were selected at random from the group that had previously undergone the first comparability test. This test consisted solely of the digital test. The participants had not done any activities aimed to improve their reading in the two weeks, and the results from the first analog test could therefore be used as a valid basis for comparison.

Furthermore, the second test was administered in the absence of a moderator.

The means and standard deviation were compared to highlight any differences between the results of the tests. In addition to the test of means, dependent samples t-test and Spearman's Rank-order correlation were utilized to analyze the results. All of the analyses were performed using the statistical analysis software SPSS version 29.0.0.0 on macOS.

The stanine scale was used to compare the distribution of scoring on the two tests.

Dependent samples t-test

The dependent samples t-test compares the means of two related groups on the same continuous, dependent variable to determine whether there is a statistically significant difference between these means. This requires that the same participants are tested more than once and represented in both groups. When choosing to analyze the data with a dependent t-test, the data needs to be checked to make sure it can be used in a dependent t-test. There are some assumptions the data need to pass to give a valid result [41].

1. The dependent variable should be measured on a continuous scale.
2. The independent variable should consist of two categorical related groups. A related group indicates that the same subjects are present in both groups.
3. There should be no significant outliers in the difference between the two related groups.
4. The distribution of the differences in the dependent variable between the two related groups should be approximately normally distributed. The reason the test only requires approximately normal data is that it is quite robust to violation of normal distribution.

The last two assumptions can be verified by utilizing SPSS. To identify potential outliers, a boxplot can be used to graphically illustrate the numeric data. Mild outliers, which are values 1.5 x the interquartile range below the first quartile or above the third quartile, are represented by circles. Extreme outliers, which are values 3 x the interquartile range below the first quartile or above the third quartile, are represented by asterisks (Figure 4.1).

Shapiro-Wilks Test of Normality can be used to test if the two related groups are approximately normally distributed. The test is a hypothesis test, where the null hypothesis is that the sample comes from a normal distribution. With a chosen degree of confidence of 95% and if the significance value is below 0.05, the data significantly deviate from a normal distribution. The Shapiro-Wilk test should not be taken as 100% reliable and should be interpreted along with the result of graphical and numerical tools. The formula used to calculate the Shapiro-Wilk statistic is [42]:

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.1}$$

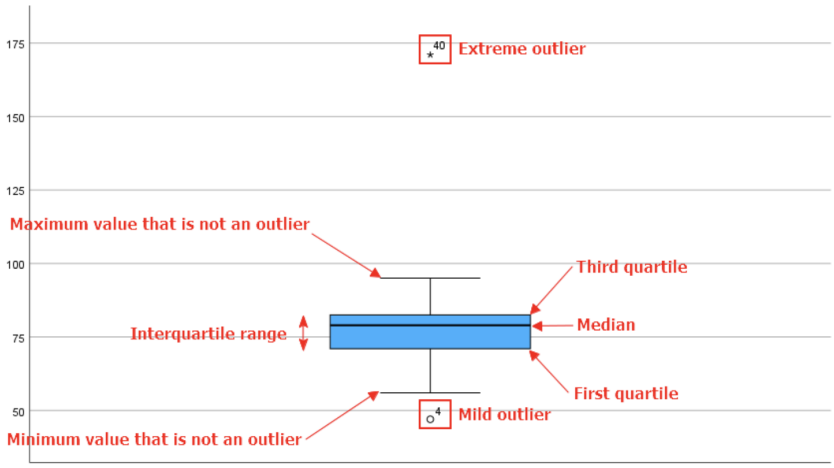


Figure 4.1: SPSS-Boxplot Example

The variables in the equation are described as [42]:

- W is the Shapiro-Wilk statistic.
- a_i are the constants generated from the covariance matrix of the order statistics of a sample of size n from a standard normal distribution.
- y_i are the ordered sample values.
- $\bar{y}$ is the mean of the sample values.
- The index i ranges from 1 to n, where n is the sample size.

The result of the dependent t-test contains a lot of different information. The first five columns represent the difference between the two related groups. The last three columns express the results of the dependent t-test. The different columns are [43]:

- **Mean:** The average difference between the pair of scores. The scores are usually from the same individuals under two different conditions or times.
- **Std. Deviation:** The amount of variation from the mean. A low standard deviation indicates that the data points are close to the mean difference, high standard deviation indicates that the data points are spread out.
- **Std. Error Mean:** Measure of how far the sample mean of the data is likely to be from the true population mean.
- **Lower and Upper:** Represents the lower and upper bound of the 95% confidence interval for the mean difference. The 95% confidence interval means that there is a 95% certainty that the true mean is in the range between the lower and upper value.
- **t:** A ratio of the difference between the mean of the two related groups and the variation that exists within. This is an indication of whether there is a significant difference between the means of the two groups. Whether the t-value is considered significant is found in the t-distribution table Figure 4.2.
- **df:** Represents the degree of freedom in the t-test. Df is N - 1 in a dependent samples t-test.
- **One-sided p and Two-sided p:** The p-values for one-tailed and two-tailed tests. The value is the probability of observing equally or more extreme results than the observed data, assuming the null hypothesis is true. A small p-value (usually less than 0.05) usually rejects

the null hypothesis. In a dependent samples t-test, the null hypothesis is usually that there is no difference between the means of the two sets of scores. The one-sided p-value test that the mean difference is either greater than or less than 0, while the two-sided p-value tests that the mean difference is not 0 without specifying a direction.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Figure 4.2: T-distribution table

Spearman's correlation

Spearman's correlation is a statistical measure used to assess the strength and direction of the relationship between two variables. Spearman's correlation measures the variable's monotonic relationship, which is any relationship that increases or decreases systematically but not necessarily in a straight line. The correlation is calculated by ranking every score from 1 to N based on how the score rank compared to the others. The score with the highest value is labeled "1" and the lowest is labeled "N". [44] Based on the two tests, each participant in the test is assigned two ranks ranging from 1 to N based on their result, e.g., the participant with ID 2 scored highest on the Paper-based test and on the PC-test and will get the ranks 1 and 1 Table 2.

The standard formula for calculating Spearman's Correlation coefficient is calculated by [44]:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4.2}$$

where d = the pairwise distance of the ranks of the variables x_i and y_i and n = the number of samples.

SPSS calculates the Spearman correlation by a more advanced formula, which considers tied ranks Equation 4.4. When no tied ranks occur, the formula is identical to the standard formula Equation 4.3.

$$\rho = \frac{T_x + T_y - \sum_{i=1}^{N} d_i^2}{2\sqrt{T_x T_y}} \tag{4.3}$$

and T_x and T_y are:

$$T_x = \frac{N^3 - N - S_{T_x}}{12} \tag{4.4}$$

$$T_y = \frac{N^3 - N - S_{T_y}}{12} \tag{4.5}$$

The average rank is assigned if t_i observations are tied in a group of ties. Each time t_i ¿ 1 occurs, the quantity $(t_i^3 t_i)$ is calculated and summed up for X and Y separately. The summations are designated by ST_x and ST_y, respectively. N is the total number of samples, and S. [45]

Spearman's correlation coefficient, often denoted as "ρ" ranges from -1 to +1, where a value of -1 indicates a perfect negative monotonic relationship between the two variables, a value of +1 indicates a perfect positive monotonic relationship and a value of 0 indicates no monotonic relationship between the variables [44]. Table 4.2 shows how to interpret the correlation coefficient [46].

Size of correlation	**Interpretation**
.90 to 1.00 (-.90 to -1.00)	Very strong positive (negative) correlation.
.70 to .89 (-.70 to -.90)	Strong positive (negative) correlation.
.40 to .69 (-.50 to -.70)	Moderate positive (negative) correlation.
.10 to .39 (-.30 to -.50)	Weak positive (negative) correlation.
.00 to .10 (.00 to -.30)	Negligible correlation.

Table 4.2: Interpretation of Correlation Coefficient

SPSS generates a table following Spearman's correlation procedure which includes the correlation coefficient. In addition to the coefficient, a p-value is calculated, which is the probability of wrongfully rejecting the null hypothesis of no correlation, shown in the Sig. (2-tailed) row. In this project, p-values <0.05 is regarded as statistically significant.

Spearman's correlation requires the data to pass some assumptions. As in dependent samples t-test, to provide a valid result, the data need to meet some assumptions [44]. The three required assumptions are listed below:

1. The two variables should be measured on an ordinal, interval, or ratio scale.
2. The two variables represent paired observations.
3. There is a monotonic relationship between the two variables.

Spearman's correlation was deemed more suitable than, e.g., Pearson's correlation for the statistical analysis owing to its non-parametric nature. Since the data might not follow normal distribution or linearity, it would not conform to the assumptions required for Pearson's correlation. Furthermore, Spearman's correlation is less vulnerable to the impact of outliers that might be present in our data [47].

4.4 Stanine scale

Stanine, Standard nine, is a method of scaling test scores into 9 different categories. A distribution that is close to being normally distributed can be converted to a stanine value. Using stanine, a percentage of the scores are classified into each stanine group [6]. Using stanine makes it easier to interpret the result as it reduces the work of trying to interpret small score differences [48]. The scale has a mean of five and a standard deviation of two. Table 4.3 shows the percentage of scores in each group and the cumulative frequency on each level. In combination with the word chain test, it is recommended to further test the pupils scoring in the lowest two levels on the stanine scale [6].

Stanine	1	2	3	4	5	6	7	8	9
Distribution by percentage	4%	7%	12%	17%	20%	17%	12%	7%	4%
Cumulative distribution	4%	11%	23%	40%	60%	77%	89%	96%	100%

Table 4.3: Stanine scale

4.5 Tools and technologies

This section lists and describes the tools and technologies used to develop the application in this project.

Git

Git is an open-source distributed version control system to track changes in source code during software development. It allows multiple developers to collaborate on the same codebase, record changes to files, and manage the development of software projects. In this project it was used to track changes and allow simultaneous development, dividing the backend and two frontend applications in separate repositories [49].

GitHub

GitHub is a web-based platform for version control and collaborative software development using Git. It allows developers to host, review and share their source-code repositories and offers additional tools such as issue-tracking and deployment pipelines. In this project, GitHub was used to host the aforementioned Git repositories and perform code reviews [50].

Docker

Docker is a platform used to create, deploy, and run applications in containers, which are minimalistic virtual computers. It allows developers to package their applications along with their dependencies into a single unit that can easily be deployed to different computing environments. In this project, Docker was used to package the API and database in separate containers [51].

Ktor

Ktor is a lightweight, asynchronous web framework for building server-side applications in Kotlin. It provides an API for handling HTTP requests and responses and is also compatible with websockets and HTTP/2. In this project, Ktor was used to create the API, letting both the dashboard and the tester's frontend communicate with the database [52].

Firebase

Firebase is a cloud-based Google-owned mobile and web application development platform that provides numerous services to facilitate application development. These services include real-time database, user authentication, file storage, application hosting, and analytics, among others. In this application, we have used Firebase for the authentication of users through a SDK [1] in the

[1] SDK: Software Development Kit

Ktor API [53].

PostgreSQL

PostgreSQL is an open-source relational database management system that supports the query language SQL. It is known as a scalable and reliable DBMS with ACID compliance, concurrency, and transaction support. In this project, PostgreSQL has been used as the main component in the data layer of the application [54].

React

React is a declarative and component-based JavaScript library for building user interfaces and is developed and maintained by Facebook. Being the most popular frontend framework as of March 2023, it has a large community and is widely supported by smaller plugins and components. In this project, React has been used as the frontend framework for both the tester's frontend and the teacher's dashboard frontend [55].

Vite

Vite is an open-source build tool for web applications aiming to provide a fast and efficient development experience. Vite supports features such as hot module replacement, a speedy development server, and highly optimized production builds. In this project, Vite has been used as a development web server and for building and packaging the frontends for deployment [56].

Netlify

Netlify is a cloud-based web development platform that provides website hosting, continuous deployment, and serverless functions. It simplifies the web development workflow by automating common tasks (like building and deploying a website) and providing an easy-to-use interface for developers. This project uses Netlify to host the two frontends, and deploys new versions when the main branch of a repository is updated [57].

TypeScript

TypeScript is a strongly typed superset of the frontend programming language JavaScript, which TypeScript compiles to. Being strongly typed, it allows developers to discover errors more quickly and debug more efficiently. In this project, TypeScript has been used along with React to program the frontends [58].

Kotlin

Kotlin is developed by JetBrains and can be described as a modern remake of Java, that also runs on the Java Virtual Machine. It is designed to be more concise, expressive, and safe than Java, while still being able to import packages written for its predecessor and run in the same runtime environments. In this project, Kotlin has been used to code the backend using Ktor [59].

Tailwind CSS

Tailwind is a utility-first CSS framework that provides pre-designed CSS classes. It is used to build responsive and customizable user interfaces swiftly. By condensing common CSS rules into parameterized class names, it allows developers to describe complex components in a single line of code. In this project, Tailwind CSS was used in both frontend applications [60].

React Query

React Query is a library for managing server state in React applications. It provides a simple and powerful way to handle data fetching, caching, and updating, making it easier to connect React frontends to a REST API [61].

Microsoft Visual Studio Code

Microsoft Visual Studio Code is a free, open-source, highly customizable, and lightweight code editor that supports multiple programming languages. It provides typical code editor functionalities like debugging, refactoring, and syntax highlighting, while also providing a vast marketplace for

user-made extensions. According to StackOverflow Developer Survey for 2022, it is the most used code editor to date. In this project, Visual Studio Code was used for frontend development [62].

JetBrains IntelliJ IDEA

IntelliJ IDEA is an integrated development environment created by JetBrains, mainly aimed at Java and Kotlin development. Being designed for Java development, it features typical IDE functionalities like debugging, refactoring, and syntax highlighting, but also more Java-oriented features such as out-of-the-box support for build tools like Gradle and Maven, and custom build settings for specific projects. In this project, IntelliJ IDEA has been used for backend development in Kotlin [63].

Chapter 5

Implementation

The implementation chapter provides an overview of the project's architecture, outlining the functional and non-functional requirements defined for this project. The two iterations of implementation are also described including the feedback received during testing.

5.1 Architecture

The current project employs the widely used 3-tier architectural pattern, which is typical for modern web applications [64]. The presentation tier comprises two distinct user interfaces: an administrative dashboard designed for teachers and a test application intended for pupils. Both frontends are developed using React JS and compiled into static HTML web pages that can be effortlessly hosted on any standard web server. The two frontends are connected to the same logical layer, which is an API built with Kotlin-Ktor. This, in turn, connects to the data tier, which is built upon a PostgreSQL database.

It is noteworthy that the project deviates from the conventional 3-tier architectural pattern by utilizing Google Firebase as an Authorization server instead of implementing a login system in the API server. When users log in or sign up, a direct request is made to Firebase via the client, and upon successful completion, a User-object that contains metadata about the logged-in user is returned. The signup process is illustrated in Figure 5.1. If a user signs up for the first time (in the dashboard-frontend), additional user information obtained from the signup form is written to the database. The User-object, which is obtained from the Authorization server/Firebase, includes an Access Token. When making requests to the Ktor API, this Access Token is supplied in the HTTP request. The API utilizes the Firebase SDK to validate the token before fulfilling the request, obtaining a User-object that can be used for further authentication and authorization of the user. This process is illustrated by the sequence diagram in Figure 5.2. By considering this, one could argue that the project also adopts the traditional OAuth Authorization pattern, where Google Firebase acts as the Authorization server, and the Ktor API acts as the resource server.

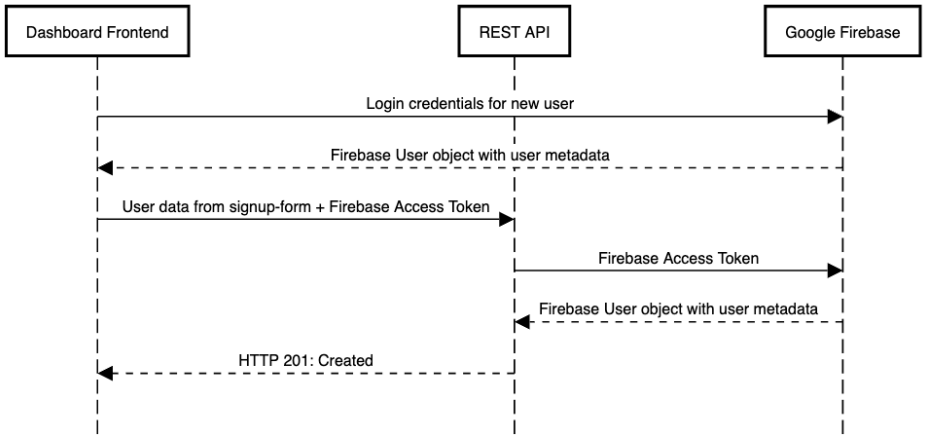


Figure 5.1: Sequence diagram of the sign-up process

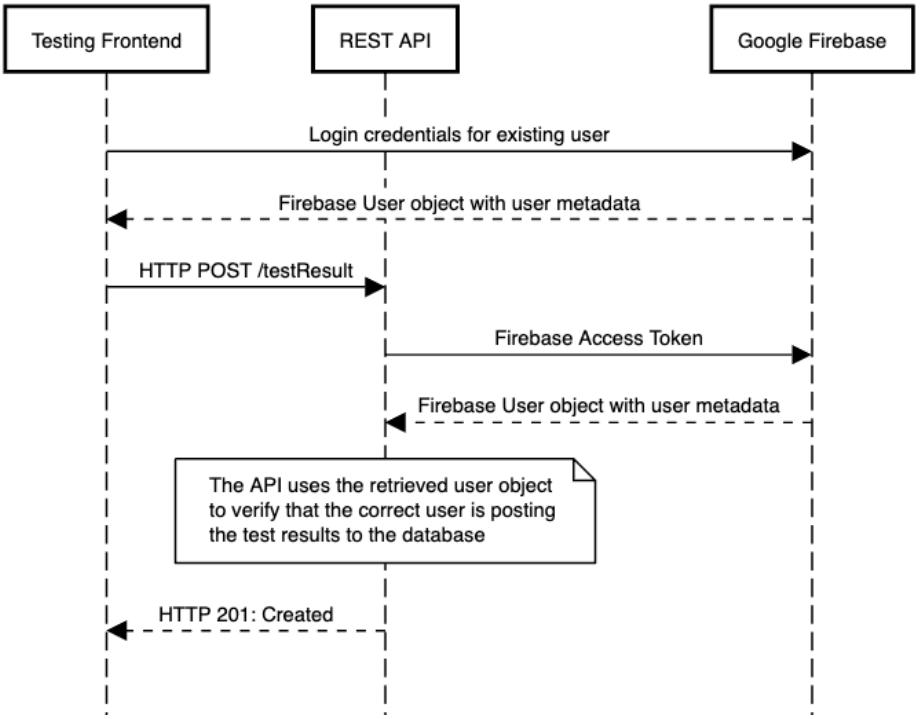


Figure 5.2: Sequence diagram showing the process of posting a test result from the Tester's frontend

5.2 Requirements

The following section presents the functional and non-functional requirements defined for the project. These requirements were initially defined in the specialization project (Appendix H) and are based on the paper-based word chain test. Each requirement is assigned a unique ID and includes a description that outlines its intended objective. Additionally, a priority level ranging from low to high is assigned to each requirement.

Id	Description	Priority
FR1	A pupil should be able to perform a word chain test	High
FR2	A teacher should be able to monitor the progress/level of each pupil	High
FR3	A teacher should be able to monitor the progress/level of each class	High
FR4	A user should be able to log in as their role (teacher or pupil)	High
FR5	A teacher should be able to create test-sessions	High
FR6	A teacher should be able to create user credentials for the pupils	High
FR7	The teacher should be able to print out a list of pupils and passwords	High
FR8	A teacher should be able to create a class or pupil	High
FR9	The system should be able to auto-generate passwords when a pupil profile is created	Medium
FR10	The system should be able to output anonymous test reports as data files	Medium
FR11	The teacher should be able to print out a detailed list of class results	Medium
FR12	A pupil should be able to train on test-tasks to familiarize themselves with the application	Medium
FR13	A system administrator should be able to create/delete user credentials for the teachers	Medium
FR14	The pupils should be able to access a picture-based walk-through	Medium
NFR1	The application should support modern web browsers and devices with keyboard and mouse for efficient assessment	High
NFR2	The application should provide immediate feedback to users for any action they perform (mouse click, keyboard input, etc.)	High
NFR3	The application should start within 2 seconds of accessing the URL	High
NFR4	At least 80% of users should find the application easy to use on the System Usability Scale (SUS). Users should either agree or strongly agree with the statement "I thought the system was easy to use", and either disagree or strongly disagree with the statements "I found the system unnecessarily complex", "I think that I would need the support of a technical person to be able to use this system", and "I found the system very cumbersome to use"	High
NFR5	At least 80% of users should find the application easy to learn on the System Usability Scale (SUS). Users should either agree or strongly agree with the statement "I would imagine that most people would learn to use this system very quickly", and either disagree or strongly disagree with the statement "I needed to learn a lot of things before I could get going with this system."	High
NFR6	At least 80% of users should either agree or strongly agree with the statement "I think that I would like to use this system frequently." on the System Usability Scale (SUS).	High

Table 5.1: Requirements

5.3 Data Model

The project selected a relational data model and database, based on the defined requirements and leveraging the apparent hierarchical structure of the data objects. A simple ER diagram of the data model is provided in figure 5.3. The *id*-attribute for each teacher- and pupil-object is a reference to the *id*-field of the corresponding Firebase User Object.

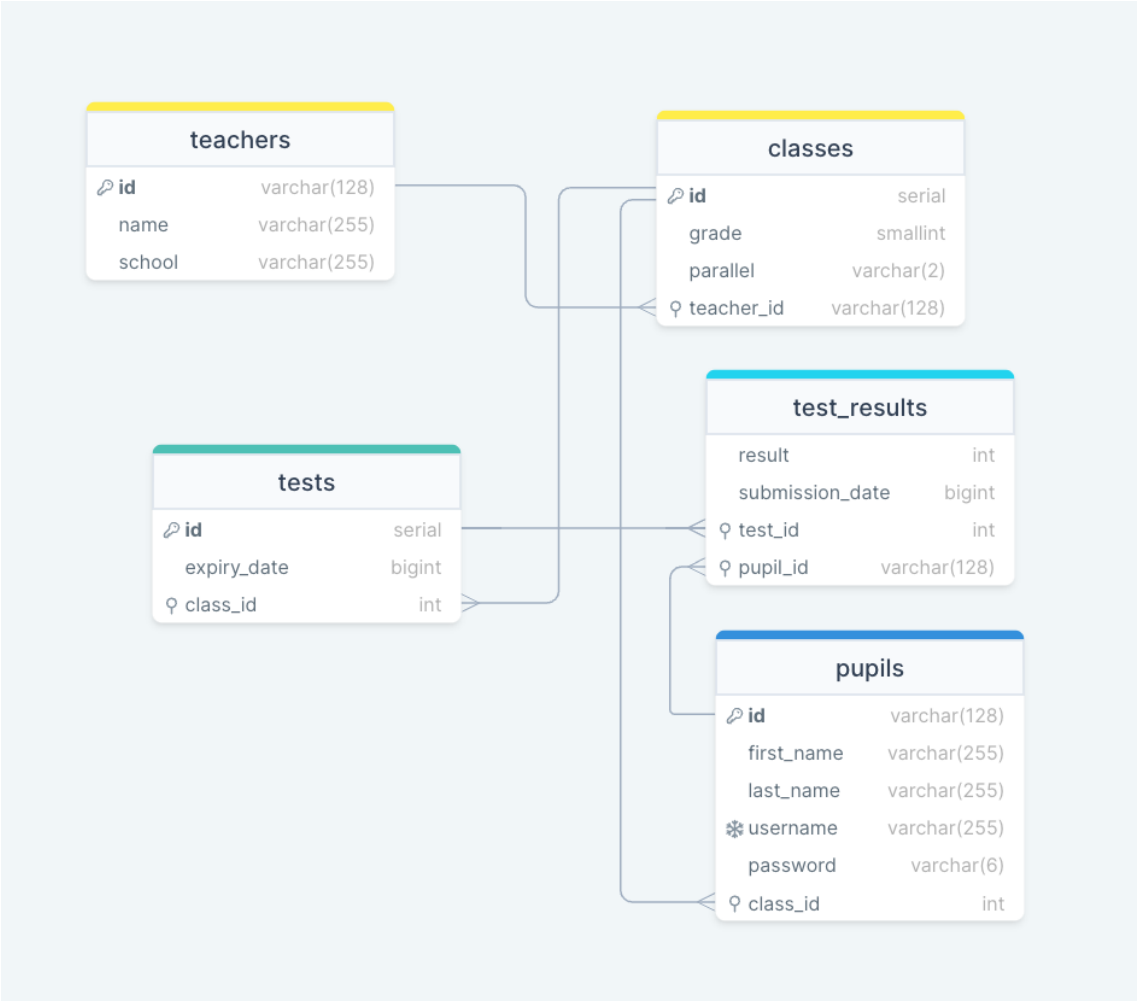


Figure 5.3: Entity Relationship diagram of the project's data model

5.4 First iteration

In the first iteration, the dashboard available for the teachers and the digital test for the pupils were implemented. The goal was to create a test that was comparable to the original test and a dashboard that provided the teachers with the desired information and results from the tests. After the first iteration was finished, a comparability test was conducted. This test resulted in both user feedback and data to compare the digital and analog tests.

5.4.1 Application design

This project consists of the implementation of two artifacts; one application designed for completing the word chain test and one dashboard application for the teachers. This section will describe the different pages and components of the two applications. The presentation of the design of the

applications is divided into two next subsections, one for each application, with each subsection presenting the pages in the order a user is likely to follow.

Dashboard design

Figure 5.4 shows the teacher's landing page, where they are given the option to log in if they have an account and sign up if they do not. By clicking the sign-up button, they are taken to the sign-up page (Figure 5.5a), where they can fill out a form with their name, school, email address, and password. On subsequent visits, a teacher can click the log-in button instead and be taken to login-page (Figure 5.5b), where they can access the application with the email address and password they provided in the sign-up form.

