

Amalie Eline Henni
Sylvi Phuong Huynh

Supporting Digital Assessment with Side-by-Side Comparison

Master's thesis in Computer Science
Supervisor: Trond Aalberg
June 2022

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



Norwegian University of
Science and Technology

Amalie Eline Henni
Sylvi Phuong Huynh

Supporting Digital Assessment with Side-by-Side Comparison

Master's thesis in Computer Science
Supervisor: Trond Aalberg
June 2022

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

Abstract

In higher education, there is an increasing workload for markers assessing exams. A growing number of students are enrolled in higher education, assessments must be finished within a time limit, some subjects already have a shortage of markers, and the assessment process is time-consuming. The evaluations are expected to be precise and fair, ensuring they reflect the students' true knowledge. These factors can create a high-pressure environment. In addition, multiple stakeholders depend on having evaluations as a part of the education, meaning that assessments are inevitable. Thus, it is desirable to support the assessment process by making it more efficient while remaining consistent. This thesis explores alternative assessment methods to support efficiency and consistency in high-stake, written exams on digital platforms in higher education. The concluding concept is the Side-by-side comparison, which displays multiple exam answers simultaneously. The answers are from different students on the same task in the exam. Then, the marker can easily compare the answers, establish a baseline for the expected knowledge level, and conduct the assessments.

To evaluate the concept and gain domain knowledge, interviews with markers in higher education were conducted. They revealed a need and room for support within digital assessment tools. Multiple interviewees had made their own systems to improve the assessment process. Regarding the Side-by-side comparison, the interviewees thought it could increase consistency, but it depended on other factors. These factors include preparation, subject level, and which answers are displayed together.

The findings from the evaluation indicated a need for further research into supporting the assessment process. This thesis's results can be used as a starting point for further exploration into markers' assessment processes and inspire expansions into alternative assessment methods.

Sammendrag

I høyere utdanning øker arbeidsmengden for sensorer som retter eksamener. Det blir flere studenter som tar høyere utdanning, evalueringer må bli gjennomført innen en gitt frist, noen fag har mangel på sensorer og retteprosessen er tidkrevende i seg selv. Evalueringene forventes å være presise og av høy kvalitet, for å sikre at de gjenspeiler studentenes kunnskapsnivå. Disse faktorene skaper mye press for sensorene. I tillegg er flere interessenter avhengig av å ha evalueringer som en del av utdanningen, som betyr at vurderinger er uunngåelige. Det er derfor ønskelig å effektivisere vurderingsprosessen gitt at den forblir konsistent. Denne masteroppgaven utforsker alternative vurderingsmetoder som kan støtte effektiviteten og konsistensen i avgjørende, skriftlige eksamener på digitale plattformer i høyere utdanning. Hovedkonseptet er Side-by-side comparison (side-ved-side-sammenligning) som viser flere eksamensbesvarelser samtidig. Besvarelsene er fra forskjellige studenter, men tilhører samme oppgave. Dette gjør at sensorer kan enkelt sammenligne besvarelsene, raskt etablere en baseline for forventet kunnskapsnivå, og gjennomføre vurderingene.

For å evaluere konseptet og få domenekunnskap ble det gjennomført intervjuer med sensorer i høyere utdanning. Intervjuene avdekket at det er behov og rom for støtte innen digitale vurderingsverktøy. Flere av intervjuobjektene hadde laget egne systemer for å forbedre vurderingsprosessen. Når det gjelder Side-by-side comparison, mente intervjuobjektene at konseptet kunne øke konsistensen, men at det var avhengig av andre faktorer. Disse faktorene inkluderer forberedelse, hvilke svar som vises sammen og type og nivå på faget.

Funnene fra evalueringen indikerte et behov for ytterligere forskning for å forbedre vurderingsprosesser. Resultatene kan brukes som et utgangspunkt for videre utforskning av sensorers vurderingsprosesser og inspirere til utvikling av alternative vurderingsmetoder.

Preface

This Master's thesis completes a five-year Master of Science degree in Computer Science at the Norwegian University of Science and Technology (NTNU). The research was done under the supervision of Trond Aalberg, whose expertise and excellent guidance have been invaluable. Special thanks to Trond for all the relevant advice, interesting discussions, and feedback throughout our work.

Additionally, we are truly grateful to everyone who participated in the interviews. Thank you for sharing insightful experiences, suggestions and perspectives. Also, we would like to thank the respondents of the usability tests for valuable feedback during the development of the web application.

Amalie Eline Henni and Sylvi Phuong Huynh

Trondheim, 9th June 2022

Contents

1. Introduction	1
1.1. Motivation	1
1.2. Research Questions	2
1.3. Methodology	2
1.4. The Foundation of the Thesis	3
1.5. Results	3
1.6. Outline	4
2. Background Theory	5
2.1. Assessment and Evaluation in Education	5
2.1.1. Types of Assessments	5
2.1.2. The Evaluation Process	5
2.1.3. Principles in Assessments	7
2.1.4. Referencing and Grading in Evaluation	7
2.1.5. Bias in Grading	9
2.1.6. Consistency in Assessments	11
2.1.7. Efficiency in the Assessment Processes	12
2.2. Digital Assessment Tools	13
2.3. Related Work	15
2.3.1. Computer-Assisted Grading Rubrics	15
2.3.2. Feedback System on Programming Assignments	16
2.3.3. Automated Essay Scoring Systems	17
3. Design	19
3.1. Methodology	20
3.1.1. Design and Creation	20
3.1.2. Experimental Research	22
3.2. Concepts	22
3.2.1. Side-by-side Comparison	23
3.2.2. The Task-by-task Assessment Method	24
3.2.3. Grouping	25
3.2.4. Consistency Check	27
4. Implementation	32
4.1. Method of Implementation	32

Contents

4.2. Technologies	33
4.3. Processing the Data Set	35
4.4. Enabling the Concept	38
4.5. Components	42
4.6. The Web Application	49
4.6.1. Initialising the Assessment	51
4.6.2. Assessing	52
4.6.3. Finalising the Assessment	56
4.7. User Testing	58
4.8. Limitations and Workarounds	64
5. Evaluation	65
5.1. Interview Preparations and Processing	65
5.1.1. Method	65
5.1.2. Preperation	66
5.1.3. Setup	67
5.1.4. Interview Questions and Walkthrough	67
5.1.5. Processing the Interviews	70
5.2. Interview Results and Discussion	70
5.2.1. Side-by-side Comparison	71
5.2.2. Comparison of Candidates	73
5.2.3. Task-by-task Assessment Method	74
5.2.4. Consistency in Assessments	75
5.2.5. Grouping	78
5.2.6. Setting Scores on Answers	80
5.2.7. Bias in Assessments	81
5.2.8. Efficiency in the Assessment Process	83
5.2.9. System Flow	84
5.2.10. Future Ideas	86
5.2.11. Summary of the Interview Results	87
5.3. Limitations in the Study	88
6. Conclusion and Future Work	90
6.1. Findings for the Research Questions	90
6.2. Future Work	93
Bibliography	93
A. Interview guide	97
B. Walkthrough guide	100
C. Application for NSD	102
D. Consent form	105

List of Figures

2.1. Inspera Assessment Navigation Menu	14
2.2. Inspera Assessment Conflict Detection	14
3.1. A Typical Evaluation Process	19
3.2. An Evaluation Process Using Side-by-side Comparison	19
3.3. The main concept and its' support concepts.	23
4.1. Inspera's Data Structure Simplified	36
4.2. Data Structure of Data Saved in the Web Application	37
4.3. Flow of the Concept	38
4.4. Overview Section: Task Page	40
4.5. Overview Section: Assessment Page	40
4.6. Overview Section: Approval Page	41
4.7. Overview Section: Completion Page	41
4.8. Task List Component	42
4.9. Grade Component	43
4.10. Answer Text Box Component	44
4.11. Approval Text Box Component	45
4.12. Candidate Text Box Component	45
4.13. Grade Text Box Components	46
4.14. Consistency Check Component	47
4.15. Process Diagram of the Web Application	50
4.16. Routing Between Pages in the Web Application	51
4.17. The Task Page	52
4.18. The Assessment Page	53
4.19. The Approval Page	55
4.20. Stripped JSON Data for Inconsistent Assessments	56
4.21. The Completion Page	57
4.22. The Assessment Page in the First Iteration of User Tests	60
4.23. The Task Page in the Second Iteration of User Tests	61
4.24. The Assessment Page in the Second Iteration of User Tests	62

1. Introduction

1.1. Motivation

Standard evaluation methods at educational institutions are tests, quizzes, essays and more. These methods are typically used as assessments before the students encounter a final exam at the end of the semester. A marker creates the exam, the students take the exam, and then it is delivered to the marker for assessment. At last, the students will receive a final score, letter grade, pass or fail on the exam within a given time limit. This evaluation summarises the extent to which the students have achieved their learning goals.