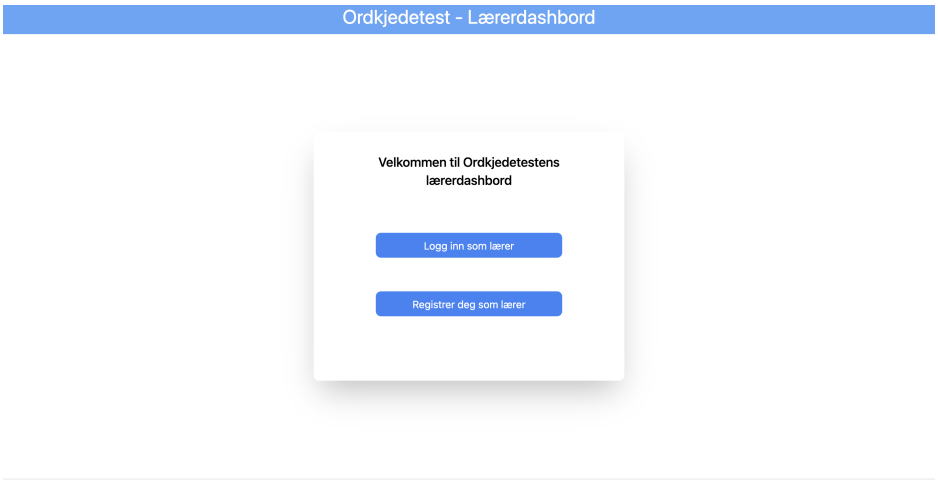
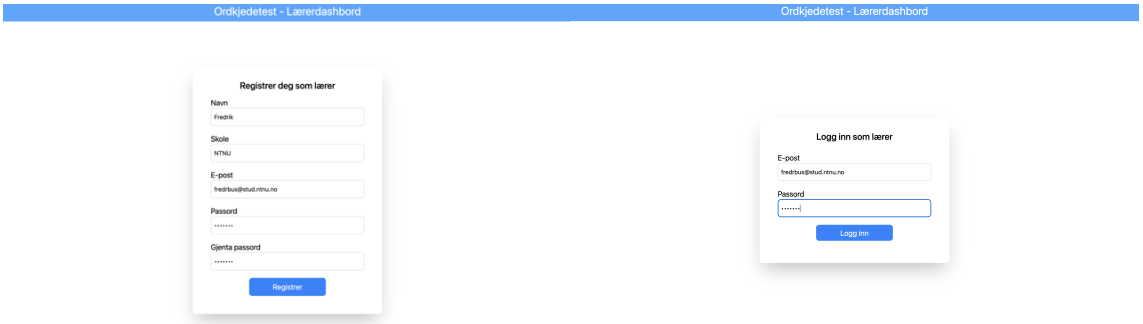


Figure 5.4: Dashboard landing page



(a) Sign-up page

(b) Log-in page

Figure 5.5: Sig-up and login-in page

After successfully logging in or signing up, the teacher is taken to the welcome page (Figure 5.6). On the left side of the page, there is a navigation pane that can be used to access the different pages of the application. This page will be visible on all the other pages as well, as long as the user is logged in. The name and school of the teacher are displayed at the bottom of the navigation pane, accompanied by a button the user can click to log out of the dashboard.

Figure 5.6: Dashboard welcome page

When clicking "Classes" on the navigation page, the browser navigates to the "Classes"-page (Figure 5.7). Initially, this will show an empty list as the user has not created any classes yet. To create a class, the teacher can click the "Create class"-button at the upper right side of the screen. This will take them to the "Create class"-page (Figure 5.8), where they can create a new class by assigning a grade (numbers 1 through 10) and a parallel (letters A through F). After successfully creating one or more classes, the teacher can navigate back to the "Classes"-page by either using the browser interface or the "Back"-button, where they can see the newly created classes (Figure 5.9).

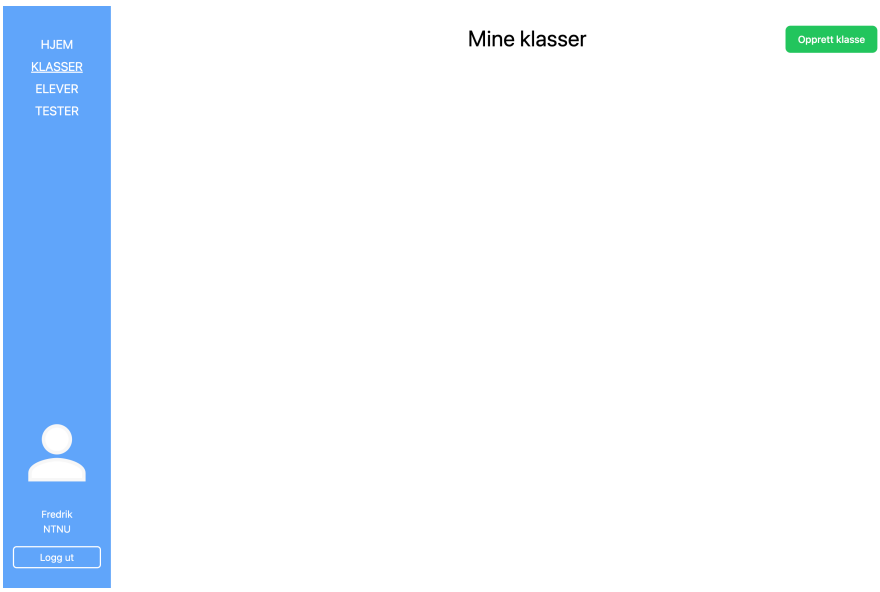


Figure 5.7: "Classes"-page, before any classes have been created

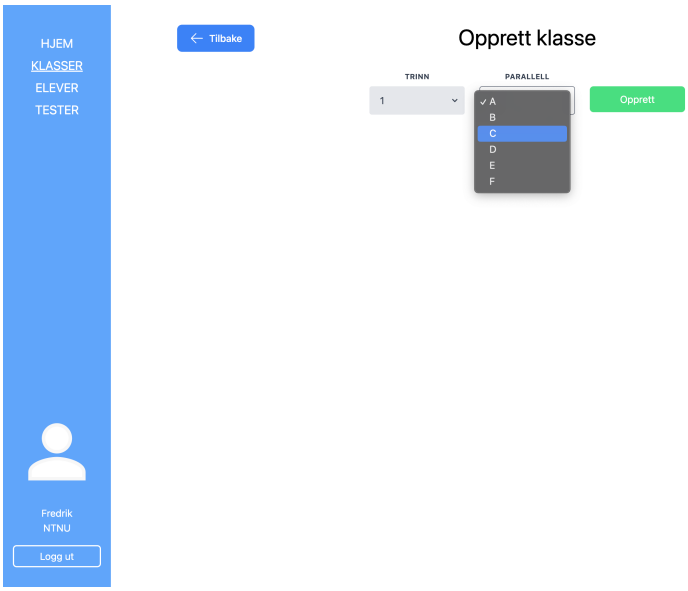


Figure 5.8: "Create class"-page

Figure 5.9: "Classes"-page, containing a table of the teacher's classes

When clicking on one of the new classes for the first time, the browser will navigate to the "Class"-page for the newly created and empty class (Figure 5.10). To populate the class with new pupils, the teacher can click the green "Create pupil"-button to the right of the class name. This will take her to the "Create pupil"-page (Figure 5.11). In this form, the teacher enters the first and last name of the pupil that should be created and appended to the class. A username in the format "firstname.lastname" will automatically be created. In the event that the same username is already in use, the application will notify the teacher so that they can alter the username and make it unique. After creating pupils for the class, the teacher can navigate back to the "Class"-page. The result section will still be empty as there have been no completed tests, but by clicking the "User info"-button, all the class pupils will be displayed by their first name, last name, username, and password in plaintext (Figure 5.12).

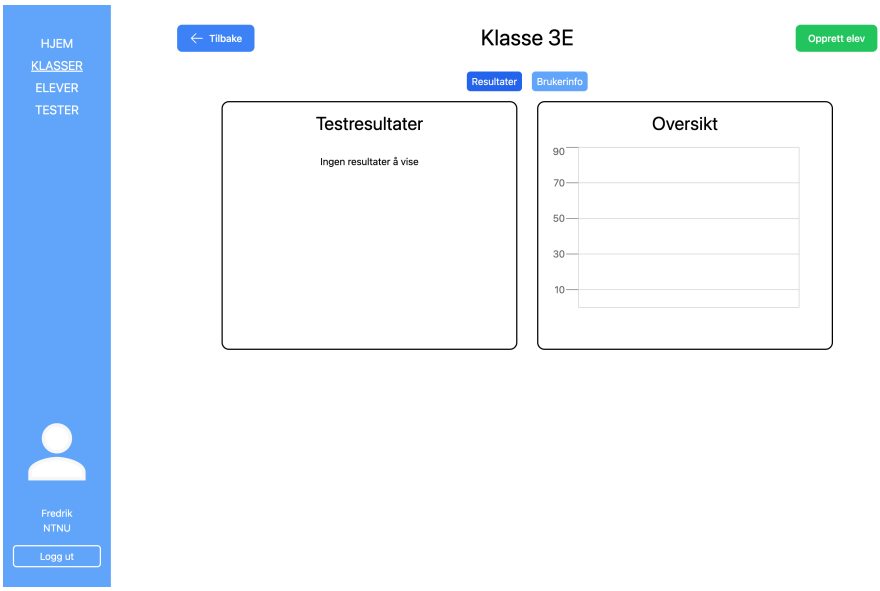


Figure 5.10: "Class"-page, class without pupils

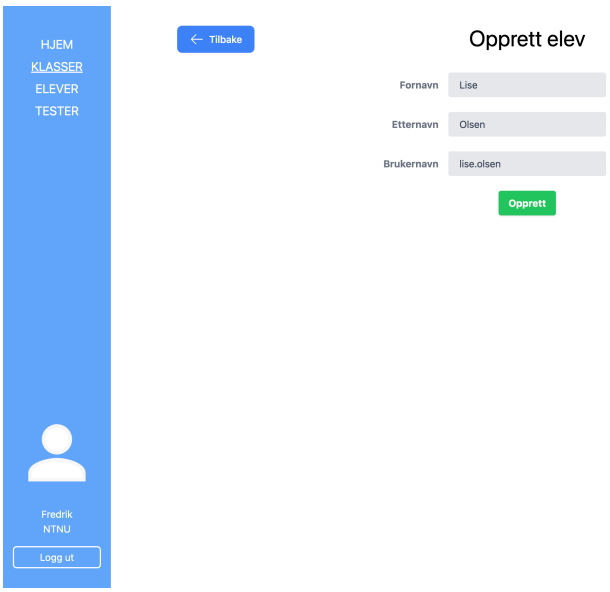


Figure 5.11: "Create pupil"-page

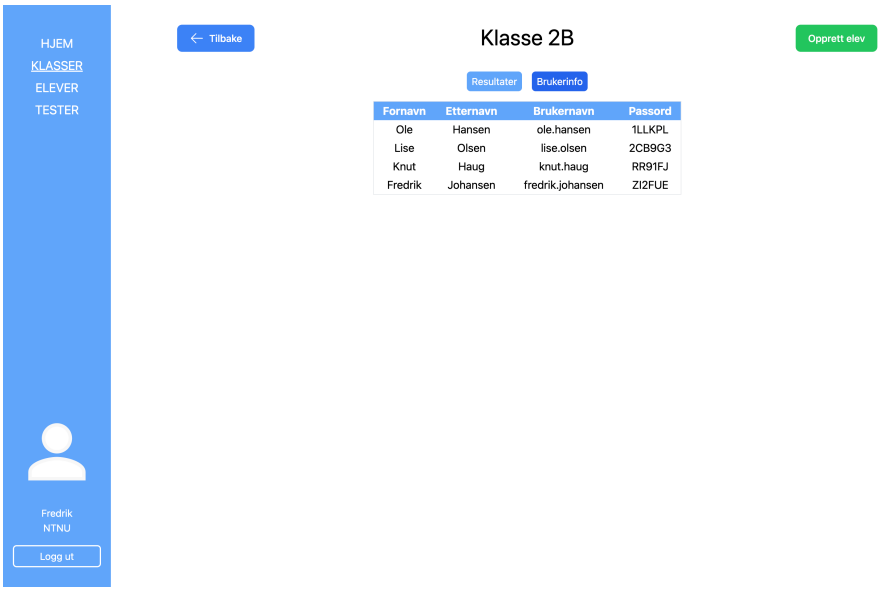


Figure 5.12: User info for pupils under "Class"-page

The teacher can access the "Tests"-page (Figure 5.13) by clicking on "Tests" in the navigation pane. On this page, all tests are shown in the table, with each row displaying the class name, expiry date, and number of pupils who have taken the tests (out of a number of eligible pupils). To create a new test, the teacher can click the green "Create test"-button to the right of the page title. This will take her to the "Create test"-page (Figure 5.14), where the teacher can select one of their classes from a dropdown menu and set an expiry date from a date picker object. Once the test is created, the pupil can log into the test application with their username and password and complete the test. After the tests are completed, the teacher will be able to analyze the results either by class or by pupil.

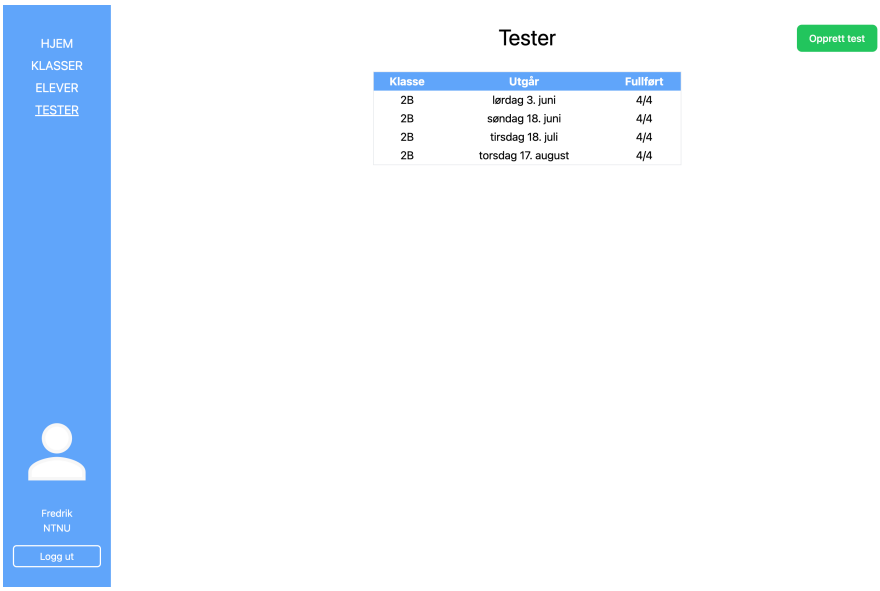


Figure 5.13: "Tests"-page

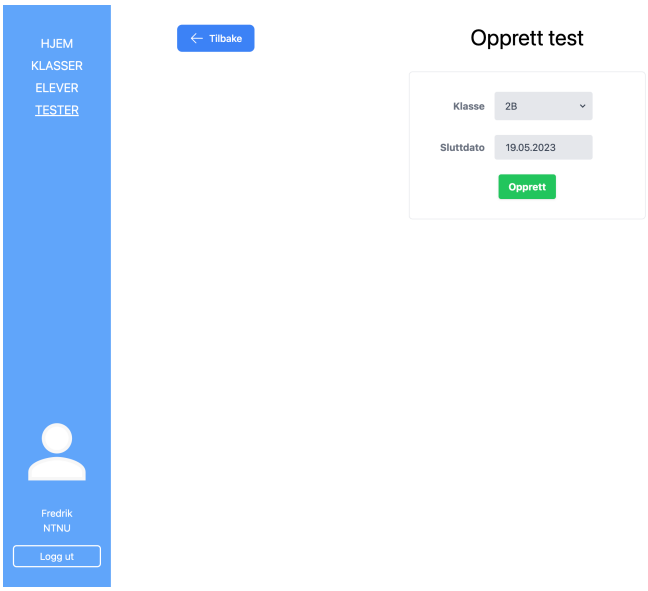


Figure 5.14: "Create test"-page

To track an individual pupil's progress, the teacher can select the pupil from the "Class"-page, or from the "Pupils"-page (Figure 5.15), which can be shown by selecting "Pupils" in the navigation pane. When clicking on a pupil in the pupil table on either page, the browser will navigate to the "Pupil"-page (Figure 5.16). This page will contain the pupil's name, class, a results table, and a results graph. The table and graph display the pupil's progress over time by showing the date and result of each test. Each row in the table will also show the progress since the previous test, making it easier to track progress from test to test.

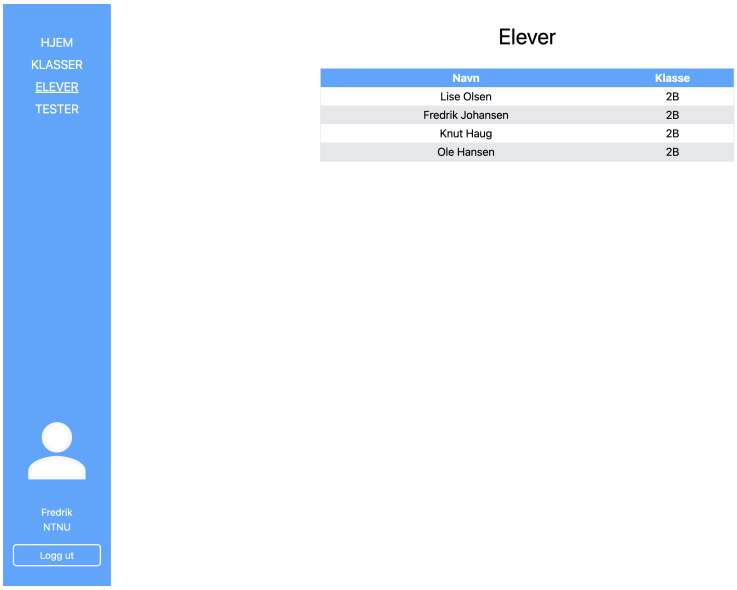


Figure 5.15: "Pupils"page

Figure 5.16: "Pupil"-page

To track the progress of a class as a whole, the "Class"-page can be revisited, now showing information about the completed tests (Figure 5.17) under the "Results"-tab. On the left side of the page, under "Test results", the teacher can select one of the conducted tests to get detailed information. This information includes the number of pupils in the low-, mid-, and high-33 percentile groups, as well as the highest score, lowest score, and mean score for the test. The expiry date and number of participants are also displayed. At the bottom of the result section, there is a table containing the names and scores of all the participating pupils. On the right side of the page, under "Overview", there is a linear graph displaying the results and progress over time, with data points for the lowest score (red), highest score (green), and mean score (blue).

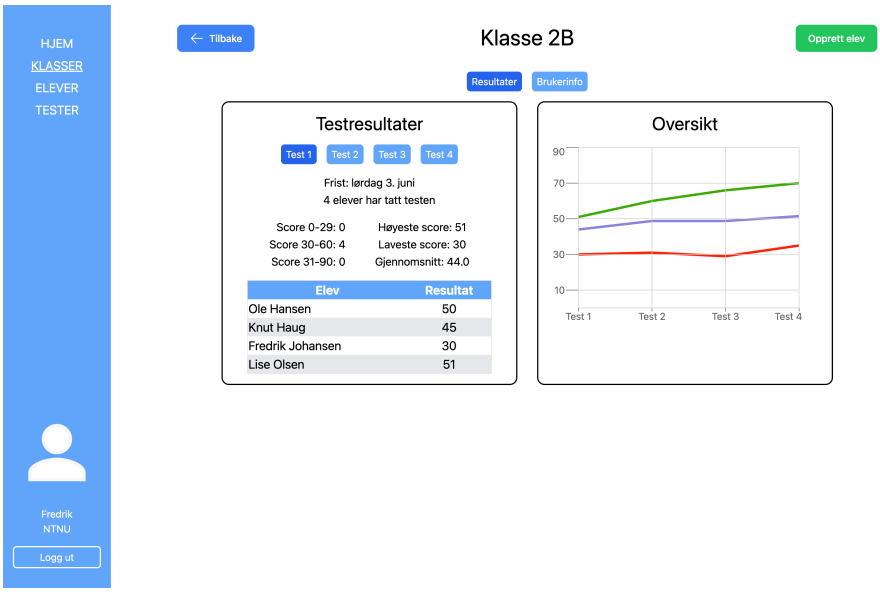


Figure 5.17: "Class"-page with test results

Word chain test design

Figure 5.18 shows the pupil landing page, where they will be directed to the login page. It is worth noting that pupils cannot create their own user accounts, but instead are provided with

pre-made accounts by their respective teachers. The page has a minimalistic design to enhance its intuitiveness and simplicity for young pupils to log in. The passwords are not hidden during input to make it easier for the pupils to see what they type. Additionally, all passwords are automatically converted to upper case letters as a design decision aimed at simplifying the login process for pupils.

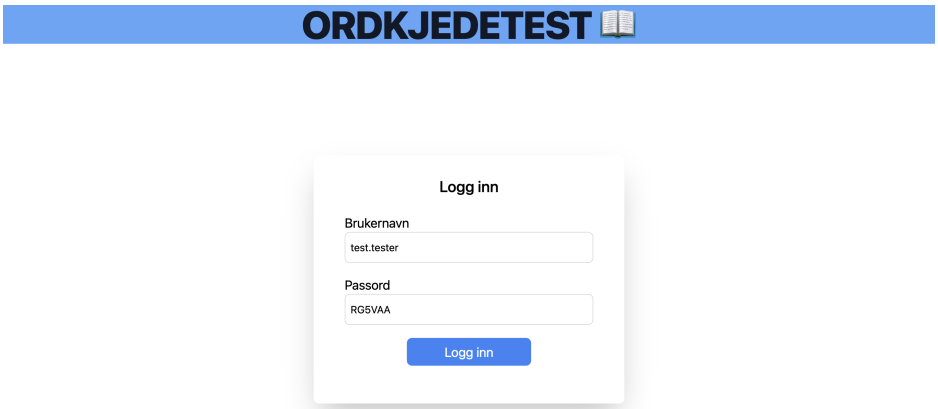
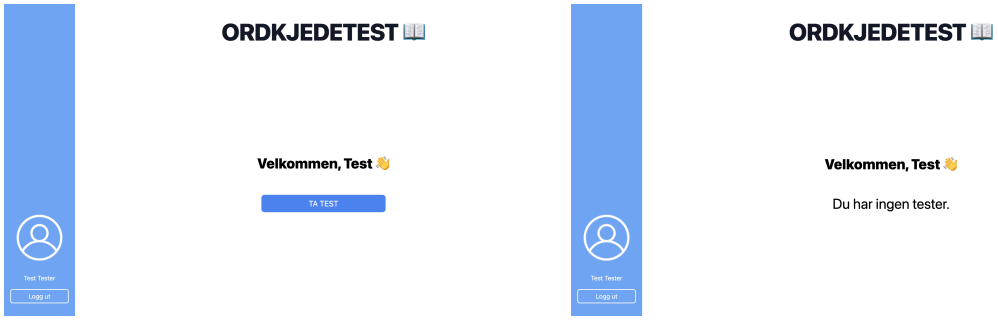


Figure 5.18: Login test-app

Upon successful log-in, the pupil will be directed to the welcome page, which is also designed to have limited options for wrongdoings. The content displayed on this page varies based on whether there is an upcoming test scheduled for the class. In the event of a pending test, a "Take Test"-button will be visible (Figure 5.19a), allowing pupils to access the test. Conversely, if no test is available, the pupil will be informed accordingly (Figure 5.19b). The welcome page features a simple sidebar that displays the pupil's name and a log-out button. To start a pending test, the user clicks the "Take Test"-button, which navigates the pupil to the "Practice"-page.



(a) With test

(b) No tests

Figure 5.19: Pupil landing page

The implementation of the "Practice"-page aims to simulate the paper-based practice that pupils undergo before taking the test. The page includes a practice word chain that allows pupils to test the functionality and become familiar with the process of dividing a word chain. The user can practice both dividing words and correcting errors. The user can click in the space between two letters and hit the space bar to insert a divider to divide words, e.g., after the first 'e' and before the second 'e' in Figure 5.20. To remove a divider, the user can click behind the desired line and hit backspace. Once the user becomes comfortable with the functionality, they can click on the "Start Test"-button to proceed to the actual test, and the timer will start.

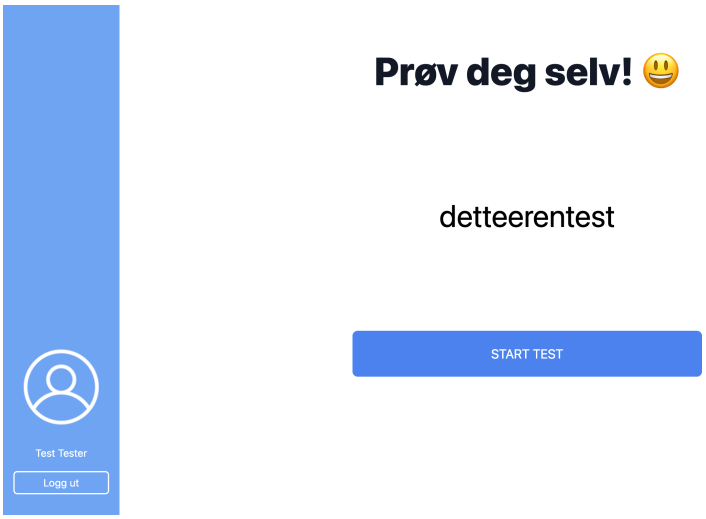
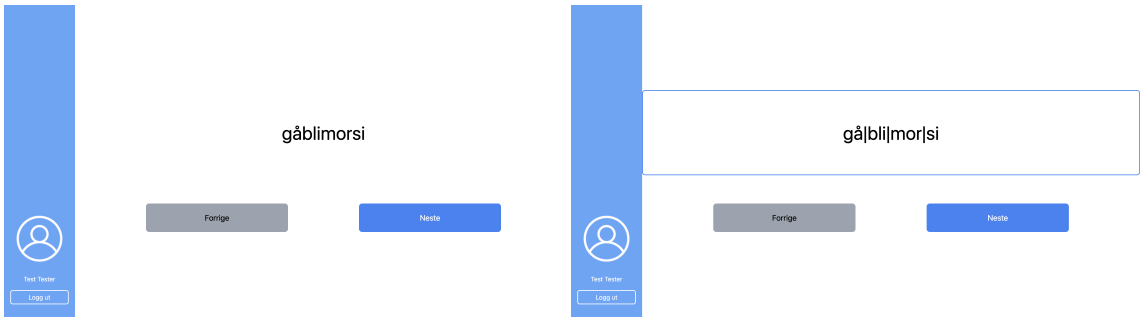


Figure 5.20: "Practice"-page

The pupil will be directed to the first word chain, as depicted in Figure 5.21a. The pupil can utilize the same functionality that they practiced on the "Practice"-page to separate the four words in the word chain, as illustrated in Figure 5.21b. Once the pupil has divided the words, they can proceed to the next word chain by selecting the "Next-button. If the pupil is not satisfied and regrets progressing to the next task, they can hit the "Previous"-button to return to the previous task.