One can notice a need for assessment support and that it can be valuable for markers to have a more efficient assessment process. When looking at the process, there are many challenges to face. Firstly, the number of exams to be assessed can exceed 1000¹ due to the many students enrolled. Secondly, the markers have limited time to assess the exams due to a legislated deadline. Additionally, a common evaluation process is to assess one candidate at a time. Then, the marker must constantly adjust to different tasks and evaluation criteria, and the context switching may decrease performance. Lastly, there are expectations that markers can ensure fairness and consistency in the evaluations. Humans are subjective by nature, so how does one know that the given assessment is correct and that personal opinions do not cloud their judgement? The main principles in assessments are for them to be fair, reliable, accurate and consistent. An action to ensure this is double marking, which is a common practice at many educational institutions. Even though it is helpful and reassuring to assess submissions twice, it is still an extra resource that requires much time and labour.

Statistics show that an increasing amount of people are choosing higher education, which consequently will result in more exams and evaluations². Based on the growing number of students and the challenges mentioned above, one can see a need to support the marker's evaluation process or get more markers. If not, this can result in overworked markers and biased assessments. The abovementioned factors indicate a need for a more efficient

¹<https://www.universitetsavisa.no/campus/1100-studenter-gikk-opp-i-dette-faget-ingen-strok/109918>

²<https://www.ielts-practice.org/band-8-5-ielts-essay-sample-there-is-a-sharp-increase-in-the-number-of-people-studying-at-university/>

1. Introduction

assessment process, and it should not go at the expense of consistency. With global digitisation, new possibilities emerge. A suggestion in this thesis is the Side-by-side comparison where multiple answers from an exam are displayed simultaneously to be assessed task by task using a digital platform.

1.2. Research Questions

This thesis explores how one can support efficiency and consistency in the evaluation process by using the Side-by-side comparison. This evaluation method is the main concept of this thesis. It will be further explored and delved into with its' supplementary concepts.

To explore this, one should understand the domain, what challenges are attached, design the concept, and test it with end-users. The following research questions have been formed to guide the work.

- **RQ1:** What are the challenges encountered when assessing?
- **RQ2:** How can assessment support be designed to support efficiency and consistency?
- **RQ3:** What are the perceptions of using Side-by-side comparison?

1.3. Methodology

The research methodology has been inspired by Design and Creation and Experimental research from [Oates \(2006\)](#). In short, the primary focus of the methodology was on conducting a preliminary study of the domain, designing a suggestion, evaluating the design and discussing the findings.

The preliminary studies helped with understanding concerns in the assessment process. It gave a foundation and uncovered the need for more research in this field. Then, a concept design that could explore the assessment processes took place. It resulted in the Side-by-side comparison. Other elements were also designed to support the main concept and ensure its quality. To evaluate the concept, there was an incentive to create a software prototype, a web application, that could showcase how the Side-by-side comparison could be used in practice. Then the web application was presented in interviews. Six interviews were held to get thoughts and perceptions on the Side-by-side comparison and the overall concept. At last, the discoveries and findings were analysed and discussed to uncover if the design suggestion accommodates the research questions.

1.4. The Foundation of the Thesis

The foundation of this master's thesis comes from the specialisation project carried out during Autumn 2021 at the Norwegian University of Science and Technology (NTNU). During this project, some research within the field of education, assessment and learning technology was conducted. Some of these findings are revised and updated. The core contribution of the specialisation project was the preliminary design of a side by side evaluation method.