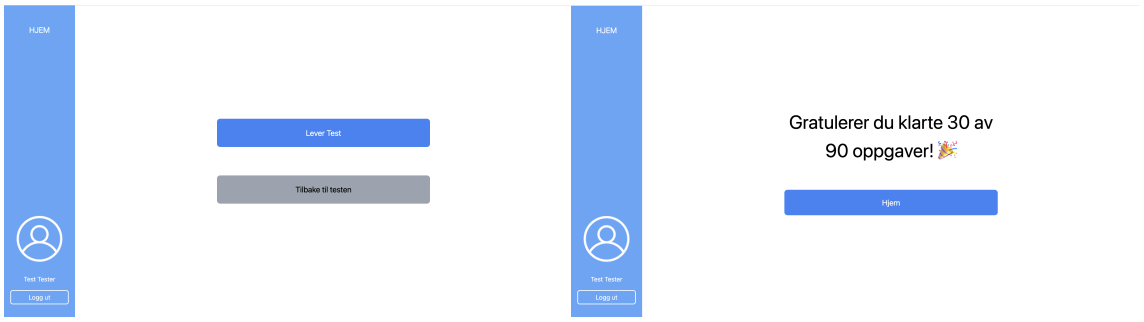
The test provides the pupil with four minutes to complete all 90 tasks. A timer will start when the test begins, and when the allotted four minutes have elapsed, the pupil will be automatically directed to the submission page. If the pupil completes all the tasks before the four-minute timeframe, they can access the submission page by selecting the "Submit Test"-button, which will appear in place of the "Next"-button on the last word chain page.



(a) No dividers (b) With dividers

Figure 5.21: A word chain page

To submit their test, the pupil can click the "Submit Test"-button displayed in Figure 5.22a, and the test score will be saved and available for the pupil's teacher. In the event that there is still time remaining, the pupil might return to the test by clicking on the "Back to test"-button. This button, however, will not be displayed if the time has elapsed. After submitting the test, the pupil is navigated to the "Result"-page, where the score is displayed (**??**). When the pupil has reviewed their score, they can return to the welcome page by clicking the "Home"-button.



(a) Submission page

(b) "Result"-page

Figure 5.22: Submission of test

5.4.2 Feedback

To improve the digital test and make it more similar to the paper test, feedback was gathered from the test participants. The feedback was evaluated, and the most helpful comments were chosen and taken into consideration for the next implementation. The seven feedback comments chosen can be seen in table 5.2.

Id	Description
C1	Remove red line under misspelled words.
C2	Add a task counter that informs the user how far they have progressed.
C3	Remove the blue line around the input field.
C4	Let the cursor remain in the same position after the user has separated the words.
C5	Add more spacing between the letters.
C6	Larger buttons to switch between tasks.
C7	Easier to divide words.

Table 5.2: Feedback comments up for changes.

In the first iteration, the occurrence of a red autocorrect line beneath misspelled words and a blue line marking the input field was deemed distracting for users. Additionally, autocorrect could be helpful for the users to divide words, so it was removed to avoid assisting the users. Another issue was the lack of feedback on test progression. It was addressed by adding a counter to enable users to track their progress more easily, similar to the number of completed words noted at the end of each line in the paper test. To enhance user experience, more space was added between letters to make it easier to add the divider at the wanted location, the cursor was made to remain in the same position after input rather than returning to the end of the line, and larger buttons for switching between tasks were implemented. Figure 5.23 shows the features commented on by the testers.

The most common feedback received was the need for a simpler way to divide words. The initial solution was a process that involved clicking between letters before pressing the space button to add a divider between them. Testers deemed this process too complicated compared to simply drawing a line, which several users felt was the primary cause of their lower scores. Simplifying the word division process would increase the similarity between the paper and digital tests, according to feedback.

Figure 5.23: Example of feedback

5.5 Second iteration

The main focus in the second iteration was to make the digital test more similar to the paper test by implementing some of the possible changes identified by the user test. In addition, the backend and teacher's dashboard were optimized for deployment to external servers. This section will present an overview of the changes made to the test frontend and the results from the second comparability test.

5.5.1 Changes to the application

A significant proportion of the feedback received was aimed at the design of the test and the potential impact of specific design choices on test performance. Specifically, comments C1, C3, C5, and C6 related to the layout of the test. In response to this feedback, several changes were implemented to improve the user experience. The red line that indicated misspelled words was removed to prevent extra help and distractions and the blue line around the input field was removed to achieve a cleaner user interface. The progression buttons were significantly enlarged to facilitate ease of task switching, even though this may not have been optimal from a design-oriented perspective. The "Next"-button, which is more frequently used, was made slightly larger than the "Back"-button to enhance accessibility. The gap between the letters was increased to make it easier to hit the wanted position. To address feedback from comment C2, a progress counter was added below the word chain to increase the similarity of the digital test to its paper-based test. All of the changes are depicted in Figure 5.24.

Figure 5.24: Visual changes

Some code changes were implemented to address C4 and C7. The cursor's position no longer shifts to the end of the chain following each click, it remains in its current position. This change made it easier for the users to remove misplaced lines. The previous method to divide words involved users clicking between the letters to indicate the correct location for the divider followed by a space-click to insert the line. The second step was removed in the second iteration, and the dividing is now done with just a click. This reduced time spent on dividing words as the user can just use the mouse-pad.

Chapter 6

Results

This chapter presents the results from the different tests conducted in this project. First the results from the usability testing, then the results from the comparability testing. The results in this section will be further discussed in chapter 7.

6.1 Usability testing

This section will present the results from the usability testing that was conducted at the end of the second development iteration.

6.1.1 Sample Demographics

An invitation to the usability test was sent to 327 teachers and special education teachers in the municipalities of Oslo, Trondheim, Bærum, Kristiansand, Stavanger, Porsgrunn, and Gjesdal. The invitation was sent by e-mail and included an informational letter (Appendix F), as well as a link to the questionnaire. A total 12 out of 327 invited teachers and special education teachers completed the usability test, where 25% (3) stated that their main occupation was "teacher", and 75% (9) stated that their main occupation was "special education teacher".

The ages of the participants can be seen in Table 6.1. Every 5-year interval in the range of 21 to 50 years old was represented by at least one and at most three participants, but there were no participants over the age of 50.

Age group	**#**	**%**
21 to 25	1	8.3%
26 to 30	2	16.7%
31 to 35	2	16.7%
36 to 40	3	25.0%
41 to 45	1	8.3%
46 to 50	3	25.0%
51+	0	0

Table 6.1: Age distribution among participants

Out of the 12 participants, 50.0% (6) stated that they had 11 or more years of occupational experience as a teacher or special education teacher, and 16.7% (2) stated that they had more than 15 years of experience. A more detailed overview of the participants' occupational experience can be found in Table 6.2.

Years of experience	#	%
0	0	0.0%
1 to 5	4	33.0%
6 to 11	2	16.7%
11 to 15	4	33.0%
15+	2	16.7%

Table 6.2: Occupational experience of the participants

The participants had experience from different grades in Norwegian elementary- and middle schools. 41.7% (5) stated that they worked with 1st to 4th-graders, 33.3% (4) answered 5th to 7th-graders, and 25.0% (3) answered 8th to 10th-graders (middle school). No participant stated that they work with more than one of the mentioned grade intervals.

When asked about how familiar they were with the current word chain tests, 75.0% (9) stated that they were either "familiar" or "very familiar". In addition, 25.0% (3) stated that they were "a little familiar", and no participant answered "very little familiar".

6.1.2 System Usability Scale

Of the 12 answers submitted to the questionnaire, the application achieved a mean SUS-score of 94.2, with a 95% confidence interval ranging from 88.4 to 99.93, reaching an A+ grading on the Sauro-Lewis Curved Grading Scale (Table 4.1). Table 6.3 shows descriptive statistics computed in SPSS 29.0.0.0 on MacOS, and the bar chart in Figure 6.1 shows the mean score for each question in the SUS-questionnaire.

	Statistic	**Std. error**
Sample size	12	
Mean	94.167	2.617
95% confidence interval		
Lower bound	88.406	
Upper bound	99.927	
Median	97.500	
Variance	82.197	
Std. Deviation	9.066	
Minimum	67.50	
Maximum	100.00	
Range	32.50	

Table 6.3: Descriptive statistics for the SUS-scores

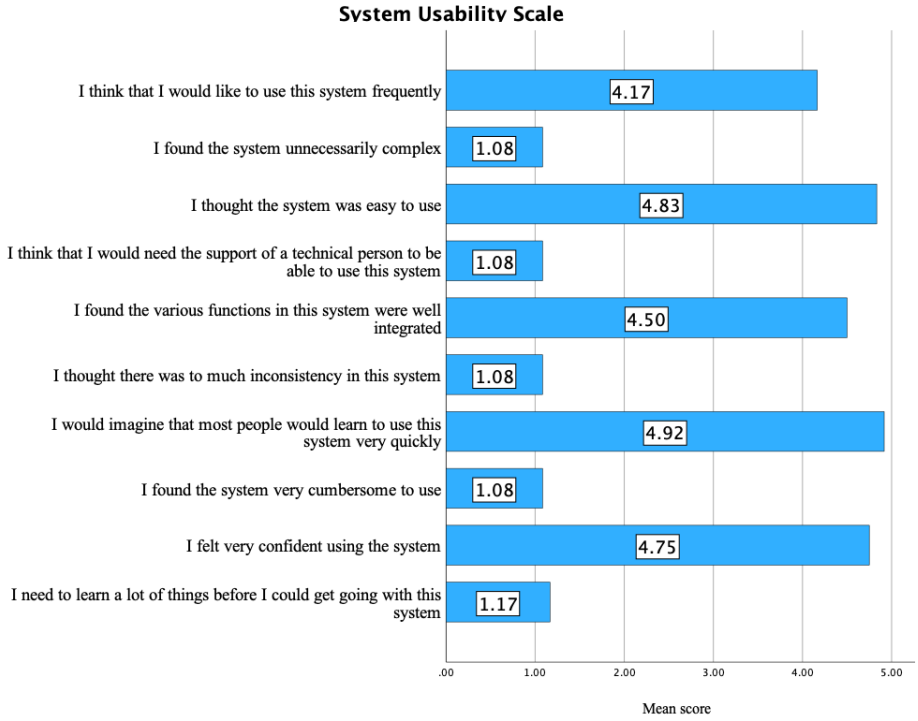


Figure 6.1: Average score per SUS-question

6.1.3 Scenarios

The first part of the usability testing questionnaire consisted of scenarios with tasks that the participants had to complete before rating how easy the tasks were to perform and optionally giving feedback in the form of textual comments. Below is a summary of both the qualitative and quantitative results from each of these tasks, in addition to linking them to their relevant functional requirement in Table 5.1. Note that some of the task descriptions have been simplified when translated to English and that some irrelevant feedback (e.g., questions about the project) has been omitted. The full task descriptions and their answers can be found in Appendix G.

Task 1: Create a teacher-user in the teacher's dashboard
Corresponding requirement: FR4
Low/High/Mean score: 4/5/4.92
Textual feedback:

- "Easy. Just like you would create a user on similar sites."
- "It was very easy to complete the task. It only took 30 seconds to create a user."

Task 2: Log out of the application
Corresponding requirement: FR4 (inherently)
Low/High/Mean score: 5/5/5
Textual feedback: None

Task 3: Log in using the user you created in Task 1
Corresponding requirement: FR4
Low/High/Mean score: 5/5/5
Textual feedback: None

Task 4: Navigate to the "my classes"-page, and create a new class.
Corresponding requirement: FR8
Low/High/Mean score: 4/5/4.92
Textual feedback: None

Task 5: Create a pupil in the newly created class.
Corresponding requirement: FR6, FR8
Low/High/Mean score: 4/5/4.75
Textual feedback:

- "I clicked on 'pupils' before realizing I could click directly on the class and create the pupil from there."
- "It is very easy to find out how to create a new pupil."
- "It was not evident that I could click on the newly created class. I first clicked on the 'pupils'-menu item."

Task 6: Navigate to the "Tests"-overview and create a test for the class you just created
Corresponding requirement: FR5
Low/High/Mean score: 4/5/4.92
Textual feedback:

- "As a special education teacher, it would be nice to have the option to create a test for only one or some of my pupils, not the whole class."

Task 7: Navigate to the class you created, and take note of the username and password of the newly created pupil
Corresponding requirement: FR7, FR9
Low/High/Mean score: 3/5/4.67
Textual feedback: None relevant

Task 8: Go to the word chain test application (supplied URL), and log in with the credentials from task 7
Corresponding requirement: FR4
Low/High/Mean score: 4/5/4.92
Textual feedback:

- "The password should be hidden."

Task 9: There should be a test available. Open the test, and try out the interface using the test-task on the first page
Corresponding requirement: FR1, FR12
Low/High/Mean score: 4/5/4.92
Textual feedback:

- "The application is easy to navigate, but 90 tests is way too much for most pupils. There should be an option to select the number of word chains for each test."

Task 10: Log back into the teacher's dashboard using the supplied credentials
Corresponding requirement: FR4
Low/High/Mean score: 3/5/4.83
Textual feedback: None

Task 11: Find the class "4B", and print out the user information for the pupils in the class using the system shortcut (CTRL/CMD + P)
Corresponding requirement: FR7
Low/High/Mean score: 3/5/4.64
Note: A rating of 1 was removed from the results as the user answered that they used an iPad, even when explicitly told to use a Windows-, Linux-, or Mac-computer.
Textual feedback:

- "Looked like the result-boxes was not perfectly aligned horizontally on the preview."
- "For once a clear and printable version of username and password! But the there should be a 'print'-button for those who do not know the system shortcut."
- "There should be a 'print'-button for those who do not use the system shortcuts. The print turned out a bit weird, as the table in the overview was outside the box."

Task 12: In class "4B", there is a pupil named Lisa Nordmann. Find this pupil, and check if she has made any progress over tha last year
Corresponding requirement: FR2
Low/High/Mean score: 4/5/4.92
Textual feedback: None

6.1.4 Participants Opinions

The last part of the usability test included a questionnaire with six questions were the participants could submit their thoughts on different aspects of the project. Four of the questions asked for text-comments, and two of the questions were in the form of a rating on the scale from 1 to 5. This section will present a brief summary of the text comments translated to English, but all the comments in their original form and language can be found in Appendix G.

Q1: Did you see any immediate potential for improvement in the application?

- "It can be demotivating for the pupils to see that there are a 100 tasks to complete. Maybe there should be a way to split it into parts so it seems like less work?"
- "No"
- "I would like information about how the pupil is performing compared to the expected level at their age."
- "I would like to be able to log in using Feide and have the results feed directly into Conexus."
- "I would like to print out using a button, and it should be possible to click directly on the pupil in the 'User info'-tab."
- "Should be possible to remove a line (from the word when conducting a test) by clicking again, not using backspace."
- "By giving the line a different color, it would be easier for the pupil to differentiate between the line they entered, and for example a lowercase 'L'."

Q2: Is there anything you thought worked especially well in the application?

- "Nice that only necessary information is given. I think this makes it easier for teachers to learn how to use it."
- "User friendly and clear."

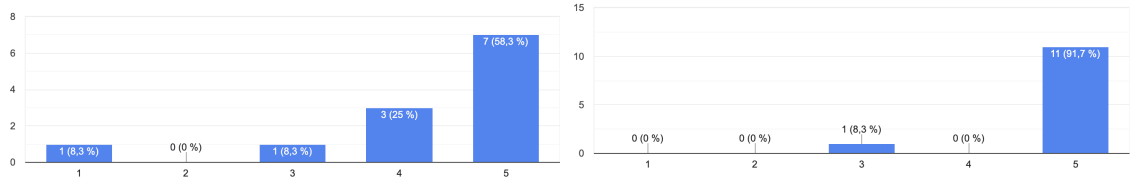
- "Intuitive and easy to use."
- "Uncomplicated, easy to get going."
- "Nice for the pupils that they only have to focus on one word chain at a time."

Q3: What are your thoughts on conducting such tests digitally, as opposed to on paper as it is today?

- "Pupils today are used to doing a lot digitally, and it makes it much easier for the teacher to detect patterns and get an overview of the pupils. I still think that the pupils could score differently than if the test was conducted on paper and that this should be taken into consideration when analyzing the results. Doing the tests digitally also avoids using unnecessary amounts of paper, in addition to saving time, something the teachers could use more of."
- "I am positive. It takes a lot of time to conduct the tests on paper, especially to correct the tests. Very nice to have the development of each pupil presented this way while also having everything stored in one system."
- "Great. Papers can be messy, but here you have all the results visually. Progress and results are saved and easy to get back to."
- "Practical and time-saving."
- "Can be harder for those with inadequate equipment or those not used to computers. On the other hand, it can be better for those who are struggling with fine motor skills and pen/pencils."
- "I think the paper is best for struggling pupils."
- "I think the test process is much easier. The results are safely stored and easily accessed. The application makes it very easy to follow the pupil and monitor their progression."
- "Easier for the teacher to conduct a digital test, and saves a lot of unnecessary photocopying and storing of paper, as well as correcting tests. I am positive about conducting tests digitally, but there is a risk of losing some information when the teacher can not see exactly what error was made in a task."
- "Good"
- "Always best to conduct such tests digitally"
- "Much better digitally, especially when the tasks appear one at a time."
- "I am fairly positive. [...] At the same time, some pupils are easily distracted when using digital tools, and there is a possibility that the program crashes. [...]"

Q4: On a scale from 1 to 5, how useful do you think it is to conduct such tests digitally?
Although 83.3% (10) of the participants answered either 4 or 5, 8.3% (1) answered 1, and another 8.3% (1) answered 3. The distribution is visualized in Figure 6.2a.

Q5: On a scale from 1 to 5, how useful do you think it is to gather the test results in a dashboard like this?
91.7% (11) of the participants answered with a top score of 5, while 8.3% (1) answered 3. The distribution is visualized in Figure 6.2b.



(a) Digitalizing word chain tests

(b) Gathering data in dashboard

Figure 6.2: Participants perceived the usefulness of digitalizing word chain tests and gathering the data in a dashboard

Q6: What are your thoughts about this project as a whole?

- "Useful project that can benefit both teachers and pupils."
- "Very good that someone focuses on this - there are many pupils in which reading/writing struggles are not detected because they are never tested. This tool makes it easier to test it on a full class, that is not how it is today."
- "Smart, but it needs to differentiate the number of word chains (the teacher must decide) and change so that the teacher can assign tests to only one pupil."
- "Very smart with digital solutions in mapping/screening-tests."
- "Nice to innovate, being part of the development in the society. Probably makes the teacher's work easier."
- "I think I'll stick to the paper based test."
- "Very good!"
- "Very positive, especially if normalized/target results are added to the application."
- "Very exciting, has my full support."
- "I am positive about the project, as long as necessary changes are made so that the test is not more difficult for the pupils to conduct digitally."

6.2 First Comparability test

The word chain test was tested on 30 university students ranging from the age of 20 to 26. The group consisted of 15 male and 15 female testers. The chosen participants were all students at the Norwegian University of Science and Technology (NTNU) because of the easy access to participants. All of the candidates answered that they regarded themselves as proficient readers. Thus it was not expected to discover any participants performing significantly worse than the others. The test group was divided into two groups, Group 1 and Group 2. Group 1 consisted of 8 male and 7 female participants, while Group 2 consisted of 7 males and 8 females. Group 1 completed the paper-based test before the digital test, and Group 2 completed the digital test first. The results of the individual participants on the test are presented in Table 2.

Table 6.4 shows the comparison of the means and standard deviation of the two groups. Group 1 scored worse (64.00) on average than Group 2 (69.27) on the analog test, and reversely Group 2 scored worse (50.40) on the digital test compared to the average of Group 2 (55.40). On average, the groups scored approximately five points higher than the other group on the test they completed second. The participants also scored a lot worse on average on the digital test (52.90) compared to the analog test (66.63). There was, in fact, only one participant that scored higher on the digital test than the mean on the analog test, with a score of 75. The lowest score on the digital test was

35. The highest-scoring tester managed to complete the whole analog test and received a score of 90, and the tester with the lowest score on the analog test scored 51.

The standard deviation on the paper-based test was also a lot higher than the standard deviation on the digital test. The average standard deviation on the paper test across both groups was 11.20 and 6.74 on the digital. The standard deviations were similar when comparing the two groups. Group 1 had a standard deviation of 6.84 on the digital test, and Group 2 had 5.83. On the analog test, Group 1 had a standard deviation of 11.36, and Group 2 had 10.76.

Group		Paper	PC
1	Mean	64.00	55.40
	N	15	15
	Std.Deviation	11.36	6.84
2	Mean	69.27	50.40
	N	15	15
	Std.Deviation	10.76	5.83
Total	Mean	66.63	52.90
	N	30	30
	Std.Deviation	11.20	6.74

Table 6.4: Mean: Average scores on groups

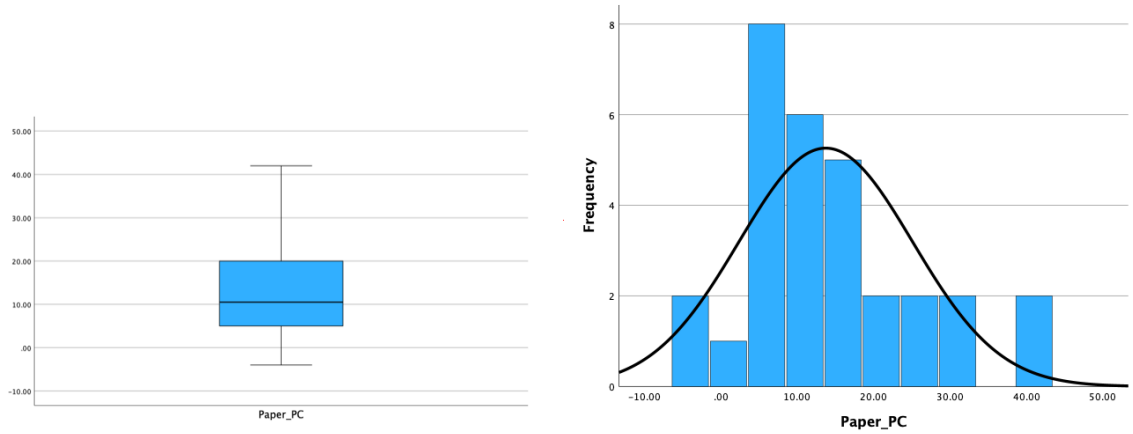
A comparison of test results was also conducted based on gender (Table 6.5). Based on the mean scores of the two genders, it was observed that males performed slightly better on both paper-based and computer-based tests. Males had an average score of 68.87 on paper-based tests, while females had an average score of 64.40. Similarly, males scored higher on the computer-based test, with an average score of 54.67, while females had an average score of 51.13. There was also a relatively high standard deviation observed in the test results, particularly in the scores obtained by males.

Sex		Paper	PC
Female	Mean	64.40	51.13
	N	15	15
	Std.Deviation	9.99	5.96
Male	Mean	68.87	54.67
	N	15	15
	Std.Deviation	12.21	7.21
Total	Mean	66.63	52.90
	N	30	30
	Std.Deviation	11.20	6.74

Table 6.5: Mean: Average scores on sex

To test if the differences in scores between the paper and digital tests are statistically significant, a paired t-test was conducted. This test included the scores of all 30 participants regardless of the group they belonged to. To ensure that the test had valid results, the data was checked to assert that it passed all the assumptions. The dependent variable, test score, was measured on a continuous scale from 0 to 90. The independent variable consisted of two related groups, paper-based test results and digital test results. There were no significant outliers in the difference between the two related groups. If there were any significant outliers, they would have been marked in the box plot Figure 6.3a, which illustrated the differences between the scores on the paper test and the digital test. The Shapiro-Wilks test resulted in a p-value of 0.03 which is less than the chosen significance level of 0.05. This suggests that there is evidence to indicate that the data deviate from a normal distribution. Figure 6.3b illustrates that the data is not perfectly normally distributed. But as the sample size was equal to 30, then the central limit theorem says that sampling distribution of sample means is approximately normally distributed [65]. Combined with

the robustness of the paired samples t-test, the assumptions were deemed as passed.



(a) Boxplot showing means

(b) Histogram showing normal distribution

Figure 6.3: Difference Paper-score and PC-score

Table 6.6 provides an overview of the results from the test. On average, the participants scored 13.73 points higher on the paper test than on the PC test, with a standard deviation of 11.37, while the standard error mean was 2.08. The lower and upper bounds represent the lower and upper bound of a 95% confidence interval. This interval suggests that there is a 95% certainty that the true mean lies between 9.49 and 17.98. The t-statistics for the test was 6.61, and the p-value was less than 0.001 on both the one-sided and two-sided tests of significance.

	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	One-Sided p	Two-Sided p
Paper - PC	13.73	11.37	2.08	9.49	17.98	6.61	29	<.001	<.001

Table 6.6: Paired Samples Test

Spearman's correlation test was conducted to assess the strength of the relationship between the scores of the two tests. The test included all 30 participants' scores. The two variables were both measured on an interval between 0 and 90, and the variables represented paired observations. Highlighted by the scatter plot in Figure 6.4, there is no apparent monotonic relationship between the two variables. Consequently, the data does not perfectly meet all the assumptions necessary to obtain a valid outcome using Spearman's correlation, and the results of a correlation test might not be accurate.

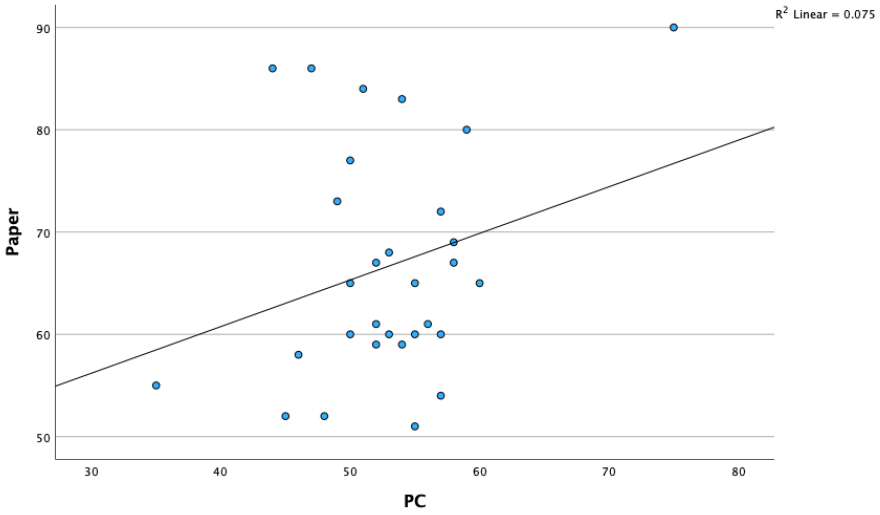


Figure 6.4: Linear regression - Comparability test 1

Regardless, a Spearman correlation test was completed to support the assumption that there was no relationship between the test scores. The results might not be perfectly accurate, but they might provide stronger support to the claim that there was no relationship between the variables. The bivariate correlation test results, presented in Table 6.7, suggest that there is a slight positive correlation between the PC-score and Paper-score variables, with a Spearman's correlation coefficient of 0.162. The p-value of 0.39 shows the statistical significance of the test.

		Paper	PC
Paper	Correlation Coefficient	1	.16
	Sig(2-tailed)		.39
	N	30	30
PC	Correlation Coefficient	.16	1
	Sig(2-tailed)	.39	
	N	30	30

Table 6.7: Correlation: Paper score and PC score

6.3 Second Comparability test

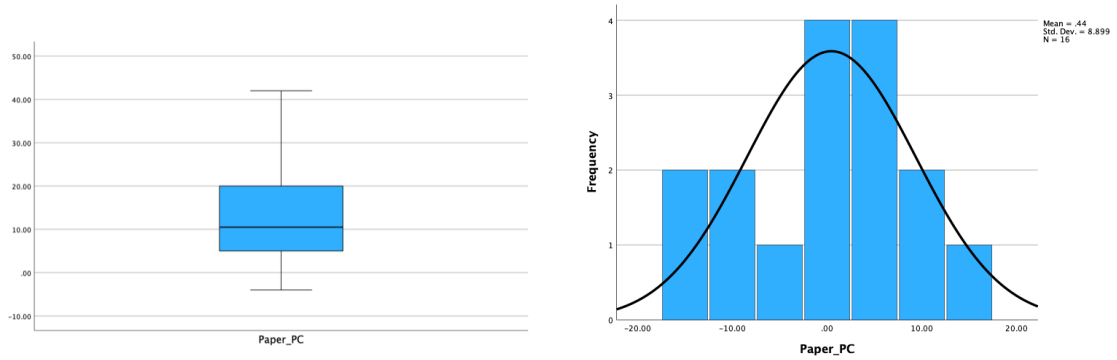
After completing the second iteration, implementing new functionality, and making changes based on the feedback and results of the first test, a second comparability test was conducted. A group of 16 participants were selected randomly from the 30 participants that took part in the first test. The only conditions for the new test group were that it had to consist of an equal number of men and women and that half of the members came from each of the two groups from the first test. Their score from the paper-based test from the first iteration was used in the comparison.

The score on the paper-based test was slightly higher on the second test (67.75) compared to the first (66.63) (Table 6.8). The reason is that the 16 participants selected scored a bit higher on average than the total average on the test. The score on the PC test increased significantly from the first test, from 53.75 to 67.31, resulting in a reduction in the difference between the paper score and the PC score from 13.73 to 0.44 points. The standard deviation of the digital test remained almost the same, being reduced from 8.32 to 8.28 when comparing the standard deviation on the scores of 16 in the new group.

	Paper 2	Paper 1	PC 2	PC 1
Mean	67.75	66.63	67.31	53.75
N	16	30	16	16
Std.Deviation	13.13	11.20	8.28	8.32

Table 6.8: Mean: Average scores on groups

Another paired samples t-test was conducted to test the statistical significance of the results from the second comparability test. The assumptions were deemed passed again. The independent and dependent variables remained the same, Figure 6.5a shows no extreme outliers, and the Shapiro-Wilks test has a p-value of .763. A p-value higher than the threshold of 0.05 indicates that there is insufficient evidence to suggest that the data significantly deviate from a normal distribution. In addition to the Shapiro-Wilks p-value, Figure 6.5b visualize the distribution. It is therefore reasonable to assume that the data does not differ from a normal distribution.



(a) Boxplot showing means

(b) Histogram showing normal distribution

Figure 6.5: Difference Paper-score and PC-score

Table 6.9 shows the results of the second paired samples t-test. After the changes were done in the second iteration, the testers scored only 0.43 points higher on the paper test than on the digital test. Compared to the first paired sample test, this is a substantial decrease. The standard deviation of paired differences was also reduced to 8.90, indicating the scores were less spread out than they were in the first test. The standard mean error is slightly bigger, at 2.23. The 95% confidence interval had a lower bound at -4.31 and 5.18, suggesting that the true difference in mean was much closer to zero than on the first test. The t-value of 0.2 supports this claim, as it is significant with a degree of freedom of 16 and chosen significance level of the two-sided test of 0.05. A p-value of 0.85 is another indicator that there is little difference between the mean of the two sets of scores.

	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	One-Sided p	Two-Sided p
Paper - PC	.43	8.90	2.23	-4.31	5.18	.20	15	.42	.85

Table 6.9: Paired Samples Test

Similar to the first test, the two variables remained on an interval scale, and the variables represented observations in pairs. Additionally, it can be observed from the Figure 6.6 that there is a clear trend in the data. The trend in the data is increasing and indicates a positive monotonic relationship. The data fulfills the assumptions, indicating that the results from the Spearman Correlation are valid.

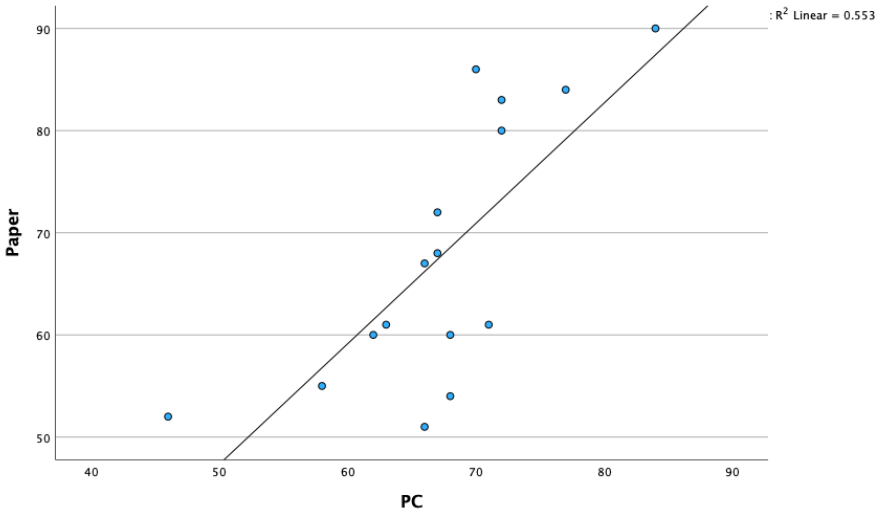


Figure 6.6: Linear regression - Comparability test 2

Following the second iteration, the correlation coefficient had increased significantly from 0.16 to 0.75. This indicates that there is an even stronger positive relationship between the results of the digital test and the paper-based test after the second iteration. The 2-tailed p-value was significantly reduced, decreasing from nearly 0.4 to <0.001, asserting that the correlation is significant at the 0.01 level (2-tailed).

		Paper	PC
Paper	Correlation Coefficient	1	.75
	Sig(2-tailed)		<.001
	N	16	16
PC	Correlation Coefficient	.75	1
	Sig(2-tailed)	<.001	
	N	16	16

Table 6.10: Correlation: Paper score and PC score

6.4 Stanine Scale

The scores on the two tests are converted to stanine values. Table 6.11 include the stanine scale for the first comparability test with 30 participants. Table 6.12 shows the stanine values for the second comparability test only including 16 participants.

Stanine	1	2	3	4	5	6	7	8	9
Number of participants	1	2	4	5	6	5	4	2	1
Paper	51	52	54-59	59-60	61-67	67-73	77-84	86	90
PC	35	44-45	46-49	50-52	52-54	55-57	57-58	59-60	75

Table 6.11: Stanine: First iteration

Stanine	1	2	3	4	5	6	7	8	9
Number of participants	1	1	2	2	4	2	2	1	1
Paper	51	52	54-55	60	61-68	72-80	83-84	86	90
PC	46	58	62-63	66	67-68	70-71	72	77	84

Table 6.12: Stanine: Second iteration

The Stanine scale is used to detect struggling pupils, and the results from the first paper-based test show that the user scoring 51 and 52 were in the two lowest groupings, which are the groups that are the subjects to being further tested for reading literacy difficulties. On the same iteration of the digital test, the two lowest groups scored 35 and between 44 and 45. In the second test, the two lowest groups remained the same, this time with only one user in the second group. The scoring of the two lowest-scoring digital groups increased to 46 and 58.

Chapter 7

Evaluation

This chapter aims to discuss the results from the usability testing and comparability testing and evaluate the fulfillment of the requirements defined in Section 5.2.

7.1 Usability Testing

This section will use the results from the usability testing to evaluate the application, mainly the teachers' dashboard, from multiple perspectives. It will begin by using the tasks/scenarios-results and SUS-scores to evaluate to which degree the application aids the teachers in completing the most vital tasks, before analyzing the textual feedback provided by participants to identify aspects of the application and its design that were well-received and areas with potential for enhancement.

7.1.1 Tasks and Scenarios

With a combined mean score of 4.86 out of 5.00 for all the scenarios in the usability test and no individual mean score lower than 4.64, the conclusion is that the application assists the teachers in completing the primary tasks to a strong degree. This claim is supported by the results from the SUS-evaluation, where 91.67% (11) of the participants answered "Strongly disagree" on the questions "I found the system unnecessarily complex", "I think that I would need the support of a technical person to be able to use this system", and "I found the system very cumbersome to use". In addition, (the same) 91.67% (11) of the participants answered "Strongly agree" on the question "I thought the system was easy to use". The lowest-scoring single task, where participants were asked to print out the list of pupils and their login credentials in a class, identified that the application could benefit from a designated "print out"-button.

7.1.2 Textual feedback

Two of the comments from the usability test praised the ease of registering and logging in with a new user, and 83.33% (10) of the participants pointed out that they appreciated the simple, clear, and intuitive design of the system. The participants' comments were more exhaustive when they were asked to voice their concerns. Below is a list of the issues/points of potential improvement/feature requests mentioned in these comments, accompanied by some notes from the developer's perspective.

- Lack of a "print out"-button

This is a feature that would be relatively easy to implement and should be considered in subsequent iterations of the application.

- It is not possible to go directly from the "user info"-page to a pupil page by clicking on a pupil in the list.

Before the pilot phase of the usability testing, this was actually possible. However, when conducting the pilot tests, it was discovered that this often lead to unintended clicks when participants tried to copy user credentials from the list, and the hyperlink function was therefore disabled.

- It can be demotivating for the pupils to see that there are many tasks to complete. Maybe there should be a way to split it into parts so it seems like less work.

As the project aimed to create an application that deviated as little as possible from the paper-based test, splitting the test into multiple parts was not an option at this point, and should be evaluated from a pedagogical perspective before being implemented in later iterations.

- It should be possible to remove a line (from the word when conducting a test) by clicking again, not using backspace.

This was attempted during the first iteration, but a solution that seemed natural and intuitive was not found. This feature request can be explored and evaluated in later iterations.

- By giving the line a different color, it would be easier for the pupil to differentiate between the line they entered ("|"), and for example a lowercase "L" or uppercase "I".

This is a feature that would be relatively easy to implement and should be considered in subsequent iterations of the application.

7.2 Comparability Testing

This section discusses the findings presented in the Section 6.2 and Section 6.3. The section aims to provide deeper insight into the results and extrapolate their implications regarding the potential for the digital test to replace the analog test.

7.2.1 First test

In the normalization test conducted with 8 schools from Rogaland, the mean and standard deviation of each grade for the paper-based test were obtained. The mean was noticeably lower in all of the grades compared to the mean from the paper-based test conducted in this project. The highest mean in the normalization test was the tenth grade with a mean score of 49.7. This is approximately 17 points lower than the average from the mean of the paper-based test in the first comparability test (Section 6.2). This could be due to the fact that the participants in this project were higher education students who all considered themselves proficient readers. None of the participants scored lower than the tenth-grade average. A normalization test with several hundred participants would be a good foundation for comparison, unfortunately, the scores are not comparable to the scores of the participants in this project. The goal of the test was to evaluate and draw comparisons between the students' performances on the traditional paper-based test and the new digital format. Although the students at NTNU do not represent the reading literacy of the entire Norwegian population, it provides a strong indication of the correlation and relationship between the test performance on the two types of tests. It was more important to ensure a sizeable sample size over a diverse participant group. Therefore, the results from the comparability testing in this project will be used as a foundation for comparison, rather than the results from the normalization test.

The standard deviation of the test showed that the standard deviation on the paper test was a lot higher than what it was on the computer-based test. This might be caused by the fact that none

of the participants had tried the digital test earlier while some participants had tried the paper-based test in their early education. Both of the standard deviations were regarded as natural, compared to the standard deviation achieved in the normalization project. Differences will occur when testing reading literacy, even for proficient readers.

The results of the first comparability test indicated that it might have been advantageous to have completed a test before taking another. The group that completed a test last scored approximately 5 points better on average compared to the group that completed the same test first. This was the case for both the digital and the paper-based tests. This could be due to a number of factors, such as a learning effect, familiarity with the test content, or simply getting more comfortable with the testing environment or process. Although, this could also be a coincidence.

The groups consisted of only 15 individuals, meaning extreme scores could easily sway the average. Table 7.1 illustrates the distribution of the highest and lowest scorers on each of the tests between the two groups.

	Group 1	Group 2
Paper	4/4 worst scores	6/8 best scores
PC	7/8 best scores	4/5 worst scores

Table 7.1: Best/Worst performers in groups

Group 1 included the four participants with the lowest scores on the paper test, whereas Group 2 included six of the eight top scorers. In contrast, on the digital test, Group 1 had seven of the top eight scores, while Group 2 contained four out of the five lowest scores. The remaining participants were distributed more evenly. The lopsided distribution of top performers might not be affected by the order of tests but rather a matter of chance. Regardless, the possible advantage of completing one test after another was leveled out when computing the total average by dividing the participants into two testing groups.

When completing the statistical analysis of the test results, all of the scores were studied as a whole. Parting the test group in two makes the analysis less affected by potential learning effects.

The difference in scores was examined using a dependent samples t-test. The difference was 13.73, which is quite a large gap between test scores. The chosen significance level for the test was 0.05, as it provides a fitting balance of type I and type II errors for this project. A two-tailed test was also deemed more fitting than the one-tailed as the alternative hypothesis is non-directional [66]. The derived t-value of 6.61 significantly exceeded the critical value associated with a 0.05 significance level and 29 degrees of freedom (Figure 4.2). Combined with a two-sided p-value below .001, implies that the null hypothesis can be rejected. There is a fair assumption that there is a significant difference between the mean scores on the digital and the analog test and that the difference did not occur by chance. An ideal result would be that the scores were equal with a difference of 0 and no standard deviation on the dependent t-test, showing that the digital test provides the exact same results as the analog test. By rejecting the null hypothesis, this was evidently not the case after the first iteration.

However, the digital test could still prove to be a valid replacement for the analog test as long as the results on the digital test were consistently higher or lower than the results on the analog test. As long as the test could separate the struggling readers from the rest, then it could be a viable option for the analog test. A strong relationship between the two scores is an indication that better readers score well on both tests and conversely, worse readers score worse. If the test managed to correctly rank the pupils, it could be a useful replacement.

The correlation test demonstrated the relationship between the two variables. By looking at the scatter plot, it was apparent that there was no clear correlation between the scores. The regression line shows only a slight monotonic relationship, in addition to Spearman's correlation coefficient of 0.162, which is a negligible correlation. A non-significant p-value of 0.393 further implies that there is a high probability that the slight correlation that occurred was by chance, and further support the claim that there was no clear relationship between the scores. The comparability test

conducted after the first iteration was not able to yield the same results as the paper-based test. The only possibility for the digital test to be an option for assessing decoding skills is if it manages to separate the weakest readers from the stronger readers. If it is able to do so, it is not vital that it does not manage to correctly rank the stronger readers.

Stanine scale

The stanine score could help make it easier to check if the digital test can detect the same struggling pupils as the analog test does. The two lowest levels are the levels where further literacy testing is recommended for the pupils falling into that level. In the first comparability test, the three participants in the two lowest levels were spread out on the stanine scale from the digital test. One of the three remained in the two lowest levels, while the two others jumped to levels 3 and 6. The only participant in level 1 on the analog test was the one jumping to level 6 on the digital test. This supports the claim that the first iteration of testing is not a valid option. There are small differences in points separating the different levels in the scale, but a jump from level 1 to level 6 is a significant increase. A similar jump on the scale from the analog test represents a point increase between 16 and 22 points. The error was maybe even more evident in the opposite direction. The lowest score on the digital test was 35, which was significantly lower than the other scores. This participant was not in the two lowest groups on the analog test. Even if it is not a huge issue to falsely detect struggling pupils as further literacy assessment would recorrect the assessment, it is not a desired outcome as it requires more resources.

The results from the first comparability test all imply that the digital tool does not provide sufficient similarity to the analog test regarding the assessment of reading literacy. The second iteration needed to introduce several changes to make the two tests more similar.

7.2.2 Second test

The most notable difference between the scores of the first test was that the participants performed considerably worse on the digital test compared to the analog one. Some of the top scorers on the analog test were amongst the lowest performers on the digital test, indicating that the technical aspect of the test had a big impact on the test results. To minimize the impact the technical abilities of the testers had on the score, the application was modified to make it easier to use the interface to complete word chains.

To assess the effectiveness of these modifications, a new test group was formed. It consisted of eight participants from each of the two groups from the first testing iteration to mitigate potential biases from the initial testing sequence. Additionally, an even gender distribution was maintained, with 8 males and 8 females, to offset any gender-based influences observed in the first round of testing, where male participants outperformed females in both formats.

The changes made to the application seemed to achieve the desired effect. Digital test scores increased to 67.31, a significant jump from the initial average of 53.75 attained by the same 16 participants on the initial digital test. For comparative purposes, the result from the paper test of the first iteration was used as a comparison in the second iteration as well. To ensure that the results still were a valid measurement of their current decoding level, the participants did not partake in any activities that should affect their reading level in the two weeks gap between the two iterations of testing. The selected group for the second iteration testing had an average score of 67.75 on the paper test, and the gap between the means was reduced significantly, from 13.73 to only 0.43. The standard deviation increased on the paper-based test on the newly formed smaller group (11.20 to 13.13). On the digital test, there was also an increase in standard deviation between the first test and second test when compared to the whole group (6.74 to 8.28). When only comparing the results of the participants in the newly formed group, it was almost identical (8.32 to 8.28). Again, non of the standard deviations were deemed unnatural from what to expect from a reading assessment compared to the normalization test [6]. To compare the averages of the paper-based test and the second digital test, another dependent sample t-test was completed.

The difference in means was reduced to 0.43, implying that the scores on the two tests were much more similar after the test. The standard deviation on the differences was also reduced to 8.90, signifying that the deviation from the mean difference was reduced, but the data is still somewhat dispersed from the mean. The lower and upper bound in the 95% confidence interval was -4.31 and 5.18, respectively. This indicates that some participants scored better on the digital and some scored better on the analog, but the difference was not a massive amount of points for the larger part of the participants. Contrary to the first score, where the difference was significantly larger.

Statistical analysis of the results from the first test implied that the null hypothesis could be rejected and that there was a significant difference between the scores. Values obtained from the second test resulted in different implications. The t-value was only .20, which is below the significant p-value for a two-tailed test with a significance level of 0.05 (Figure 4.2). The two-sided p-value was 0.85, which is well above the significant value of 0.05. These two values imply that any difference observed in scores is likely due to chance and not a real effect. Compared to the first test results, where there was no clear relationship between the scores, the second test implies that there are reasons to believe that there is a relationship. The mean difference is a good indicator that users score similarly on the two tests. However, the digital test would still not be a valid replacement for the analog test if each tester does rank similarly compared to the other testers. To validate this, Spearman's correlation was used.

Compared to the first scatter plot, there was a more apparent correlation between the two scores. This is reflected by a correlation coefficient that increased drastically. The coefficient was 0.75 after the second iteration of implementation, which corresponds to a "Strongly positive correlation" (Table 4.2). The p-value was $<.001$, which is significant, implying that there is a high chance of a correlation between the two scores of the two tests. Both of these values imply that the two tests have fairly the same test basis. The strong correlation argues that the two tests place the participants in roughly the same order, and the p-value tells us that this correlation is significant. The increase in correlation indicated that the test is closer to the analog test but still not a perfect replacement.

As aforementioned, the most vital function of the test is to single out the struggling pupils. It is not so vital that the best readers get a perfect equivalent test result on the digital test as the schools do not need to detect the best pupils but the ones who require extra attention. Looking at the scatter plot in Figure 6.6, two of the lower scores on the paper test scored much better on the digital test. One of the participants scored 51 on the paper test and 66 on the digital test. A score of 66 on the digital test is almost on average, but 51 on the paper test is almost 17 points and approximately 25% below the average. To be a perfect replacement, the digital test needs to be able to single out the struggling readers.

Stanine scale

The stanine scale can be used to further investigate the distribution of results. For example, the participant mentioned in the previous paragraph was in stanine level 1 on the analog test and in stanine level 4 on the digital test, and the pupil in level 2 on the analog test dropped down to level 1 on the digital test. The best readers on the analog test did perform fairly well on the digital test, with none of the participants in the top four levels performing worse than level 5 on the digital. Unfortunately, it was not the ability to correctly separate the best from the rest, but rather the worst from the rest the test was designed to do. It was fine margins separating the lowest-performing readers, with none of them really showing signs of worrying results. Still, such a significant boost in score from analog to paper-based, as described in the previous paragraph, might be worrying. The stanine scale shows some tendencies that the digital test might not be able to detect struggling pupils, but a sampling set of only 16 participants makes it hard to be certain of how well the test performs.

7.3 Requirements

This section evaluates the system based on the requirements defined in Section 5.2. The requirements will be labeled based on their level of fulfillment, which are "Attained", "Partly attained", and "Unattained".

7.3.1 Functional requirements

The fulfillment of the functional requirements is presented in Table 7.3. Some of the requirements (FR1, FR2, FR4, FR5, FR6, FR7, FR8, FR9, and FR12) have been tested and validated through usability testing, while others are self-evident. FR13 was not attained as adding the role of a system administrator would add another layer of complexity to the data model, complicate the registration process, and the system would need a separate administration panel for the role to function properly. FR10 and FR14 were not attained as they would take more time to implement than the benefits they would bring to the system justify. Note that all the functional requirements with a priority of "High" in Table 5.1 have been attained, and none are labeled "Partly attained".

Id	**Description**	**Fulfillment**
FR1	A pupil should be able to perform a word chain test	Attained
FR2	A teacher should be able to monitor the progress/level of each pupil	Attained
FR3	A teacher should be able to monitor the progress/level of each class	Attained
FR4	A user should be able to log in as their role (teacher or pupil)	Attained
FR5	A teacher should be able to create test sessions	Attained
FR6	A teacher should be able to create user credentials for the pupils	Attained
FR7	The teacher should be able to print out a list of pupils and passwords	Attained
FR8	A teacher should be able to create a class or pupil	Attained
FR9	The system should be able to auto-generate passwords when a pupil profile is created	Attained
FR10	The system should be able to output anonymous test reports as data files	Unattained
FR11	The teacher should be able to print out a detailed list of class results	Attained
FR12	A pupil should be able to train on test tasks to familiarize themselves with the application	Attained
FR13	A system administrator should be able to create/delete user credentials for the teachers	Unattained
FR14	The pupils should be able to access a picture-based walk-through	Unattained

Table 7.2: Fulfillment of functional requirements

7.3.2 Non-functional requirements

The fulfillment of the non-functional requirements is presented in Table 7.3. NFR1, NFR2, and NFR3 were validated through visual confirmation during comparability testing. The usability test revealed that NFR4 and NFR5 were attained while NFR6 was not, something that will be discussed in Chapter 8.

Id	Description	Fulfillment
NFR1	The application should support modern web browsers and devices with keyboard and mouse for efficient assessment	Attained
NFR2	The application should provide immediate feedback to users for any action they perform (mouse click, keyboard input, etc.)	Attained
NFR3	The application should start within 2 seconds of accessing the URL	Attained
NFR4	At least 80% of users should find the application easy to use on the System Usability Scale (SUS). Users should either agree or strongly agree with the statement "I thought the system was easy to use", and either disagree or strongly disagree with the statements "I found the system unnecessarily complex", "I think that I would need the support of a technical person to be able to use this system", and "I found the system very cumbersome to use"	Attained
NFR5	At least 80% of users should find the application easy to learn on the System Usability Scale (SUS). Users should either agree or strongly agree with the statement "I would imagine that most people would learn to use this system very quickly", and either disagree or strongly disagree with the statement "I needed to learn a lot of things before I could get going with this system."	Attained
NFR6	At least 80% of users should either agree or strongly agree with the statement "I think that I would like to use this system frequently." on the System Usability Scale (SUS).	Unattained

Table 7.3: Fulfillment of non-functional requirements

Chapter 8

Discussion, Conclusion and Further Work

This chapter will first discuss the results, relevance, and limitations of the usability testing and comparability testing before concluding on the research questions provided in Section 1.2.

8.1 Discussion

This project aimed to create a digital word chain test. Remote unmoderated usability testing was conducted by sending e-mail invitations to teachers and special education teachers, and comparability tests were conducted on fellow students on campus to assess the similarity between the digital and analog results. Initially, it was intended to conduct these tests on elementary school classes through moderated usability testing involving both teachers and pupils and by comparing the outcomes of the digital test with those of the paper-based test. However, the intended testing could not be carried out as planned due to a cancellation caused by an internal miscommunication within Trondheim Kommune, the local municipality responsible for granting access to the classes.

8.1.1 Usability testing

An invitation to the usability test was sent out to 327 teachers and special education teachers from seven different Norwegian municipalities, of which 12 chose to participate in the usability test, resulting in a participation rate of only 3.7%. Despite the low participation rate, the test managed to yield a 95% confidence interval of the SUS-score that sat comfortably within the A+ grade on the Sauro-Lewis CGS. In addition, a sample size of 12 is generally considered more than enough for a usability test to be capable of detecting a vast majority of problems in a system [67]. With this in mind, we are confident that the usability testing gave an accurate SUS-rating, and that a large majority of the possible usability problems in the system were unveiled. However, a sample size of 12 meant that it was unfeasible to draw any conclusions on how demographic factors like age and occupational experience affected usability and perceived usefulness. In addition, the usability test was only conducted on teachers and special education teachers and heavily focused on the teacher's dashboard. As a consequence, no data was collected on how pupils interact with the application.

Only 75.0% (9) of the participants agreed or strongly agreed with the statement "I think that I would like to use this system frequently", causing NFR6 to be unattained. This stands in contrast to the other SUS-questions where the system scored higher. One reason for this could be how the question is worded. Since paper-based tests are not that frequently used today, teachers might naturally lean towards giving lower ratings, as the system's usability wouldn't affect their usage frequency. Considering this, it seems that this SUS question may not have been a good basis for

a non-functional requirement.

The method of inviting users by email introduced some limitations to the usability test. It is possible that teachers who are comfortable with technology and enthusiastic about using digital tools are more likely to participate in a usability test for a product that aims to digitize a tool they already use. This introduces a risk of self-selection bias, where technically adept participants are over-represented. This can, in turn, inflate usability scores and perceived usefulness, as these participants are more capable of solving the given tasks or are more enthusiastic about digitalization than teachers in general. Meanwhile, a nonresponse bias can have the same effect on the test results. This bias would work in the opposite way when teachers that are not technically adept or disinterested in digitalizing word chain tests refrain from participating. This, in turn, could mean that fewer participants struggle with tasks or rate their perceived usefulness in the low numbers, which make the ratings artificially high.

8.1.2 Comparability testing

The word chain test is designed to detect young pupils with reading literacy difficulties. To ensure that the digital replacement for the test was a valid alternative to the analog test, it would be beneficial to test the performance of pupils in elementary and middle school. Unfortunately, due to the already mentioned miscommunications within Trondheim Kommune, none of the schools approached were able to participate. Compounding this issue, the word chain test application was not compatible with iPad, thereby excluding a significant number of other potential schools in different municipalities. With limited time, the comparability testing was solely conducted with participants pursuing higher education, all of whom were in their twenties. This could have influenced the results of the tests, as all the participants were more technically adept than what one could assume of a young pupil and therefore more effective in dividing words on the digital interface. The results from the second comparability test showed that the two tests were quite close on their test basis and that the digital might be a valid option with some further work. The certainty of this result is compromised by only being tested on more highly educated participants and not on pupils in early education.

In addition to only being tested on university-level students, the sample size of the two tests was not particularly big. When the analog test was normalized in 1997, between 159 and 240 pupils from each grade participated in the test to find the mean and standard deviation. With a larger sample size, the results would most likely have been closer to a normal distribution, and the statistical analysis would yield a more confident result. Especially the second comparability test had few participants, and a few unusual data points might affect Spearman's correlation and the paired samples t-test result. It was also difficult to utilize the stanine scale with few testers with no struggling readers. Although the correlation was deemed significant in the second Spearman's correlation test and the dependent samples t-test indicated a significant reason to assume that the mean scores were similar, it is difficult to assert that the tests are equal with small sample size.

It is also possible that the scores on the digital test in the second comparability test were higher due to the learning effect. As seen when comparing the results in the first comparability test, the participants scored higher on average on the second test they completed. This effect might influence the score on the second comparability test. Ideally, a new group of participants should have partaken in the second test to mitigate this effect.

The results indicate that the digital test has the potential to assess students' reading literacy as effectively as the analog test. However, it is difficult to conclude due to the limitations imposed by the restricted sample size and the potential influence of the learning effect on the scores.

8.2 Conclusion

This project has produced a digital word chain test with an accompanying dashboard for monitoring the results of classes and pupils. In terms of usability, the applications scored well on all tasks and

were praised for their simple design by the teachers who participated in the testing. Still, we have no data on how actual children would interact with the system. The comparability tests suggested that the system was an adequate alternative to test children's literacy skills, although with some uncertainty. A more detailed conclusion will be drawn by answering the research questions below.

RQ1: Are the results from the digitalized word chain test comparable to the results of the paper-based test?

The results from the comparability tests showed some promising results for the digital test to be able to replace the paper-based test. The difference in scores on the two types of tests became fairly similar after the second iteration. A high correlation between the two types of scores showed that the digital test did manage to rank the participant to a great extent. Some of the scores highlighted some potential flaws, but with a small sample size, it is hard to tell if it is noise in the data or an actual issue. Either way, comparing the results after the first iteration to the results after the second iteration, there was a positive trend in similarity.

There might still be some improvement needed in the sense of making it easier to divide the words to level out any potential difference in the technical ability of the pupils. It is difficult to actually conclude what needs to be done to make it fit for assessing children without testing on children. Similarly, it is not credible to conclude the performance of the test without testing on younger pupils.

RQ2: What are some challenges when creating digital word chain tests?

Some teachers raised concerns about the validity of the tests when conducted digitally. In addition to the overall comparability discussed in RQ1, a digital test can add more factors to the assessment results, like differences in hardware and technological skills. A vital part of the word chain test is that all pupils are taking the test under the same conditions, using pen and paper, to ensure that the test result accurately measures the pupil's literacy skills. When digitalizing the word chain test, one should therefore take measures to ensure that differences in hardware and technological skills are minimal factors in the result of the test.

The usability questionnaire also revealed that some teachers use the word chain test outside the recommended framework, especially when assessing children who are struggling with their literacy. For example, it was suggested that one should add the possibility of reducing the number of word chains and assigning a test to only one pupil instead of a whole class. Although the goal of this project was to create an application as similar as possible to the official word chain test and framework, one should consider that a too rigid system could impose limitations that render it less useful or completely useless to some users.

RQ3: Do teachers see the value and show interest in using a digitalized version of the word chain test?

The feedback from the usability tests was almost exclusively positive, with only one teacher stating that they would prefer to use the paper-based word chain test. Key motivational factors for using a digital word chain test seem to be less use of paper, easier organization of a test, automatic correction, and better overview of the pupils' progress. However, there was a notable difference in the perceived usefulness of digitalizing the word chain test, and the perceived usefulness of gathering and displaying result data in a dashboard. The feedback from the teachers indicates that there is a great interest in digitalizing the monitoring, storing, and structuring of the result data, but somewhat more skepticism toward digitalizing the assessment method itself.

8.3 Further work

In this section, the project's potential future work will be discussed, along with suggested steps to be taken moving ahead.

The usability test helped identify multiple flaws and deficiencies that should be investigated and fixed/added as features. A few examples are:

- A print-out button for pupil user credentials
- Custom number of word chains per test
- Assigning tests to single pupils

In addition, usability and accessibility issues regarding the testing interface were questioned by some teachers, along with their input on how it could be enhanced. Therefore, it is suggested that any further development of this application investigates if any possible adjustments can enhance the efficiency of the testing interface.

A large limitation of this project has been the lack of testing on real pupils in classes. Further studies should therefore conduct test-sessions in a real elementary- or middle school class, with accompanying interviews of pupils and teachers. Before any results from the digital word chain test can be used to evaluate the reading proficiency of pupils, the test has to be normalized on real pupils through exhaustive testing and comparison to the analog test by or with help from experts in pedagogy.

Lastly, the architecture of the system has the potential to support multiple types of assessment tools and tests, as most parts of the dashboard and backend are not specific to the word chain test. Future projects can therefore investigate the possibilities for adding new types of tools and tests to the system. Moreover, it is worth considering the option of incorporating the system with other assessment tools commonly utilized in schools, such as tools for assessing reading proficiency and screening for dyslexia.

Bibliography

[1] OECD, *PISA 2018 Assessment and Analytical Framework*. 2019, p. 308. DOI: 10.1787/b25efab8-en. [Online]. Available: https://www.oecd-ilibrary.org/content/publication/b25efab8-en.

[2] H. Sigmundsson, A. Dybfest Eriksen, G. S. Ofteland and M. Haga, 'Gender gaps in letter-sound knowledge persist across the first school year', *Frontiers in psychology*, vol. 9, p. 301, 2018.

[3] F. Jensen, A. Pettersen, T. S. Frønes *et al.*, 'Pisa 2018', *Norske elevers kompetanse i lesing, matematikk og naturfag. Oslo: Universitetsforlaget*, 2019.

[4] H. Sigmundsson, A. D. Eriksen, G. S. Ofteland and M. Haga, 'Letter-sound knowledge: Exploring gender differences in children when they start school regarding knowledge of large letters, small letters, sound large letters, and sound small letters', *Frontiers in Psychology*, vol. 8, 2017, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01539. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01539.

[5] M. Csikszentmihalyi, 'Flyt og læring (flow and learning)', in *Læring og Ferdighetsutvikling*, H. Sigmundsson, Ed. Tapir akademisk forl., 2008, pp. 119–129.

[6] T. Høien and G. Tønnesen, *Håndbok til ordkjedetesten*. Stavanger: Stiftelsen Dysleksiforsking, 1997.

[7] M. Mossige and A. C. Begnum, 'Ordkjedeprøve som måleinstrument av leseferdigheter', *Psykologi i kommunen*, no. 2, 2018.

[8] *Ikt i trondheimsskolen*, Feb. 2022. [Online]. Available: https://ww08w.trondheim.kommune.no/tema/skole/satsingsomrader/ikt-i-trondheimsskolen/.

[9] E. Kongsnes, 'Slik vil stavanger gjøre det med chromebook i skolen', *Stavanger Aftenblad*, Aug. 2020. [Online]. Available: https://www.aftenbladet.no/lokalt/i/BR5zdE/slik-vil-stavanger-gjoere-det-med-chromebook-i-skolen.

[10] *Pisa 2018: Insights and interpretations, url=https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20In* 2019.

[11] D. L. Mueller, 'Teacher attitudes toward reading.', *The Journal of Reading*, 1973.

[12] G. Amirova, 'Why is reading important', 2019.

[13] A. T. Sharon, 'What do adults read', *Psychometrika*, vol. 1972, 1972.

[14] H. Sigmundsson, M. Haga, G. S. Ofteland and T. Solstad, 'Breaking the reading code: Letter knowledge when children break the reading code the first year in school', *New Ideas in Psychology*, vol. 57, p. 100 756, 2020.

[15] L. C. Ehri, 'Learning to read words: Theory, findings, and issues', *Scientific Studies of reading*, vol. 9, no. 2, pp. 167–188, 2005.

[16] C. Hulme and M. J. Snowling, 'Learning to read: What we know and what we need to understand better', *Child development perspectives*, vol. 7, no. 1, pp. 1–5, 2013.

[17] T. Horowitz-Kraus, R. Schmitz, J. S. Hutton and J. Schumacher, 'How to create a successful reader? milestones in reading development from birth to adolescence', *Acta Paediatrica*, vol. 106, no. 4, pp. 534–544, 2017.

[18] Organisation for Economic Co-operation and Development, *Pisa 2018 results (volume i): What students know and can do - student performance in reading, mathematics and science*, https://www.oecd.org/pisa/publications/PISA2018_CN_NOR.pdf, Accessed on May 13, 2023, 2018.

[19] Connexus, *Lus*, Mar. 2023. [Online]. Available: https://www.conexus.net/produkter/innhold/lus/.

[20] *Hva er lus - leseutviklingsskjema*, 2008. [Online]. Available: https://web.archive.org/web/20080525101904/http://www.utdanningsetaten.oslo.kommune.no/satsingsomrader/norsk_matematikk_naturfag_engelsk_og_2_fremmedsprak/lese_skrive_sprak/lus/.

[21] Utdanningsdirektoratet, Mar. 2023. [Online]. Available: https://www.udir.no/eksamen-og-prover/prover/kartlegging-gs/#a135996.

[22] Utdanningsdirektoratet, *Nasjonale prøver*, Nov. 2022. [Online]. Available: https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/om-nasjonale-prover/.

[23] J. K. Maney, 'Chapter 16 - the role of technology in education: Reality, pitfalls, and potential', in *Handbook of Educational Policy*, ser. Educational Psychology, G. J. Cizek, Ed., San Diego: Academic Press, 1999, pp. 387–415. DOI: https://doi.org/10.1016/B978-012174698-8/50043-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780121746988500436.

[24] L. Stošić, *The importance of educational technology in teaching*, 2015. DOI: https://doi.org/10.23947/2334-8496-2015-3-1-111-114.

[25] Dec. 2022. [Online]. Available: https://www.uis.no/nb/kunnskapssenter-for-utdanning/ressurser/bruk-av-nettbrett-i-barneskolen.

[26] Mar. 2022. [Online]. Available: https://www.uis.no/nb/skole/fordeler-og-ulemper-med-nettbrett-dette-mener-elevene.

[27] A. R. Brown and B. D. Voltz, 'Elements of effective e-learning design', *The International Review of Research in Open and Distributed Learning*, vol. 6, no. 1, 2005. DOI: 10.19173/irrodl.v6i1.217.

[28] B. Faghih, D. M. R. Azadehfar and S. D. Katebi, 'User interface design for e-learning software', *The International Journal of Soft Computing and Software Engineering*, vol. 3, no. 3, pp. 786–794, Mar. 2013. DOI: 10.7321/jscse.v3.n3.119.

[29] H. B. Hutchinson, B. B. Bederson and A. Druin, 'Interface design for children's searching and browsing', *U. of MD HCIL Technical Report*, 2005.

[30] J. C. Read and B. Cassidy, 'Designing textual password systems for children', in *Proceedings of the 11th International Conference on Interaction Design and Children*, 2012, pp. 200–203.

[31] A. R. Hevner, S. T. March, J. Park and S. Ram, 'Design science in information systems research', *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004, ISSN: 02767783. [Online]. Available: http://www.jstor.org/stable/25148625 (visited on 8th May 2023).

[32] K. Peffers, T. Tuunanen, C. E. Gengler *et al.*, 'Design science research process: A model for producing and presenting information systems research', *CoRR*, vol. abs/2006.02763, 2020. arXiv: 2006.02763. [Online]. Available: https://arxiv.org/abs/2006.02763.

[33] T. de Boer, 'From trunk-based to merge requests: A field study at adyen', 2021.

[34] K. Moran, *Usability testing 101*, Dec. 2019. [Online]. Available: https://www.nngroup.com/articles/usability-testing-101/.

[35] K. Whitenton, *Unmoderated user tests: How and why to do them*, Oct. 2019. [Online]. Available: https://www.nngroup.com/articles/unmoderated-usability-testing/.

[36] A. Bangor, P. T. Kortum and J. T. Miller, 'An empirical evaluation of the system usability scale', *Intl. Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

[37] J. Brooke *et al.*, 'Sus-a quick and dirty usability scale', *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[38] J. Sauro and J. R. Lewis, *Quantifying the User Experience: Practical Statistics for User Research*, eng. Burlington: Elsevier Science & Technology, 2012, ISBN: 0123849683.

[39] J. R. Lewis, 'The system usability scale: Past, present, and future', *International Journal of Human–Computer Interaction*, vol. 34, no. 7, pp. 577–590, 2018.

[40] I. Etikan, 'Comparison of convenience sampling and purposive sampling', *American Journal of Theoretical and Applied Statistics*, vol. 5, no. 1, p. 1, 2016. DOI: 10.11648/j.ajtas.20160501.11.

[41] [Online]. Available: https://statistics.laerd.com/spss-tutorials/dependent-t-test-using-spss-statistics.php.

[42] A. King and R. Eckersley, *Wilk test*, 2019. [Online]. Available: https://www.sciencedirect.com/topics/mathematics/wilk-test.

[43] 2023. [Online]. Available: https://libguides.library.kent.edu/spss/pairedsamplesttest.

[44] [Online]. Available: https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php.

[45] 2021. [Online]. Available: https://www.ibm.com/docs/en/SSLVMB_28.0.0/pdf/IBM_SPSS_Statistics_Algorithms.pdf.

[46] P. Schober, C. Boer and L. A. Schwarte, 'Correlation coefficients: Appropriate use and interpretation', *Anesthesia & analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[47] J. C. De Winter, S. D. Gosling and J. Potter, 'Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data.', *Psychological methods*, vol. 21, no. 3, p. 273, 2016.

[48] R. L. Thorndike, *Applied psychometrics*. Houghton Mifflin, 1982.

[49] J. D. Blischak, E. R. Davenport and G. Wilson, 'A quick introduction to version control with git and github', *PLoS computational biology*, vol. 12, no. 1, e1004668, 2016.

[50] M. Maheshwari. 'Introduction to github'. Accessed on May 19, 2023. (Jun. 2022), [Online]. Available: https://www.geeksforgeeks.org/introduction-to-github/.

[51] Docker, *What is a container?*, Accessed on May 19, 2023. [Online]. Available: https://www.docker.com/resources/what-container/.

[52] KtorIO. 'Ktor github repository'. (Mar. 2023), [Online]. Available: https://github.com/ktorio/ktor.

[53] George Batschinski. 'Firebase - a comprehensive guide'. Accessed on May 19, 2023. (n.d.), [Online]. Available: https://blog.back4app.com/firebase/.

[54] PostgreSQL. 'About postgresql'. Accessed on May 19, 2023. (2023), [Online]. Available: https://www.postgresql.org/about/.

[55] ReactJS. 'Reactjs legacy documentation'. Accessed on May 19, 2023. (2023), [Online]. Available: https://legacy.reactjs.org/.

[56] ViteJS. 'Vite github repository'. (Feb. 2023), [Online]. Available: https://github.com/vitejs/vite.

[57] Netlify. 'Netlify'. Accessed on May 19, 2023. (n.d.), [Online]. Available: https://www.netlify.com/.

[58] Microsoft. 'Typescript'. Accessed on May 19, 2023. (2023), [Online]. Available: https://www.typescriptlang.org/.

[59] JetBrains. 'Kotlin github repository'. Accessed on May 19, 2023. (Jan. 2023), [Online]. Available: https://github.com/JetBrains/kotlin.

[60] Tailwind Labs. 'Tailwind css'. Accessed on May 19, 2023. (2023), [Online]. Available: https://tailwindcss.com/.

[61] TanStack. 'Tanstack query v3'. Accessed on May 19, 2023. (n.d.), [Online]. Available: https://tanstack.com/query/v3/.

[62] Microsoft. 'Visual studio code github repository'. Accessed on May 19, 2023. (Sep. 2022), [Online]. Available: https://github.com/microsoft/vscode.

[63] JetBrains. 'Jetbrains intellij idea'. Accessed on May 19, 2023. (2023), [Online]. Available: https://www.jetbrains.com/idea/.

[64] IBM. 'Three-tier architecture'. Accessed on May 21, 2023. (2023), [Online]. Available: https://www.ibm.com/topics/three-tier-architecture.

[65] S. Turney, *Central limit theorem: Formula, definition & examples*, Nov. 2022. [Online]. Available: https://www.scribbr.com/statistics/central-limit-theorem/.

[66] A. Banerjee, U. Chitnis, S. Jadhav, J. Bhawalkar and S. Chaudhury, 'Hypothesis testing, type i and type ii errors', *Industrial Psychiatry Journal*, vol. 18, no. 2, p. 127, 2009. DOI: 10.4103/0972-6748.62274.

[67] L. Faulkner, 'Beyond the five-user assumption: Benefits of increased sample sizes in usability testing', *Behavior Research Methods, Instruments, & Computers*, vol. 35, pp. 379–383, 2003.

Appendix

A LUS

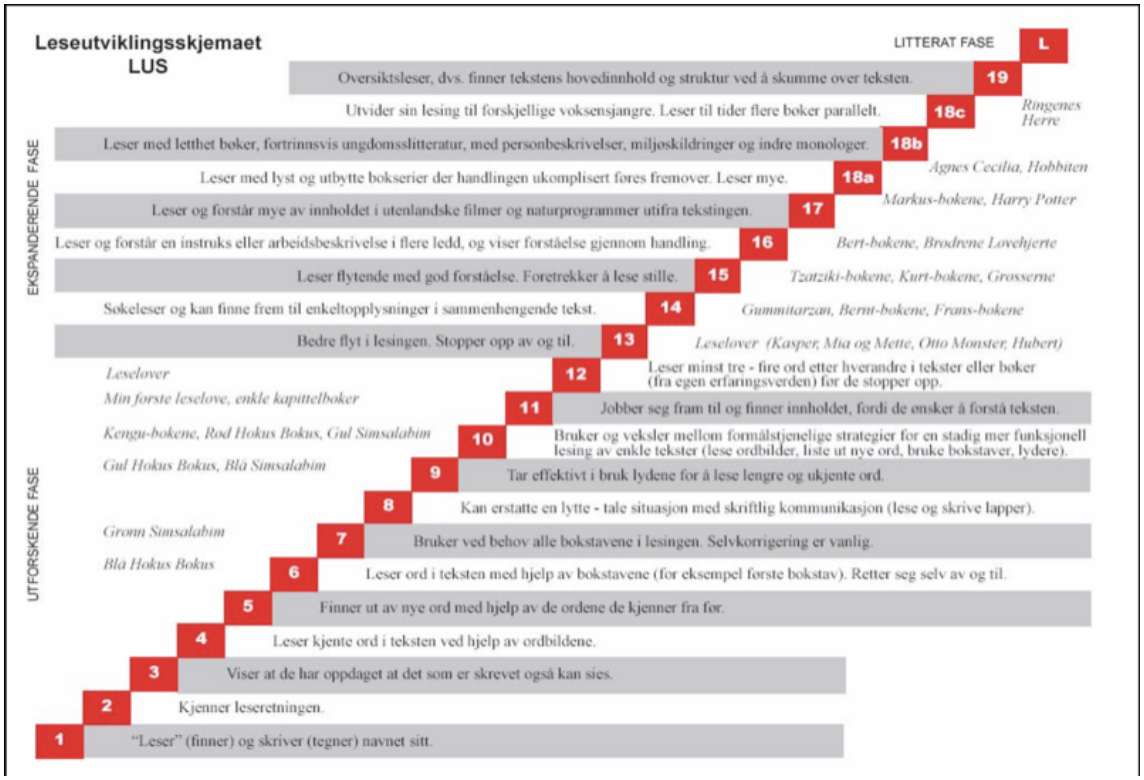


Figure 1: LUS

B Github Repositories

Tester's frontend: https://github.com/eirikolav/reco-2023-tester

Teacher's dashboard: https://github.com/fredrbus/reco-2023-dashbord

Backend: https://github.com/fredrbus/reco-2023-backend

C Feedback from the users

Id	Comment
1	It was difficult to keep track of one's progress and maintain motivation when one did not know how far one had come. This was easier on paper as you got a sense of how many word chains you had completed.
2	A red underline is not needed for misspelled words.
3	I spent a lot of time deleting a misplaced lines because the cursor jumps to the end of the word.
4	I wish I could use the arrow keys to go to the next word. It would be faster that way.
5	Difficult to hit between the letters when in a hurry.
6	The blue line around the words doesn't need to be there.
7	My hand got tired from taking the test.
8	It was annoying that the cursor moved after each click.
9	It might be better to have larger buttons to switch tasks. It takes some extra effort to hit the button as it is now.
10	I found it was a bit clumsy way to separate the words from each other.

Table 1: Feedback from user test after first iteration.

D Test result

First comparability test

Id	Group	Paper-score	PC-score	Sex
1	1	54	57	Female
2	1	90	75	Male
3	1	72	57	Male
4	1	60	53	Male
5	1	69	58	Male
6	1	52	45	Female
7	1	52	48	Male
8	1	51	55	Male
9	1	60	50	Female
10	1	84	51	Female
11	1	59	52	Female
12	1	65	60	Female
13	1	60	57	Female
14	1	67	58	Male
15	1	65	55	Male
16	2	61	52	Female
17	2	55	35	Female
18	2	68	53	Female
19	2	86	47	Female
20	2	73	49	Female
21	2	83	54	Male
22	2	80	59	Male
23	2	86	44	Male
24	2	58	46	Male
25	2	77	50	Male
26	2	61	56	Male
27	2	67	52	Male
28	2	60	46	Female
29	2	59	54	Female
30	2	65	50	Female

Table 2: Results first comparison test

Second comparability test

Id	Group	Paper-score	PC-Score	Sex
1	1	54	68	Female
2	1	90	84	Male
3	1	72	67	Male
7	1	52	46	Male
8	1	51	66	Male
10	1	84	77	Female
13	1	60	62	Female
14	1	67	66	Male
16	2	61	63	Female
17	2	55	58	Female
18	2	68	67	Female
19	2	86	70	Female
21	2	83	72	Male
22	2	80	72	Male
26	2	61	71	Male
28	2	60	68	Female

Table 3: Results second comparison test

E Intructions on how to complete the digital word chain test.

Test preperations

Before completing the test, the teacher or administrator for a class needs to set up a test. The following list describes the steps needed to set up a test.

1. The facilitator navigates to: https://reco-dashboard.netlify.app
2. If the test facilitator does not have an account they set up an account. If they already are registered, they can log in.
3. If the class scheduled for a test does not exist for the teacher, the class needs to be created. Then all the students in the class need to be added to the class. If the class exists, check if all students attending the test are registered. If they are not, add the missing students.
4. When all the students are added to the class, create a test for the students. Set the expiration date to the last date the test is relevant for the class.
5. Print out the user credentials for the class completing the test.

Test completion

1. The pupils navigate to: https://reco-test.netlify.app on their computers.
2. The teacher hands out the user credentials to all the pupils. ¬The teacher specifies that the test is not to be started until further instructions.
3. The pupils log into their accounts.
4. The teacher shows the first word-chain: "musfemrihar" on a display in the classroom. The teacher informs the students how they divide words.
5. The teacher demonstrates how to complete the first example task and goes through the four words "mus—fem—ri—har" and tells the pupils to draw the same line on their sheet.
6. Subsequently, the pupils are instructed to try for themselves on the next tasks and remembered of the number of words and lines required. After some time, the teacher will show the correct solution and goes through common mistakes, such as writing the line after "gå" and not "går".
7. The pupil will try again to complete the next word chain, but without any further instruction. The teacher will show the correct solution after completion.
8. All the pupils will have 30 seconds to test themselves on the three word chains on the practice part of the test. The teacher shows the solution and the pupils control their answers. The teacher reminds the students that the goal of the test is to complete as many word chains as possible in four minutes and three sheets in total with word chains. Then asks if there are any further questions.
9. Reminds them that there are in total 90 word chains and the duration of the test is 4 minutes. 3 Then say: Start
10. View the results at the teacher dashboard.

F Informational letter - Usability Testing

Brukertesting – digital ordkjedeprøve

Hei, og tusen takk for at du vil teste ut applikasjonen vår! Vi vil gjerne begynne med å fortelle litt om oss, prosjektet vårt, og brukertesten.

Hvem er vi?

Vi er to studenter som går femte og siste året på Datateknologi ved NTNU i Trondheim. Høsten 2022 og våren 2023 har vi brukt på å planlegge og utvikle en applikasjon som utforsker muligheten for å bruke datateknologi til å fremme leselæring hos barn.

Hva har vi laget?

Gjennom høsten 2022 og våren 2023 har vår prosjektoppgave og masteroppgave dreid seg rundt digitalisering av leseevalueringstester, og blitt realisert i form av en digital ordkjedetest. Ordkjedetesten, distribuert av forlaget Logometrica, er en papirbasert lesetest der elevene skal dele lange liksom-ord (eks. «havgårdspeilflis») inn i fire mindre ord (ek. «hav—gård—speil—flis») ved å sette blyantstreker på papiret. Vi har valgt å digitalisere denne testen ved å lage et dashbord for lærere og en test-applikasjon for elever. I lærerdashbordet oppretter man klasser, oppretter elever i klassene, og setter opp tester elevene kan ta. Systemet vil generere brukernavn og passord til elevene, som de kan bruke til å logge inn på testapplikasjonen og gjennomføre en digital versjon av ordkjedetesten, der man setter streker ved å trykke med musen. Disse testene vil da rettes automatisk når eleven leverer prøven eller tiden går ut. Når elevene i klassen har tatt en test, vil resultatene bli synlig for læreren i dashbordet. Her vil man få en detaljert oversikt over både per elev og klassevis. Etter hvert som klassen tar flere tester, vil man også få informasjon om fremgangen til hver elev og klassen som helhet gjennom detaljerte grafer og tabeller.

Hva er hensikten med dette prosjektet?

Grunntanken bak paraplyprosjektet vi skriver oppgave under er å bruke datateknologi til å fremme leselæring hos barn. Ideen bak å digitalisere ordkjedetesten er å gjøre slike lesetester mer tilgjengelige, da skolene ikke trenger å skrive ut flere A4-ark til hver elev, og læreren ikke må rette opp mot 7200 ord per skoleklasse på 20 elever. Med dashbordet sikter vi også på å lettere lagre fremgangen til hver enkelt elev,

slik at det er lettere å fange opp de som strever med lesing, eller ikke følger den progresjonskurven de burde.

Hvorfor trenger vi din hjelp?

Vi som datateknologistudenter kan programmere et program, men har ikke god nok innsikt i hva dere som pedagoger og lærere foretrekker og ser nytten av i deres arbeid. Derfor vil vi gjerne ta dere gjennom noen tenkte oppgaver i applikasjonen vår, og se hvor godt programmet tilrettelegger for å utføre disse oppgavene.

Hva vil brukertesten gå ut på?

Brukertesten vil bli gitt i form av en lenke til applikasjonen vår, og et spørsmålsskjema. I dette skjemaet vil vi gi enkle oppgaver med så lite informasjon som mulig, og for hver oppgave du utfører skal du gi en tilbakemelding på hvor enkel/vanskelig oppgaven var, samt eventuell annen tilbakemelding.

Testen vil bestå av to deler. I den første vil du bli bedt om å opprette en egen bruker, og opprette egne data. I andre del vil du få tilgang til en testbruker vi har laget på forhånd, med forhåndslastet data. Kort fortalt kan man si at del 1 tester mest funksjonalitet, mens del 2 demonstrerer/tester mer av nytteverdien vi tror en slik applikasjon vil gi etter bruk over lengre tid. Når alle oppgavene er utført vil vi stille 10 standardiserte spørsmål som evaluerer applikasjonen som helhet etter det som kalles System Usability Scale. Helt til sist vil vi stille noen spørsmål om deg og din erfaring som lærer, hva din subjektive mening om applikasjonen er, potensialet du ser i en applikasjon, og eventuelle andre tanker du måtte ha.

Hva trenger du for å gjennomføre brukertesten?

Alt du trenger er en datamaskin med internettilgang. Operativsystem er ikke så viktig, MacOS, Windows eller Linux går fint. Applikasjonen vil ikke fungere som tiltenkt på tablet (f.eks. iPad) eller mobil.

Greit å vite:

- Applikasjonen er en prototype, laget for bruk på en typisk bærbar PC i et nettleservindu som dekker hele skjermen. Den er ikke laget for, og vil ikke fungere som tenkt, på en skjerm som er smalere enn dette (eksempelvis om nettleservinduet er gjort smalere for å få plass til spørreskjemaet ved siden av).
- Når du oppretter en bruker, vil det du oppgir som e-post adresse, navn og skole være fullt synlig for oss i klartekst på vår database. Denne datainnsamlingen skal gjerne være så anonym som mulig, så du må gjerne bruke en e-post adresse som ikke finnes (men den må være unik), og navnet på deg og skolen din kan gjerne være oppdiktet. Passordet ditt vil ikke være synlig for oss (det krypteres og sendes til en autentiseringstjeneste eid av Google), men kan (på dette tidspunktet) heller ikke endres, så det er viktig at du husker hva slags innloggingsdata du oppgir til de neste stegene av testen.

Til sist vil vi igjen si tusen takk for at du vil teste applikasjonen vår, og om du skulle lure på noe er det bare å ta kontakt med oss på fredrbus@stud.ntnu.no eller eirioa@stud.ntnu.no, eller vår faglærer og veileder John Krogstie på john.krogstie@ntnu.no.

Med takknemlig hilsen, Eirik Olav Aa og Fredrik Busklein

G Usability test questionnaire

Brukertesting - Digital Ordkjedetest

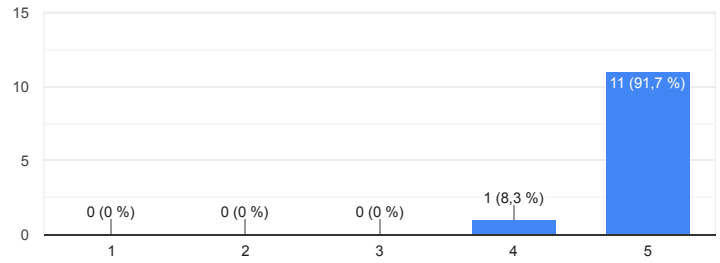
12 svar

Publiser analytics

OPPGAVE 1: Lag en lærer-bruker i lærerdashbordet. Hold gjerne informasjon om e-post, navn og skole anonymt, men husk hva du har oppgitt. Når du har kommet til en side der det står "Velkommen, ...!" er du i mål.

Kopiér

12 svar



OPPGAVE 1: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

2 svar

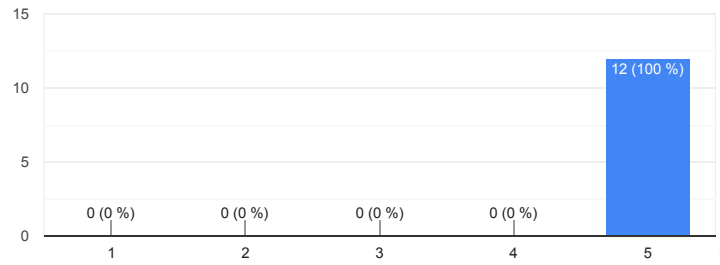
Enkelt. Slik man oppretter bruker på andre liknende sider.

Det var veldig enkelt å utføre oppgaven. Det tok bare 30 sekunder å lage lærer-bruker.

OPPGAVE 2: Logg nå ut av lærerdashbordet.

Kopiér

12 svar



OPPGAVE 2: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

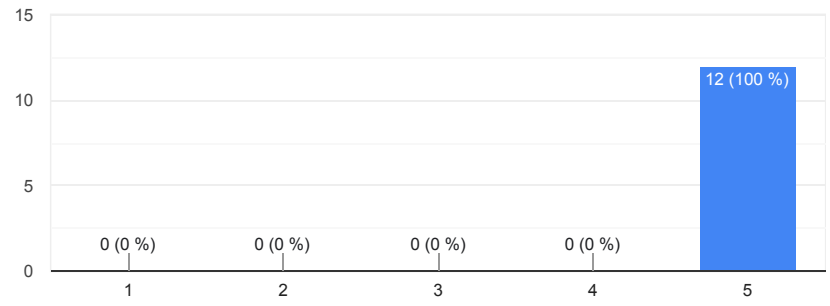
0 svar

Det finnes foreløpig ingen svar på dette spørsmålet.

OPPGAVE 3: Fra forsiden av lærerdashbordet, logg nå inn igjen, med informasjonen du oppga i oppgave 1.

Kopiér

12 svar



OPPGAVE 3: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

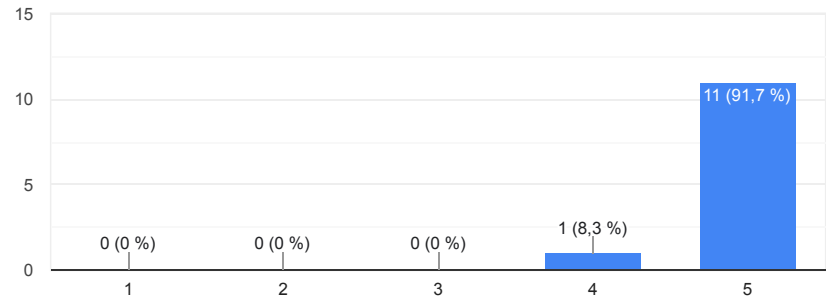
0 svar

Det finnes foreløpig ingen svar på dette spørsmålet.

OPPGAVE 4: Naviger til dine klasser, og opprett en ny klasse. Trinn og parallell er valgfritt.

Kopiér

12 svar



OPPGAVE 4: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

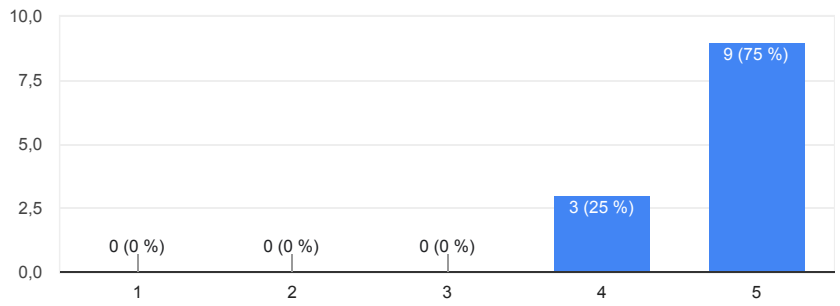
0 svar

Det finnes foreløpig ingen svar på dette spørsmålet.

OPPGAVE 5: I den nye klassen, opprett en ny elev, med et valgfritt navn og brukernavn. OBS: Her kan det hende at en elev med samme navn eksisterer fra før av. Da kan du endre brukernavnet i nederste felt.

Kopiér

12 svar



OPPGAVE 5: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

3 svar

Trykka først på elever så gikk jeg tilbake inn på klassen og såg man kunne oprette elev der.

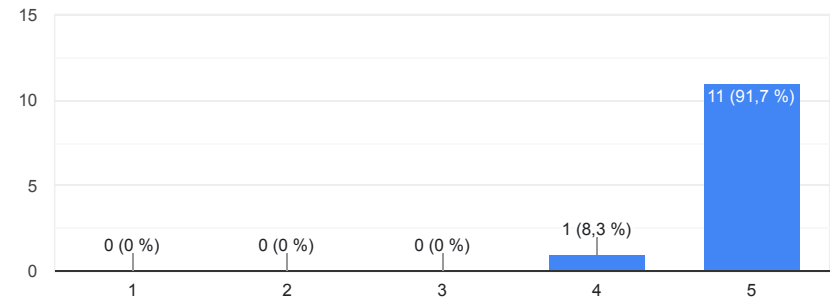
Det er veldig lett å finne ut hvordan man skal opprette en ny elev

Var ikke tydelig at jeg kunne klikke på klassen jeg hadde opprettet. Gikk først i menyen "elever"

OPPGAVE 6: Naviger til testoversikten, og lag en ny test for klassen du nettopp laget. Fristen for denne testen setter du til et vilkårlig tidspunkt i fremtiden.

Kopiér

12 svar



OPPGAVE 6: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

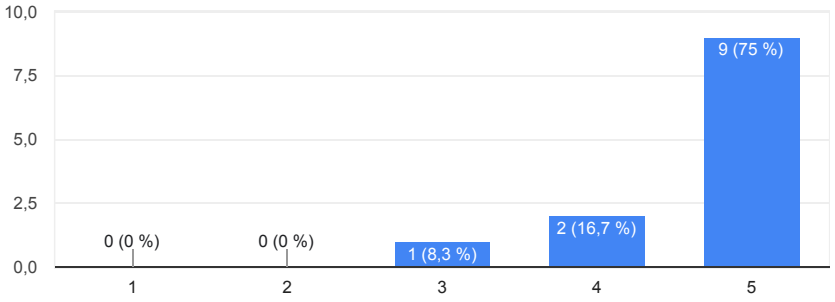
1 svar

Tips kan være å kunne hake av for hele klassen, eller hake av kun for enkeltelever. Jeg som spesialpedagog har ofte kun får elever, og ønsker ikke å dele tester osv. med hele klassen. Må kunne velge om hele klassen eller enkeltelever skal ta testen (hake av for hele klassen, eller trykke og hake av for enkeltelever).

OPPGAVE 7: Naviger til klasseoversikten, og trykk deg inn på klassen du har laget. Under brukeroversikten finner du informasjon om eleven du har laget. Noter deg brukernavn og passord på denne eleven.

Kopiér

12 svar



OPPGAVE 7: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

2 svar

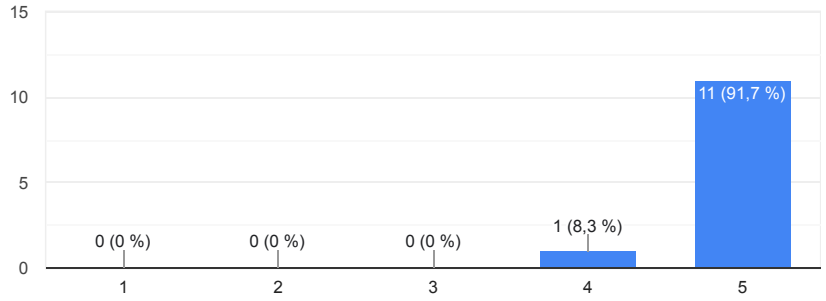
Bruk samme ord som står på siden, brukerinfo - ikke brukeroversikten.

Må man noterer seg brukernavn og passord for alle elever?

OPPGAVE 8: I en ny fane eller nytt vindu, naviger til testapplikasjonen (https://reco-test.netlify.app/) og logg inn med elevbrukeren du noterte deg i forrige steg.

Kopiér

12 svar



OPPGAVE 8: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

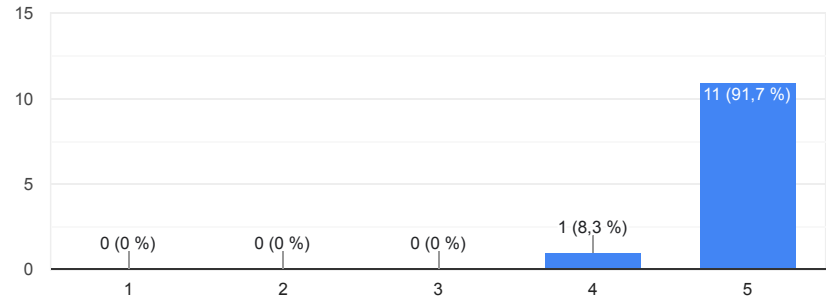
1 svar

Passordet burde bli skjult når man skriver det inn. Eventuelt en egen knapp for å vise passordet.

OPPGAVE 9: Hvis steg 6 gikk som det skulle, skal det nå ligge en test tilgjengelig. Åpne denne testen, og prøv ut øvingsoppgaven på forsiden (trykk mellom bokstavene i testen for å sette streker, og fjern streker med backspace eller delete)

Kopiér

12 svar



OPPGAVE 9: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

2 svar

Programmet er enkelt å intuitivt å navigere i.
Men dere bør virkelig legge inn at læreren kan differensiere antatt ordkjeder. 90 ordkjeder er ALT for mye for de fleste. 10 kan være for mye for noen. Det må kunne bestemmes av læreren hvor mange ordkjeder som skal deles ut og ikke kun være en standard på lengde og oppgaver.

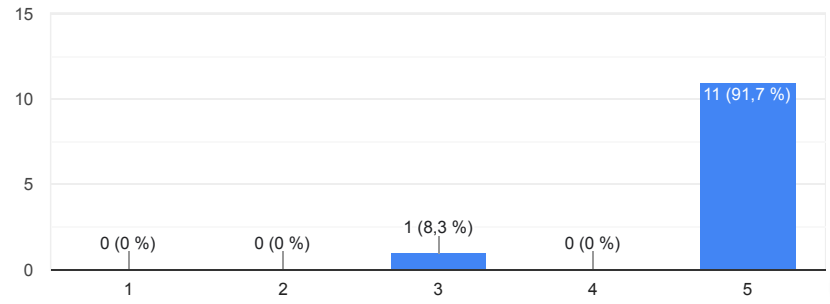
Har kun prøvd på pc, ikke iPad

BRUKERTEST - DEL 2

OPPGAVE 1: Logg inn på *lærerdashbordet (https://reco-dashboard.netlify.app)* med brukerinformasjonen som er gitt over

Kopiér

12 svar



OPPGAVE 1: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

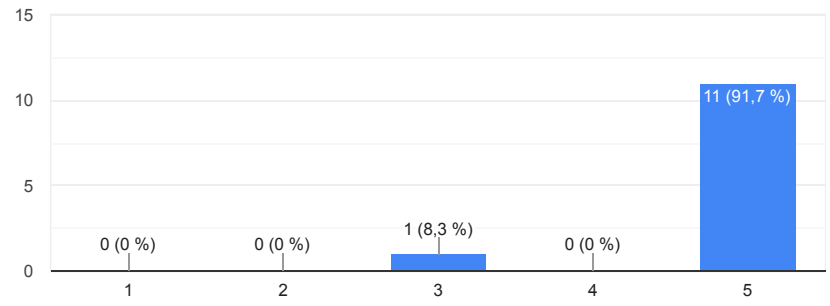
0 svar

Det finnes foreløpig ingen svar på dette spørsmålet.

OPPGAVE 1: Logg inn på *lærerdashbordet (https://reco-dashboard.netlify.app)* med brukerinformasjonen som er gitt over

Kopiér

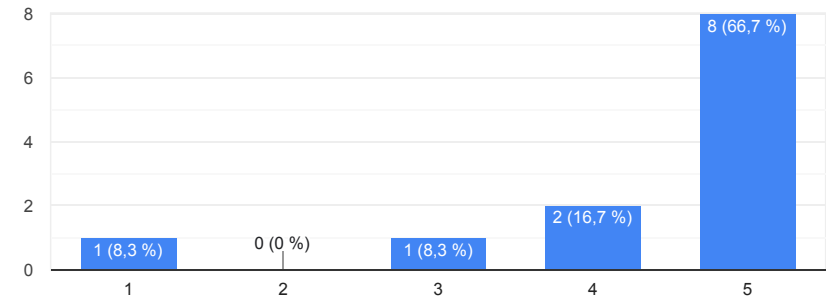
12 svar



OPPGAVE 2: Finn klassen 4B, og print ut informasjonen om elevene og deres passord med command + P for Mac, eller Ctrl + P på Windows PC. Det er ikke nødvendig å printe ut på papir, hvis du får opp en utskriftdialog/vindu kan du trykke "avbryt" og anse oppgaven som ferdig.

Kopiér

12 svar



OPPGAVE 2: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

4 svar

Så ut som at resultatboksene ikke var helt midtstilt på forhåndsvisningen.

For en gangs skyld en oversiktlig og utskriftsvennlig versjon av brukernavn og passord! Men burde vært en "skriv ut" knapp for de som ikke kjenner kommandoen ctrl + p

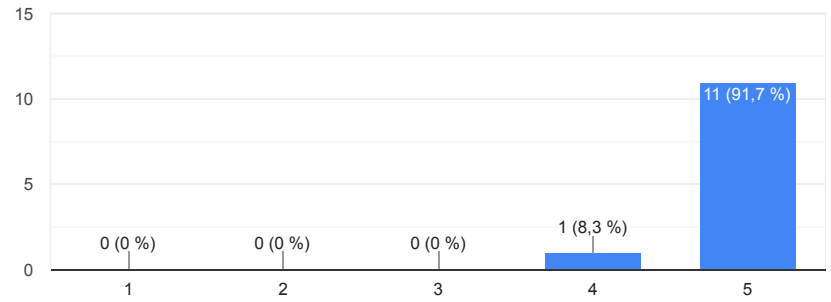
Det burde vært en skriv ut-knapp på skjermen. Ikke alle bruker funksjonstasene på PC'en til dette. Utskriften ble litt rar. Tabellen i oversikten ble liggende utenfor boksen.

Bruker ipad. Fungerte ikke

OPPGAVE 3: I klasse 4B er det en elev som heter Lisa Nordmann. Finn denne eleven, og undersøk om hun har gjort noe fremgang i løpet av det siste året.

Kopiér

12 svar



OPPGAVE 3: Har du noe konkret tilbakemelding du vil legge til om denne oppgaven? (Valgfritt)

0 svar

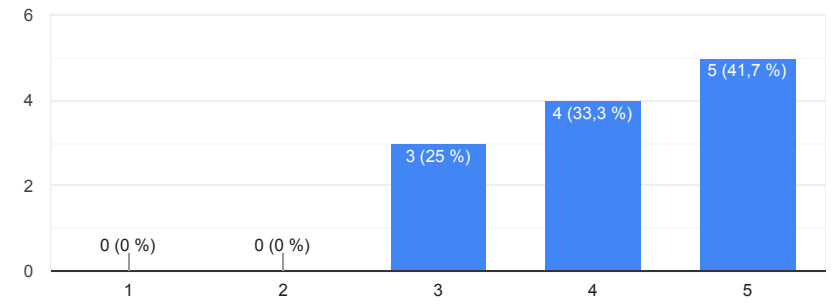
Det finnes foreløpig ingen svar på dette spørsmålet.

System Usability Scale

Jeg kunne tenke meg å bruke dette systemet ofte.

Kopiér

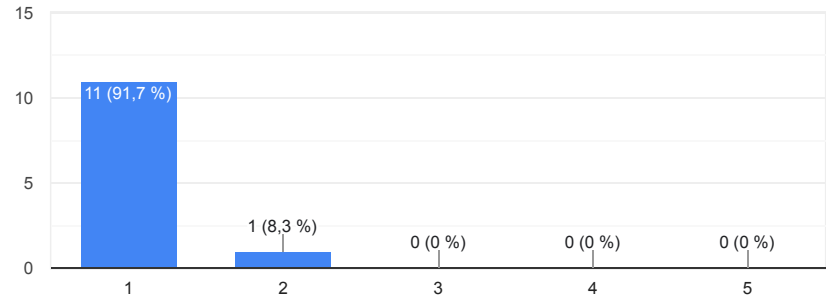
12 svar



Jeg synes systemet var unødvendig komplisert.

Kopiér

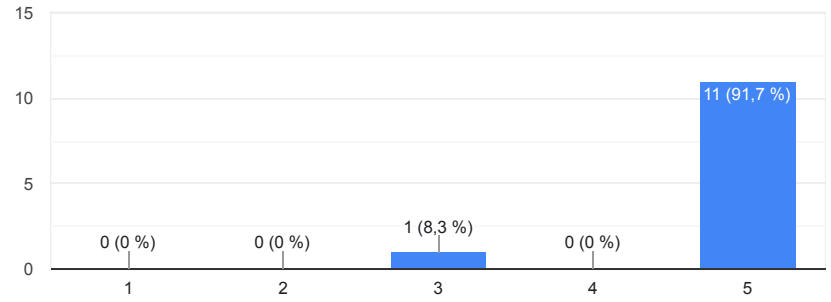
12 svar



Jeg synes systemet var lett å bruke.

Kopiér

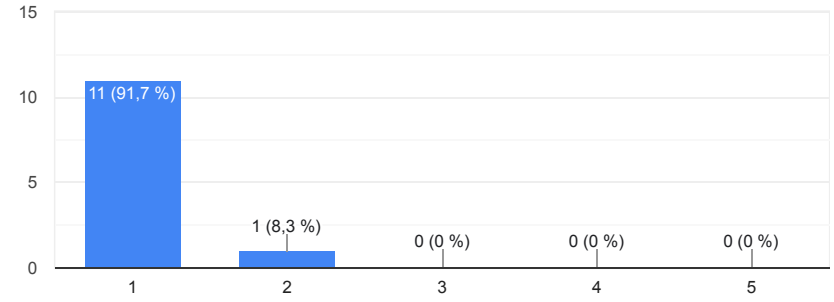
12 svar



Jeg tror jeg vil måtte trenge hjelp fra en person med teknisk kunnskap for å kunne bruke dette systemet.

Kopiér

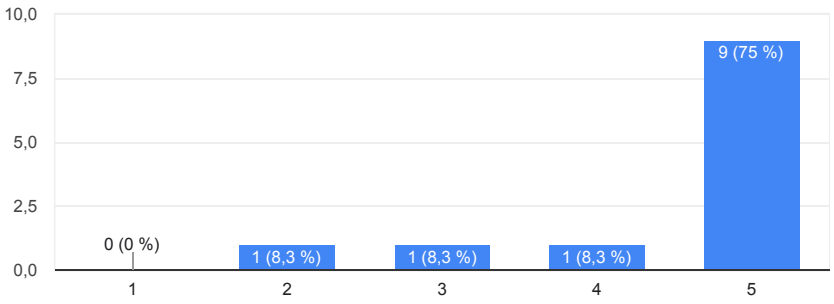
12 svar



Jeg syntes at de forskjellige delene av systemet hang godt sammen.

Kopiér

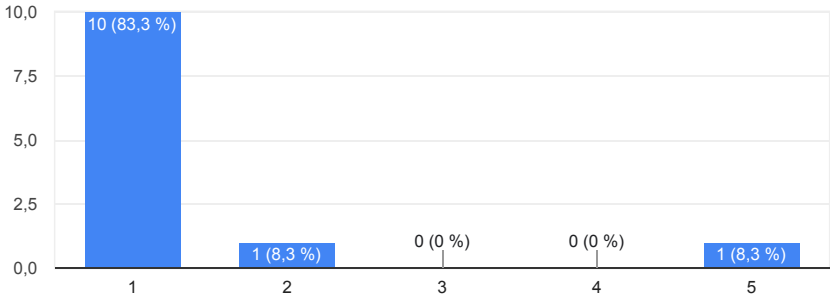
12 svar



Jeg syntes det var for mye inkonsistens i systemet. (Det virket "ulogisk").

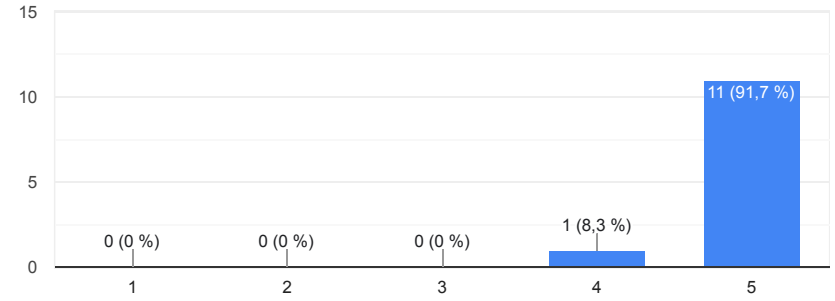
Kopiér

12 svar



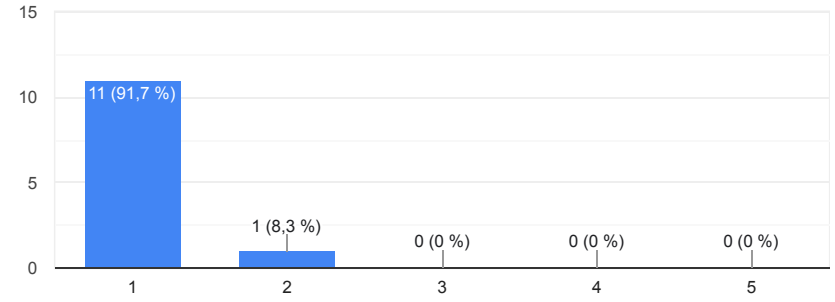
Jeg vil anta at folk flest kan lære seg dette systemet veldig raskt. Kopiér

12 svar



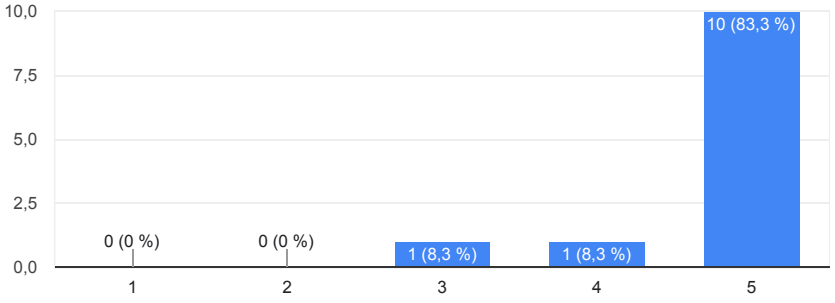
Jeg synes systemet var veldig vanskelig å bruke. Kopiér

12 svar



Jeg følte meg sikker da jeg brukte systemet. Kopiér

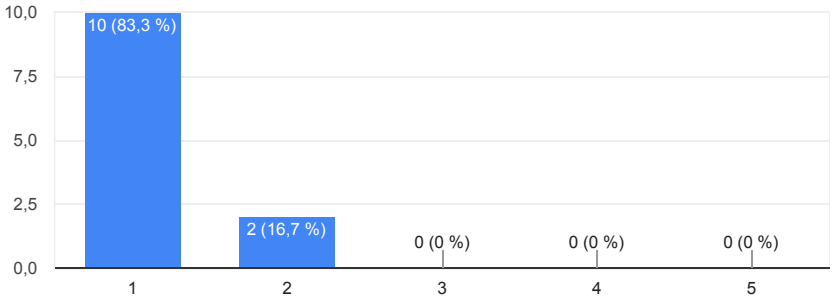
12 svar



Jeg trenger å lære meg mye før jeg kan komme i gang med å bruke dette systemet på egen hånd.

Kopiér

12 svar

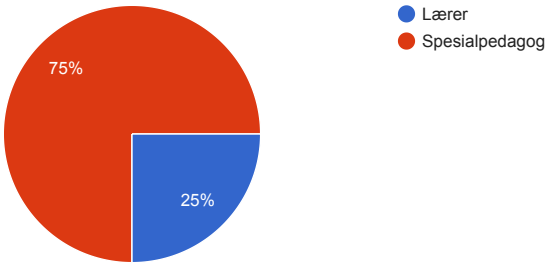


Informasjon om deg

Er du hovedsaklig lærer eller spesialpedagog?

Kopiér

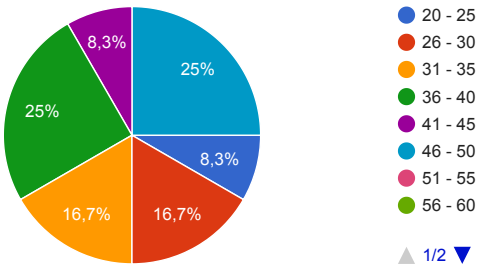
12 svar



Hvor gammel er du?

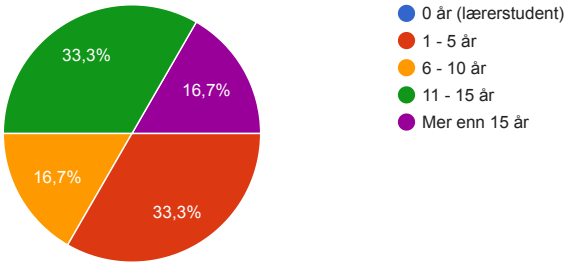
Kopiér

12 svar



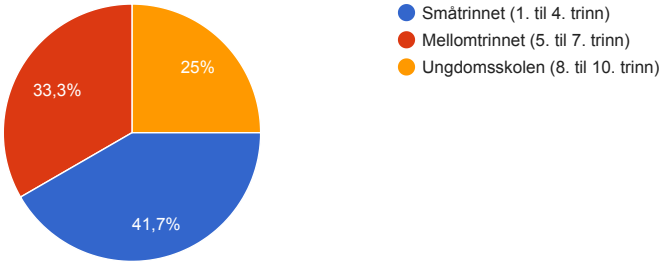
Hvor lenge har du vært lærer/spesialpedagog? Kopiér

12 svar



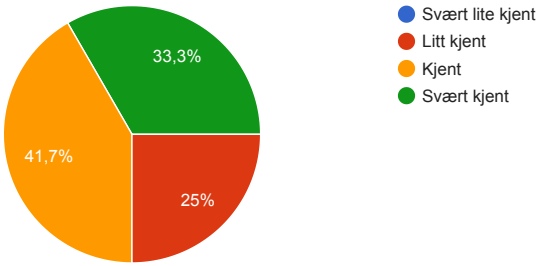
Hvilke skoletrinn er du eller har du vært lærer/spesialpedagog for? Kopiér

12 svar



Hvor godt kjent er du med den papirbaserte (nåværende) ordkjedetesten? Kopiér

12 svar



Dine tanker om prosjektet

Ser du noe umiddelbart forbedringspotensiale ved applikasjonen?

12 svar

Nei

Vil tro at elevene kan bli noe demotivert av å se at det er 100 oppgaver de skal igjennom. Kanskje mulig å dele opp så det virker mindre?

Har lagt inn kommentarer i de første delene av undersøkelsen.

Ikke umiddelbart

nei

Nei

Ønsker meg at resultatene fra testen også sier noe om forventet nivå ut fra alder på elev, og når resultatene er så svake at eleven trenger ekstra oppfølging. Jmf. staminenivå i papirutgaven av ordkjedetesten.

Logge inn med feide og at resultatene legger seg direkte inn i conexus

Utskrift helst en button. Når jeg er i "brukerinfo" burde det også være mulig å klikke seg inn i eleven.

Fjerne streken man setter ved å trykke på den en gang til, ikke måtte trykke backspace

Forslag til å gjøre testen mer forståelig for elevene er å forandre fargen på streken eleven setter mellom ordene. For noen kan streken lett se ut som en stor I eller liten L, og at det derfor kan ødelegge for resultatet. Eksempelvis kan streken være rød slik at elevene enklere ser hvilke ord de har skilt ut. Ved gjennomføring i papirform benytter eleven ofte penn, og dermed ser de enklere hva som er en strek de selv har satt og hva som er bokstaver.

Er det noe du syntes fungerte spesielt bra med applikasjonen?

12 svar

Synes det er bra at kun nødvendig informasjon er gitt, tror en enkel og ryddig utforming av nettsiden gjør det lettere for alle lærere å kunne lære seg og enkelt bruke det.

Brukervennlig og oversiktelig

intuitivt og enkelt å bruke

Oversiktlig og enkel

Virker brukervennlig. Lett å finne frem, ikke for mye valg.

enkelt å bruke

Det var oversiktlig, ukomplisert og enkelt å komme i gang

Enkel å bruke, oversiktlige menyer.

Fint å få det digitalt

Den var oversiktlig. Også for elever.

Enkel og oversiktlig

Fint for elever at de kun trenger å fokusere på én ordkjede av gangen. Dette gjør kanskje at de ikke blir like forstyrret av alle de andre ordkjedene som kommer, slik noen kanskje kan bli ved gjennomføring på papir.

Hva er dine tanker rundt å gjøre slike tester digitalt, fremfor på papir slik det er i dag?

12 svar

Elevene i dag er vandt til å gjøre mye digitalt, og det gjør det mye lettere for lærerne og se mønster og få oversikt over resultatene til elevene. Likevel tror jeg elevene kunne scoret anerledes om testen ble gitt på papir, og at dette derfor bær tas i betrakning når man analyserer resultatene. Å gjøre testene digitalt sparer også unødvendig bruk av papir i tillegg til at jeg tror det vil ta mindre tid, som lærere alltid trenger mer av.

Positivt. Tar mye tid på papir - spesielt å få resultater. Gull å få presentert utviklingen til hver enkelt elev på denne måten og samtidig ha alt lagret i et system.

Flott. Papir flyter, her får man opp resultatene visuelt. Framgang og resultater lagres og er enkelt å finne tilbake til.

Praktisk og tidsbesparende

Kan oppleves vanskelig for de med dårlig teknologisk utstyr, eventuelt når de ikke er vant med å bruke PC. For andre derimot som ikke har god fingermotorikk med blyant/penn, kan dette absolutt være bra.

Jeg tenker at papir er bedre for svake elever

Jeg synes at hele test-prosessen blir mye enklere. Resultatene blir lagret på et trygt sted. Informasjon man trenger er lett tilgjengelig. Applikasjonen gjør det veldig enkelt å følge med på eleven og se elevens progresjon.

Enklere for lærer å gjennomføre en digital oppgave, sparer masse unødvendig kopiering og lagring av papir, sparer rettearbeid. Jeg er positiv til å gjøre tester digitalt. Faren er at man mister noe informasjon, i og med at lærer ikke kan se svarene til elevene og dermed ikke vet noe om hvilke feil som gjort.

Bra

Alltid best digitalt med slike tester.

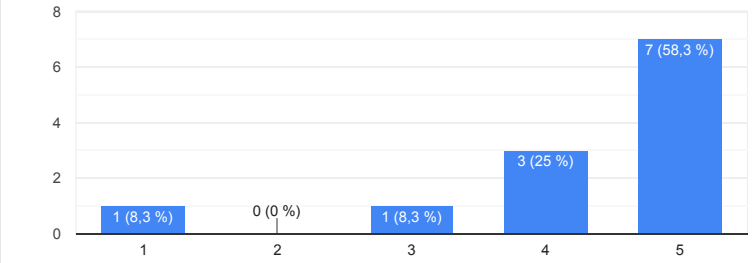
Mye bedre digitalt, spesielt når ordkjedene kommer en og en

Er ganske positiv til det. Digital gjennomføring forenkler rettingsarbeidet og gjør det enklere å se elevens fremgang. Videre er det papirbesparende og det kan også kanskje oppleves enklere for eleven at de bare får se én ordkjede av gangen. Samtidig er det fort gjort at noen elever kan bli forstyrret av det digitale, eller at det er fare for at programmet henger seg opp. Den svarte skillestreken kan også være for lik bokstavene slik at det er vanskeligere for elevene å se ordene de skiller ut. Men om dette endres, er jeg svært positiv til å gjennomføre slike tester digitalt.

Hvor nyttig tror du det er å gjøre en slik test digitalt?

Kopiér

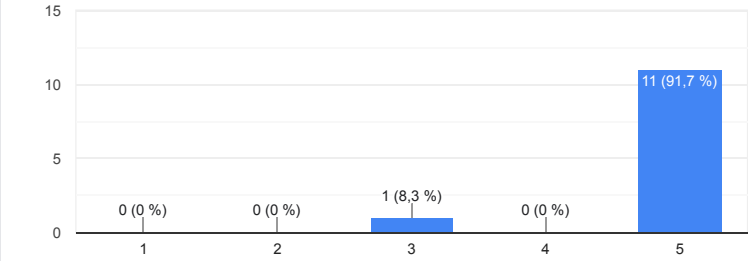
12 svar



Hvor nyttig tror du det er å samle prøveresultatene i et slikt dashbord?

Kopiér

12 svar



Hva er dine tanker om prosjektet som helhet?

12 svar

Nyttig prosjekt som kan komme både elever og lærere til gode :)

Veldig bra at noen har fokus på dette - flere elever man oppdager lese/skriveproblemer sent (eller ikke i det hele tatt) fordi de ikke blir testa. Her er det lett å gjøre det på full klasse - det er ikke vanlig slik det er i dag.

Smart, men trengs å differensiere antall oppgaver (læreren må kunne plukke ut antall oppgaver) og endres på tildeling av oppgaver (dele til klasse eller enkeltelever).

Svært fornuftig med digitale løsninger på kartlegginger/screeninger

Fint å tenke nytt, være med på utviklingen som skjer i samfunnet. Letter nok arbeidet til lærer.

tenker at jeg vil fortsette å bruke papirversjonen

Flo

Veldig positiv, særlig hvis det blir lagt inn stamini--nivå (eller lignende) i programmet.

Veldig bra!

Bra!

Veldig spennende! Får min fulle støtte

Er positiv til prosjektet, så lenge det gjøres nødvendige endringer slik at testen ikke blir vanskeligere for elevene digitalt.

NTNU
Kunnskap for en bedre verden

Department of Computer Science

TDT4501 - Datateknologi, Fordypningsprosjekt

Specialization Project

Authors:
Eirik Olav Aa
Fredrik Busklein

December 2022

Abstract

Contribution: This specialization project investigates the possible advantages of digitalizing word chain-tests used in Norwegian elementary schools and aims to identify key design choices when designing such an application with children as the primary user. These findings will create a basis for our master's thesis, which will focus on implementing and testing the application. **Background**: For the last 25 years, Norwegian schools have used paper-based word chain-tests to assess the decoding skills of young pupils. Although tested and proven as a sound method of reading literacy assessment, paper-based tests require manual grading and storing of results and provide no information on student progress on their own. While Norwegian elementary schools are currently handing out laptops to their pupils, this study aims to improve reading literacy in children by capitalizing on current digitalization and providing teachers with more detailed, accurate, and automatically generated feedback. **Methodology**: To find the most suitable design choices for the application, we used Oates' theory on researching Information Systems and Computing to answer a set of research questions. We will continue this research approach in our master's thesis. **Findings**: Our research found that a digitalized word chain-test could aid in teaching young pupils to read by giving teachers detailed information about their pupils' decoding skills. Teachers can then use this information to adapt challenges to each pupil's reading level for more effective learning. In addition, we found a number of design elements that had to be considered when designing an application for children, which we incorporated into our prototype.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter described the motivation for this study, the description of the project, the three research questions, and a report outline for the project.

1.1 Motivation

The definition of reading literacy has changed over the course of time. Once seen as simply a skill acquired during the first years of school, it is now understood as an ever-expanding set of knowledge, skills, and strategies built upon through interactions with other people in various contexts. In the PISA 2018 Assessment and Analytical Framework, reading literacy is defined as "...understanding, using, evaluating, reflecting on and engaging with texts in order to achieve one's goals, to develop one's knowledge and potential and to participate in society" [1]. Apart from being a requirement for efficiently obtaining knowledge in further education, Sigmundsson et al. also argue that developing reading literacy as a skill helps intellectual, emotional, and social development in children [2].

Reading literacy is an essential skill set on many levels, and according to PISA 2018, 19% of all Norwegian 15-year-olds struggle with reading, compared to 15% in 2015. When comparing boys and girls, we see that 26% of the boys belong to this group versus 12% of the girls [3]. A study on Norwegian 5-6-year-olds concluded that a gap in reading literacy exists between the genders already when children start school [4]. According to Csikszentmihalyi, an approach to reducing this gap is to assess the literacy level of each child and use this assessment to provide them with challenges of an appropriate level, and follow closely on their progress [5].

Since the late 1990s, Norwegian schools have used word-chain tests ("ordkjedetester") to assess reading literacy in Norwegian children from third to tenth grade. In these tests, each pupil is given a list of word chains put together by four words, and their task is to identify the separate words within each word chain by writing a line between them. The goal is to correctly decipher as many word chains as possible within a given time. The test has been normalized twice (1997 and 2007) at schools in Rogaland county to set a benchmark for reading literacy at individual grade levels [6][7]. Currently, the tests are conducted with pen and paper and graded manually by the teacher.

Several Norwegian municipalities have recently started issuing laptops to their pupils [8][9]. Meanwhile, Norwegian schools spend resources on buying test sets from publishers, printing the tests on paper for each pupil, grading each test, and recording test results digitally or by hand. With the introduction of laptops at an early stage in elementary schools, we see the opportunity to capitalize on technological advancements and create a platform that can benefit both teachers and pupils. By digitalizing word chain tests, we hope to reduce the resources spent on assessing the pupils' reading literacy and enhance assessment by providing teachers and pupils with detailed results from each test.

1.1.1 Project description

The scope of our task is to develop an IT tool that can help children develop their reading- and reading comprehension skills. This project will therefore aim to create an e-learning tool for assessing and monitoring pupils' reading literacy from third to tenth grade using word chain tests. By reinforcing the assessment of each pupil's reading literacy, the teachers can better provide each pupil with challenges of an appropriate level and more effectively help them improve their reading skills.

1.2 Research Questions

To best design such a tool for children, we have chosen to focus our research on the following research questions:

Research Question 1: How can digital systems help teachers in gaining a better overview of the reading- and reading comprehension level of individual pupils?

Research Question 2: What design elements should be applied in a user interface with children as the primary users?

Research Question 3: What are the challenges in designing a login-based application for children?

1.3 Report Outline

Chapter 2 - Research Approach: Describes the research questions and research method of this project.

Chapter 3 - Background: Provides theories from relevant areas in different topics related to this project. This include theory on how to use digital tools to optimize learning, designing of user interfaces for children, and login-based applications for children.

Chapter 4 - Users and Use Cases: Classifies and describes the identified users of the app, and gives a couple of examples of typical use cases.

Chapter 5 - Requirements: Specify all the requirements for the system, both functional and non-functional requirements.

Chapter 6 - Prototype: Contains screenshots of the essential application frames of the prototype.

Chapter 7 - Conclusion and future work: Gives a conclusion to the work done in the project and discuss future work.

Chapter 2

Research Approach

The following sections aim to give insight to the research method used in this project, and in the following master thesis, as well as the evaluation method which will be used to evaluate the application.

2.1 Research Method

Oates proposes a method for researching [10], which includes the 6Ps of research. One of these is the *process*. This specialization project follows Oates' theory on the process of researching Information Systems and Computing. Figure 2.1 shows the model of the research process and the different components included in the model. The highlighted boxes indicate the components included in this project.

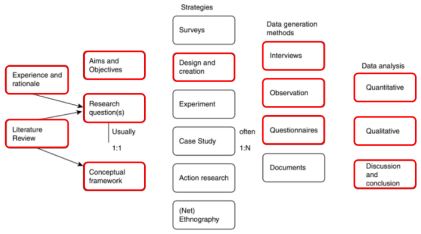


Figure 2.1: Model of research process

The research strategy chosen for this project is the *Design and creation* strategy. This will be the research strategy for both the specialization project and the master thesis. This strategy focuses on developing new IT products or artifacts. In the specialization project, we are creating a digital prototype for the complete system we will produce as the master thesis. The data will be collected through questionnaires with users and through interviews with teachers on how they experienced the app. The prototype will be usability tested at the beginning of the master thesis. The data collected will be both qualitative and quantitative.

2.2 Evaluation

The application will be evaluated qualitatively through interviews and questionnaires with teachers conducting word chain tests with their classes. The questionnaire will also include an evaluation based on the System Usability Scale, a simple tool for measuring usability in a reliable manner [11]. Usability will also be tested on pupils by conducting usability tests based on the functional- and non-functional requirements described in Chapter 5. In addition, we hope to gather enough data through tests to be able to compare the application with the conventional hand-written tests.

Chapter 3

Background

The following chapter will provide some relevant background material to the Specialization Project. This includes theory on how to use digital tools to adapt teaching to the pupils, how to design applications for children, and the login-based challenges for children.

3.1 Use of Digital Tools to Optimize Learning

This section aims to explain how a digital system can be used in order to optimize the teaching for young pupils learning how to read. The section covers the process of learning to read, why supervision is important in teaching, how digitalization can help teachers, and an explanation of an existing reading evaluation test.

3.1.1 Learning to Read Words

Learning to read is one of the most vital parts of early education. Having difficulties or not being able to read can have adverse consequences on a child's future education because a great deal of today's education depends on being able to read [12]. There are two main components that must be mastered to be a competent reader: decoding of words and reading comprehension [6]. The decoding part of reading is the process where the reader links letters to sounds. After the pupil has learned to decode enough letters, they can start to comprehend written words [13]. Being able to read words is one of the most important steps in the process of becoming an adequate reader, as words are the basic units readers use to create meaning out of text [14].

Consequently, learning how to decode complete words is essential for young pupils to be competent readers. Reading involves much more than having efficient decoding skills, but poor decoding skills will be an obstacle in developing proficient reading skills [15]. It is therefore important that each pupil is given sufficient attention and it is made sure that the pupils master the skill of reading words in the first couple of years of education.

3.1.2 Supervision in Teaching

Mihaly Csikszentmihalyi proposes a theory on the term Flow and the importance for pupils to experience flow in learning. There are two conditions that need to be fulfilled to experience flow: perceived challenges that stretch (neither overmatching nor underutilizing) existing skills, clear proximal goals, and immediate feedback about the progress made [16]. Most of the time, the pupils experience school either as an environment where either their skill level is too limited and their challenges are too challenging, which leads to insecurity, or their skills are limited and the challenges are too easy, which leads to boredom and apathy [5].

In the modern classroom, there is a relatively high pupil-teacher ratio in Norway. After the introduction of "Læremormen" in 2019, there are up to 15 pupils per teacher in first to fourth grade [17]. Too many pupils limit the teacher and reduce their ability to adapt their teaching to all the children. Struggling pupils will have a great benefit from extra attention and specialized tuition. An experiment conducted in a seventh-grade class in Trondheim showed that pupils that are far behind the rest performed a lot better and attained increased motivation [18]. When teachers are able to tailor-made tasks for each pupil, both struggling and excelling, the pupils will have a massive educational benefit.

To be able to match each pupil with the right challenges requires a complete understanding of the pupils and their level. With the help of digital tools, this can become an easier task for teachers.

3.1.3 Digitalization

As aforementioned, one of the two components required to experience flow was immediate feedback about the progress. Applying a digital solution in classes makes it possible to give every pupil immediate feedback on their work, as well as display the progress they have made. Digitalization is also beneficial to teachers. By making the grading of tests automated, they will save both time and work. Planning and recording tasks require less space and are more easily accessible [19]. Having test results and pupil data more accessible makes it easier for teachers to utilize the data to gain a better overview of a class and adapt the teaching accordingly.

3.1.4 Word-chain test

One of the tests Norwegian schools have used is the word-chain test. The word-chain test is a screening test that assesses pupils' decoding skills. Some pupils experience great difficulty in acquiring sufficient decoding skills [6]. The word-chain test makes it easier for teachers to detect struggling pupils.

The word-chain test is used in all grades from the second up to the ninth. In total, the test itself takes 4 minutes to complete. In the four minutes, the pupils try to complete as many "word chains" as possible. A word chain consists of 4 different words put together. The length of the words in the test varies from 2 to 7 letters and can be nouns, verbs, adjectives, adverbs, prepositions, or number words. In total, there are 90 chains similar to the examples of word-chains. To complete a word chain, all the words in the chain must be separated by a line; see example of completed chain. Unfinished, wrongly separated, or omitted chains are considered wrong; see examples of unapproved word-chains. The teacher will inform the pupils about the different requirements along with other relevant information in advance of the test.

ordpilvedhvem treoverlivse surminstfriku

Word-chains

ord|pil|ved|hvem

Completed word-chain

ord|pil|vedhvem tre|overliv|se surm|inst|fri|ku

Unapproved word-chains

3.2 Designing User Interfaces for Children

When designing an assessment tool for children, one must consider that young users may lack the cognitive abilities we assume of adults. Having their reading skills evaluated can be stressful enough, so the user interface should put an as little cognitive load on the pupils as possible. In addition, if the pupils have difficulties using the assessment application, the teacher will have to spend extra time helping them, which can stall the whole class. The following paragraphs will describe our findings on creating a learning environment with a low cognitive load by balancing multiple aspects of assessment- and learning environment design.

In *Elements of Effective e-Learning Design*, to keep pupils motivated, Brown and Voltz suggest creating a scenario, a context in which the tasks take place and have meaning [20],. Assuming that young children will have difficulties following text instructions and engaging in a textual scenario, the application can benefit from the extended use of multimedia such as images, icons, and music. Although not directly conflicting, this must be done while keeping the findings from *User Interface Design for E-Learning Software* in mind, which argues that the optimal environment for learning in electronic applications is well organized and eliminates unnecessary distractions, like music and animated figures [21]. The findings in these two articles and the assumptions of a higher need for audiovisual aids suggest that application designers must find a way to use such aids without cluttering the learning environment.

Another aspect of designing interfaces for children is the balance of flat hierarchical structure and low cluttering. While Faghih et al. emphasize the need for a learning environment with as few distractions as possible [21], *Interface design for children's searching and browsing* [22] has a different angle on the cognitive load on children linked to the hierarchical structure of information. While browsing and searching for information in a user interface, adults can easily handle and take advantage of utilities like search bars, filtering, categories, and custom queries. Such utilities allow the designers to minimize clutter and reduce the cognitive load the interface poses on the user by organizing content hierarchically. However, young children lack the cognitive capacity to utilize such features, demanding a different way to search and browse information. A study conducted on elementary-school-aged children by researchers at the University of Maryland in 2005 indicates that a flat, non-hierarchical interface was easier and faster to use in most cases, especially for younger children [22]. However, this also suggests a more cluttered interface, as the application must display more information at the time. This conflicts with the findings of Faghih et al., which state that only relevant information should be displayed [21].

Although not directly conflicting, the findings in these articles discuss aspects of user interface design that must be balanced against each other when designing for children. Considering their assumed lack of reading abilities, children can benefit from using design elements that may negatively affect an older user group. The findings by Hutchinson et al. also show a significant gap in understanding between first -and fifth-graders regarding how to browse and search for information using a graphical user interface [22]. To summarize, there seems to be no correct conclusion to the design problem but rather a set of design decisions that must be considered and weighed against each other in the context of the target user group.

3.3 Login-Based Applications for Children

The application will need an authentication system to ensure that the results recorded during the assessment belong to the correct pupils. Usually, a typical username and password-based login system would be sufficient, but this may not be the optimal solution for young pupils. In their paper *Designing Textual Password Systems for Children*, J. Read and B. Cassidy examine password usage and habits of young children and propose design requirements for password systems aimed at this target group. Two of their main proposals relevant to this application are 1) keeping passwords short and 2) keeping them simple by avoiding the requirement of using both letters and numbers [23]. These proposals conflict with the usual safety-oriented password requirements usually found in applications but can, in turn, significantly increase usability. When designing the authentication system for the application, it is important to consider this trade-off between security and usability.

Chapter 4

Users and Use Cases

This chapter describes the identified users the application is designed for. It will also specify a couple of use cases that explain some typical scenarios of the use of the app.

4.1 Users

The application is designed for the use of three different user groups. The three identified groups are: pupils, teachers and administrators. The groups and their needs are described in Table 4.1.

User Group	Description	Needs
Pupil	Assessment taker. The only user who conducts tests.	• An application that is easy to understand and use. • Instructions on how to complete the assessment. • Immediate feedback based of their results to stay motivated.
Teacher	Supervisor of the assessment. Has overview of the results of the classes and pupils.	• Can initiate new assessments. • Responsible for supervision under the assessments. • Can view the results of pupils and classes, and their progress.
Administrator	Responsible for the system implementation at a given school/grade. Not necessarily a teacher.	• Grant privileges and user credentials to teachers. • View all classes and pupils. • View the different tests, and verify results.

Table 4.1: User Groups

4.2 Use Cases

ID and Name	UC1 - Schedule word chain-test
Users	Teacher
Description	Teacher sets up and schedule a test
Precondition	Teacher is logged in to the system.
Postcondition	Test is scheduled and the pupils have received their login information.
Normal flow	1. Teacher navigates to tests overview, and clicks "create test". 2. Teacher schedules a test for a given class and a given date. 3. On the day of the test, the teacher navigates to the class overview, and clicks "Print usernames and passwords", and prints out the user credentials

Table 4.2: Schedule word chain-test

ID and Name	UC2 - Perform word chain-test
Users	Pupil
Description	Pupil perform and complete a word chain-test
Precondition	The pupil have access to the system.
Postcondition	Each pupil has completed the word chain-test.
Normal flow	1. Each pupil gets a small piece of paper with their user credentials, and uses this to log into the application 2. The pupils log into the system using their credentials 3. The pupils navigate to "tests", and click on the scheduled test. A timer starts. 4. Each pupil performs a task, and moves on to the next one 5. When the timer runs out for a pupil, the test is completed. The system displays a page with feedback on the test. 6. The test is concluded when all pupils have taken the test, or the scheduled date expires

Table 4.3: Perform word chain-test

ID and Name	UC3 - View class result after test
Users	Teacher
Description	Teacher view the board showing the results for the class on a test.
Precondition	Teacher is logged in to the system and at least one test is finished
Postcondition	The system displays the data from the completed test.
Normal flow	1. Teacher navigates to class overview, and clicks on the class that completed the test. 2. Teacher selects the completed test on the left side of the dashboard.

Table 4.4: View class results

ID and Name	UC4 - View pupil's results and progress
Users	Teacher
Description	Teacher uses the pupil overview to monitor the results and progress of a selected pupil.
Precondition	Teacher is logged in to the system, and the pupil has finished one or more tests.
Postcondition	The system displays the pupil's results and progress.
Normal flow	1. Teacher navigates to pupil overview, and finds the pupil by class or search bar. 2. Teacher clicks on the pupil's icon 3. The system displays information about all the test, and progress made by the pupil.
Alternate flow	1. Teacher navigates to a class overview, and finds the pupil in the class list or results table. 2. Teacher clicks on the pupil's icon 3. The system displays information about all the test, and progress made by the pupil.

Table 4.5: View pupil's results

ID and Name	UC5 - View own results
Users	Pupil
Description	Pupil uses system to see own results and progress
Precondition	Pupil is logged in to the system, and has finished one or more tests.
Postcondition	The system displays the pupil's results and progress.
Normal flow	1. Pupil clicks on "My profile" in the navigation bar. 2. Pupil views results and progress on the displayed table and graph.

Table 4.6: Pupil views own results

Chapter 5

Requirements

This section describes the different requirements identified for the system. This includes both functional and non-functional requirements. The requirements are based of the paper based word chain-test

5.1 Functional Requirements

The following functional requirements are listed in descending order based on their priority.

FR1: A pupil should be able to perform a word chain test.

FR2: A teacher should be able to monitor the progress/level of each pupil.

FR3: A teacher should be able to monitor the progress/level of each class.

FR4: A user should be able to log in as their role (teacher or pupil).

FR5: A pupil should be able to view their past results.

FR6: A teacher should be able to create, edit and delete a class or pupil.

FR7: A teacher should be able to create test-sessions.

FR8: A teacher should be able to create/delete user credentials for the pupils.

FR9: The teacher should be able to print out a list of pupils and passwords.

FR10: The system should be able to auto-generate passwords when a pupil profile is created.

FR11: The system should be able to output anonymous test reports as data files.

FR12: The teacher should be able to print out a detailed list of class results.

FR13: A pupil should be able to train on test-tasks to familiarize themselves with the application.

FR14: A system administrator should be able to create/delete user credentials for the teachers.

FR15: The pupils should be able to access a picture-based walk-through.

FR16: Multiple teachers should be able to administer the same class.

5.2 Non Functional Requirements

NFR1: The application should be usable from any modern web browser, and assessment can be done efficiently from any device with keyboard and mouse.

NFR2: Any action from the user (mouse click, keyboard input, etc.) should give immediate feedback to the user.

NFR3: The application should start fast, and be available to the user within 2 seconds after accessing the URL.

NFR4: The application should be easy to use. On the System Usability Scale (SUS), at least 80% of the users should either agree or strongly agree with the statement "I thought the system was easy to use", and either disagree or strongly disagree with the statements "I found the system unnecessarily complex", "I think that I would need the support of a technical person to be able to use this system", and "I found the system very cumbersome to use".

NFR5: The application should be easy to learn. On the System Usability Scale (SUS), at least 80% of the users should either agree or strongly agree with the statement "I would imagine that most people would learn to use this system very quickly", and either disagree or strongly disagree with the statement "I needed to learn a lot of things before I could get going with this system."

NFR6: The application should feel more useful to the users than the current analog tests. On the System Usability Scale (SUS), at least 80% of the users should either agree or strongly agree with the statement "I think that I would like to use this system frequently."

Chapter 6

Prototype

Insight from Chapter 3, use cases from Chapter 4, and requirements from Chapter 5 laid the foundations for the provided prototype. The primary tool used to create the prototype was Figma[1]. In the following chapter, figures 6.1 to 6.17 present screenshots of the essential application frames. However, the online document also provides clickable elements displaying application flow and will be used during further development.

[1] Figma is an online design tool for wireframing and prototyping. *www.figma.com*

Figure 6.1: Teacher login page

Figure 6.2: Pupil login page

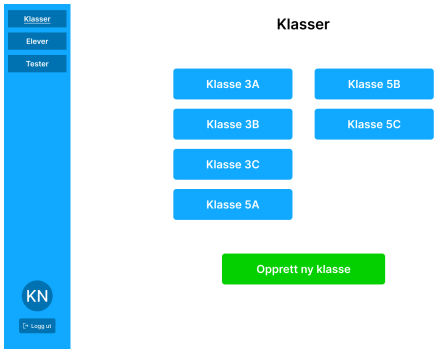


Figure 6.3: Teachers overview of classes

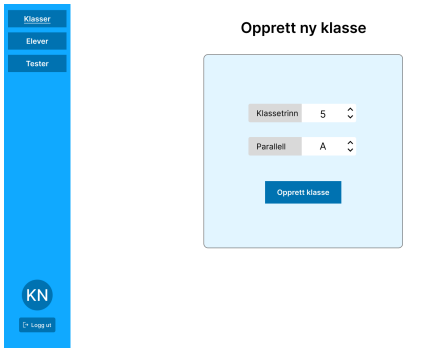


Figure 6.4: Creating a class

Figure 6.5: Teachers overview of pupils, sorted by class with search options

Figure 6.6: Creating a new pupil

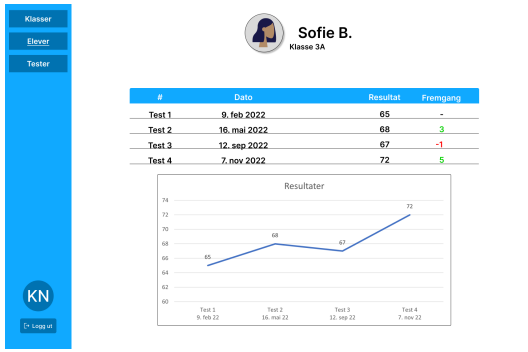


Figure 6.7: Teacher's overview of a selected pupil

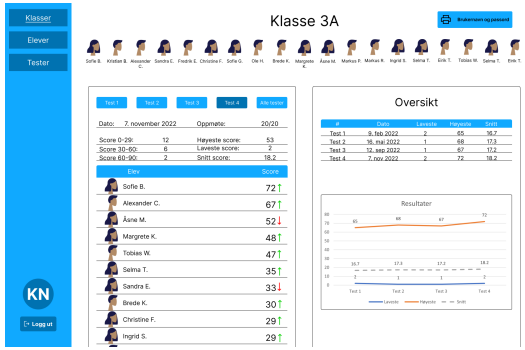


Figure 6.8: Teacher's overview of a selected class and specific test

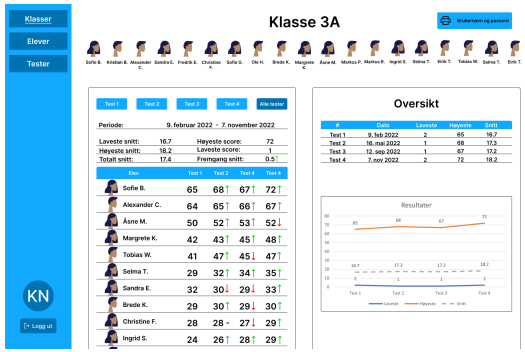


Figure 6.9: Teacher's overview of a selected class, with a summary of all tests

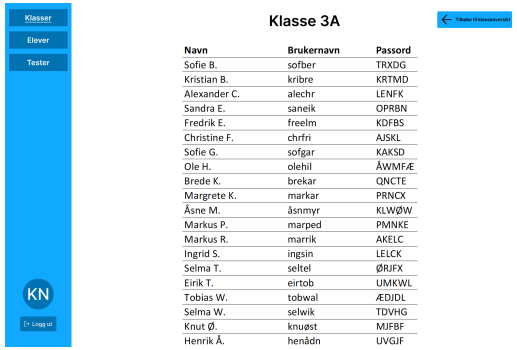


Figure 6.10: Printable page with login details of all pupils in class

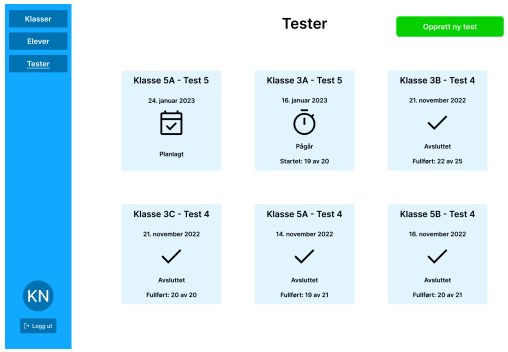


Figure 6.11: Teachers overview of all tests

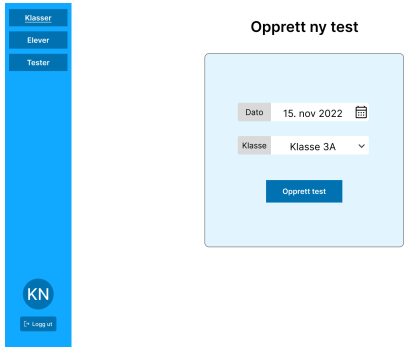


Figure 6.12: Creating a new test

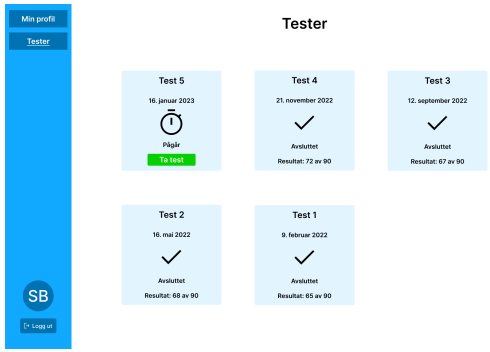


Figure 6.13: Pupil's overview of tests

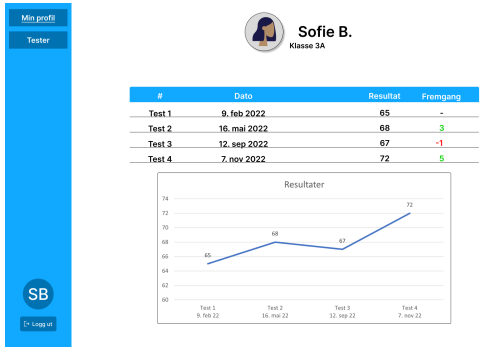


Figure 6.14: Pupil's overview of own results and progress

Figure 6.15: Pupil taking test - before words are separated

Figure 6.16: Pupil taking test - after words are separated

Figure 6.17: Test finished

Chapter 7

Conclusion and Further Work

This section describes the results of the research and study and aims to answer the research questions and explain further work and the following stages of the project.

7.1 Conclusion

Research Question 1: How can digital systems help teachers in gaining a better overview of the reading- and reading comprehension level of individual pupils? When learning to read, it is essential to learn how to decode words. Poor decoding skills will be an obstacle in developing proficient reading skills [14]. It is important to get challenges that stretch existing skills, have clear proximal goals, and immediate feedback about progress when learning [15]. In a class with up to 15 pupils, it can be hard for a teacher to have a complete overview of the pupil's level, and provide them with suitable challenges. By digitalizing the word chain-test, the teacher would get test results and data about the pupils more accessible, making it easier to adapt the teaching. The word chain-test assesses pupils' decoding skills, by having results structured digitally, it will be easier for teachers to recognize pupils struggling with decoding and help them accordingly.

Research Question 2: What design elements should be applied in a user interface with children as the primary users? When designing a user interface with children as the primary users, one must consider and weigh multiple design choices against each other. For example, as Faghih et al. argue, the application should be rid of any unnecessary distractions [21]. In contrast, Brown and Voltz advocate using multimedia elements like sound and animations to engage the pupils better [20]. In the context of this application, which will be an assessment tool, the focus will be on the former argument, and the user interface will be as minimalistic as possible. In addition, keeping the findings of Hutchinson in mind [22], we will also focus on keeping the hierarchical structure of the application flat.

Research Question 3: What are the challenges in designing a login-based application for children? As Read and Cassidy argue in their paper *Designing Textual Password Systems for Children*, passwords for children should be kept short and simple [23]. These suggestions sacrifice the security of the application for enhanced usability. A tradeoff like this is more than welcome in this application, as we identify loss of usability and potential stalling of class progress as a more significant risk than unauthorized access to a pupil's account. In our prototype, we have taken this principle further by autogenerating the pupils' passwords in lengths of 5 letters and storing them in plaintext.

7.2 Further Work

The literature study and the prototype will be the foundation of the digital system that will be developed, tested, and evaluated. The main part of the master thesis will be to implement the prototype created in this specialization project, and all the functional requirements defined in Section 5.1. After the implementation, the system will be tested by both elementary school pupils and teachers. The system will be evaluated both by observations, but also through questionnaires and interviews with the users.

Bibliography

[1] OECD, *PISA 2018 Assessment and Analytical Framework*. 2019, p. 308. DOI: https://doi.org/https://doi.org/10.1787/b25efab8-en. [Online]. Available: https://www.oecd-ilibrary.org/content/publication/b25efab8-en.

[2] H. Sigmundsson, A. Dybfest Eriksen, G. S. Ofteland and M. Haga, 'Gender gaps in letter-sound knowledge persist across the first school year', *Frontiers in psychology*, vol. 9, p. 301, 2018.

[3] F. Jensen, A. Pettersen, T. S. Frønes *et al.*, 'Pisa 2018', *Norske elevers kompetanse i lesing, matematikk og naturfag. Oslo: Universitetsforlaget*, 2019.

[4] H. Sigmundsson, A. D. Eriksen, G. S. Ofteland and M. Haga, 'Letter-sound knowledge: Exploring gender differences in children when they start school regarding knowledge of large letters, small letters, sound large letters, and sound small letters', *Frontiers in Psychology*, vol. 8, 2017, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01539. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01539.

[5] M. Csikszentmihalyi and H. Sigmundsson, 'Flyt og læring (flow and learning)', in *Læring og Ferdighetsutvikling*. Tapir akademisk forl., 2008, pp. 119–129.

[6] T. Høien and G. Tønnesen, *Håndbok til ordkjedetesten*. Stavanger: Stiftelsen Dysleksiforsking, 1997.

[7] M. Mossige and A. C. Begnum, 'Ordkjedeprøve som måleinstrument av leseferdigheter', *Psykologi i kommunen*, no. 2, 2018.

[8] *Ikt i trondheimsskolen*, Feb. 2022. [Online]. Available: https://ww08w.trondheim.kommune.no/tema/skole/satsingsomrader/ikt-i-trondheimsskolen/.

[9] E. Kongsnes, 'Slik vil stavanger gjøre det med chromebook i skolen', *Stavanger Aftenblad*, Aug. 2020. [Online]. Available: https://www.aftenbladet.no/lokalt/i/BR5zdE/slik-vil-stavanger-gjoere-det-med-chromebook-i-skolen.

[10] B. J. Oates, M. Griffiths and R. McLean, *Researching Information Systems and Computing*. Sage, 2022.

[11] J. Brooke *et al.*, 'Sus-a quick and dirty usability scale', *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[12] C. Hulme and M. J. Snowling, 'Children's reading comprehension difficulties: Nature, causes, and treatments', *Current Directions in Psychological Science*, vol. 20, no. 3, pp. 139–142, 2011. DOI: 10.1177/0963721411408673. eprint: https://doi.org/10.1177/0963721411408673. [Online]. Available: https://doi.org/10.1177/0963721411408673.

[13] H. Sigmundsson, M. Haga, G. S. Ofteland and T. Solstad, 'Breaking the reading code: Letter knowledge when children break the reading code the first year in school', *New Ideas in Psychology*, vol. 57, p. 100 756, 2020.

[14] L. C. Ehri, 'Learning to read words: Theory, findings, and issues', *Scientific Studies of reading*, vol. 9, no. 2, pp. 167–188, 2005.

[15] C. Hulme and M. J. Snowling, 'Learning to read: What we know and what we need to understand better', *Child development perspectives*, vol. 7, no. 1, pp. 1–5, 2013.

[16] J. Nakamura and M. Csikszentmihalyi, 'The concept of flow', in *Flow and the foundations of positive psychology*, Springer, 2014, pp. 239–263.

[17] E. T. Rognerød, 'Lærernormen har ført til bedre lærertetthet', *Utdanningsforbundet*, Sep. 2021. [Online]. Available: https://www.utdanningsforbundet.no/nyheter/2021/minstenormen-har-fort-til-bedre-larertetthet/.

[18] M. Csikszentmihalyi and H. Sigmundsson, 'Flyt og læring (flow and learning)', in *Læring og Ferdighetsutvikling*. Tapir akademisk forl., 2008, pp. 69–79.

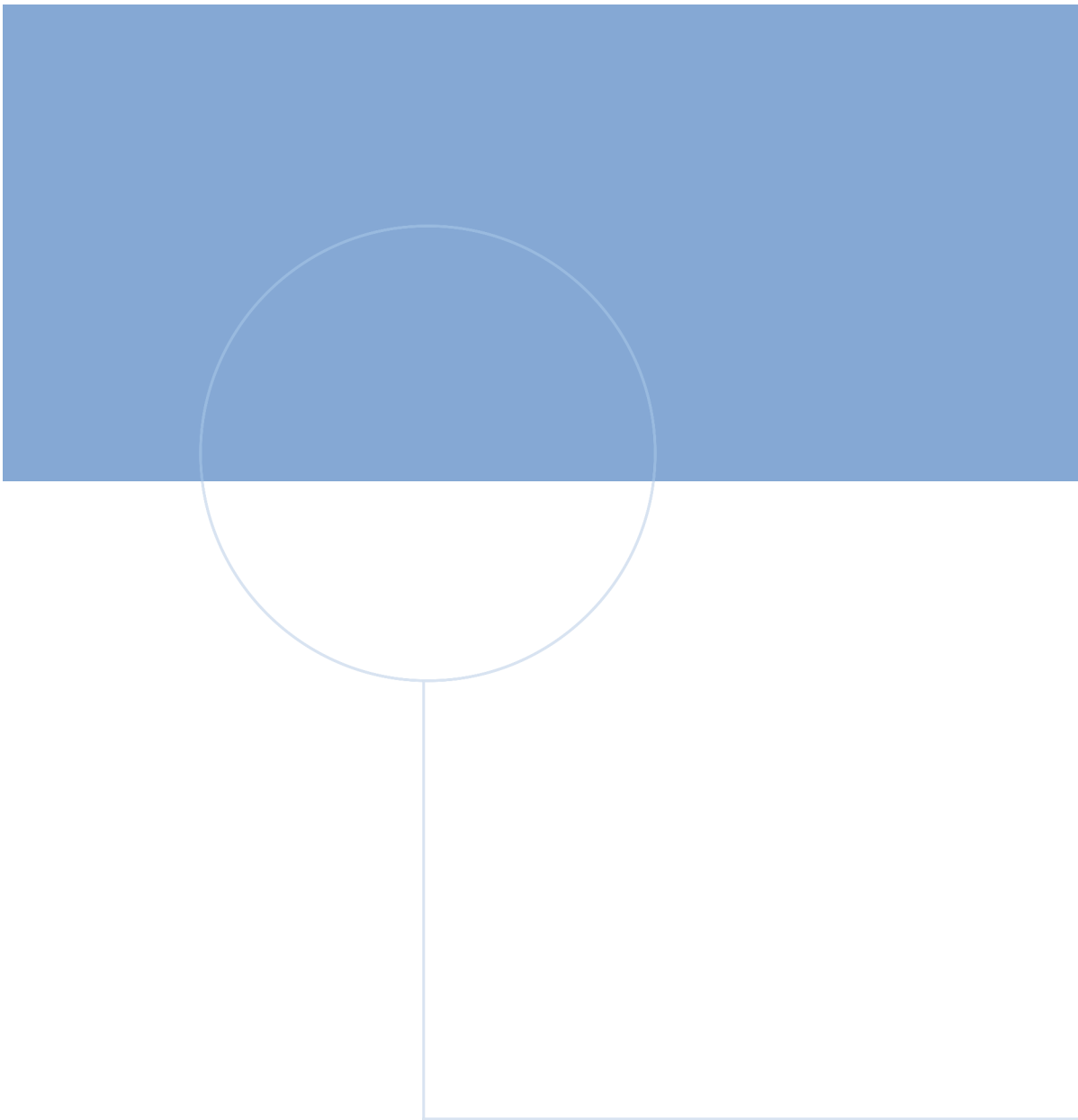
[19] S. D. Hicks, 'Technology in today's classroom: Are you a tech-savvy teacher?', *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, vol. 84, no. 5, pp. 188–191, 2011.

[20] A. R. Brown and B. D. Voltz, 'Elements of effective e-learning design', *The International Review of Research in Open and Distributed Learning*, vol. 6, no. 1, 2005. DOI: 10.19173/irrodl.v6i1.217.

[21] B. Faghih, D. M. R. Azadehfar and S. D. Katebi, 'User interface design for e-learning software', *The International Journal of Soft Computing and Software Engineering*, vol. 3, no. 3, pp. 786–794, Mar. 2013. DOI: 10.7321/jscse.v3.n3.119.

[22] H. B. Hutchinson, B. B. Bederson and A. Druin, 'Interface design for children's searching and browsing', *U. of MD HCIL Technical Report*, 2005.

[23] J. C. Read and B. Cassidy, 'Designing textual password systems for children', in *Proceedings of the 11th International Conference on Interaction Design and Children*, 2012, pp. 200–203.



NTNU

Norwegian University of
Science and Technology