1.5. Results

This thesis explores how the digital assessment process of written exams in higher education can be supported to become more efficient and consistent. Concerning RQ1, a detailed preliminary study was conducted to gain valuable knowledge within the field and discover what markers find challenging with the assessment process. To address RQ2, the main concept was developed. It is an alternative method where markers are displayed multiple answers from different students on the same task simultaneously when assessing exams. It is called Side-by-side comparison. The concept is implemented as a single-page web application and was constructed to see the feasibility and enable a more thorough evaluation. RQ3 was investigated by conducting interviews with markers in the target group, where they were asked about their experience, thoughts on the concept, and insight into the web application's assessment flow. The interview subjects had various experiences with the digital assessment process of exams in higher education, and they had different opinions regarding the concept and the elements complementing it. Several respondents saw using Side-by-side comparison as an advantage because it could ensure that similar submissions would get the same score, resulting in fairness and consistency. However, some feared that the Side-by-side comparison could encourage internal ranking, which they considered unfair. Although there were no unanimous answers, all seemed intrigued by the focus of this thesis and were optimistic regarding further exploration of the assessment process.

1.6. Outline

- In [chapter 2](#), the background for the project is discussed, enlightening RQ1. It involves theory regarding assessment, bias, different scoring methods and other work conducted by researchers to support efficiency in grading processes.
- In [chapter 3](#), RQ2 is contemplated, and a suggestion is presented. The design, design process, and concepts are explained.
- In [chapter 4](#), the implemented product for this thesis is described. It concerns the approach for development, a description of the data set, the technologies being used, and a thorough walkthrough of the implemented web application. Additionally, it concerns the user testing conducted.
- In [chapter 5](#), RQ3 is addressed and the process for evaluating the concept is described. This involves the method, preparation, setup, interview questions, and after work regarding the interviews. Then the results from the interviews are presented and discussed. Lastly, this chapter discusses improvements and limitations in the study.
- In [chapter 6](#), the research questions are addressed, and future work to explore the concepts is described.

2. Background Theory

2.1. Assessment and Evaluation in Education

In education, one must consider many aspects and metrics when performing evaluations. It includes externals' requirements and demands, students' expectations, and personal factors. It is thus a demanding process, which requires much consideration and effort, and any support could be valuable. Today, society is characterised by more digital solutions, and the field of assessments and education is keeping up. Previously, most evaluation processes took place on paper, but with the digital technology shift, new opportunities arise.

2.1.1. Types of Assessments

There are two assessment types practised in education. One type is formative assessments which are low-stake, in-process evaluations or informal evaluations designed to assist the learning process by providing feedback to the learner¹. The other assessment type is summative assessments. They are high-stake evaluations and usually result in final decision-making when completing courses or curricula. Typically these assessments occur at the end of educational activities. The score, pass or fail, or a letter grade reflects to what extent the curricular goals have been achieved during the course and represent the conclusions of the marker's judgments². In addition, summative assessments communicate the learners' abilities to external stakeholders.

2.1.2. The Evaluation Process

In educational settings, evaluation determines students' knowledge levels in subjects. It is a systematic method to study how well the students have achieved the learning objectives³.

¹<https://www.edglossary.org/formative-assessment/>

²<https://www.edglossary.org/summative-assessment/>

³<https://www.ruralhealthinfo.org/toolkits/rural-toolkit/4/evaluation-importance>

2. Background Theory

In some cases, the markers are a part of the whole evaluation process, from preparing the exams to giving the students their feedback or score. For others, the evaluation process starts when the students have performed the exams and the exams are sent to assessment. There exist different guides, practices and routines for the evaluation process. Some institutions might have specific and own practices.

Firstly, the exams need to be planned and created. One has to decide on the appropriate type of exam, such as multiple-choice questions, completion tasks, written or oral exams, longer essays, or short text answers. An exam can consist of one or multiple types, and the type needs to allow the candidates to show their competence. Additionally, it must fit the assessment method the marker will use later in the evaluation process. After that, one must align the exam to the learning outcomes and clarify which objectives the exam addresses. Then it is time to create questions, and these should reflect the learning objectives and have clear instruction. When all of this is done, one should ask a colleague to read through the exam to make sure that everything is unambiguous and clear before letting the students perform the exam⁴.

After the exam has been conducted, the assessment of the exams begins. When performing assessments, there are different assessment methods for markers to apply. One can, for instance, evaluate submissions with a candidate by candidate or a task by task method. The latter does not add value for essay exams, but it can when applied for exams with several tasks. When evaluating candidate by candidate, markers can give a holistic scoring where grades are based on the overall quality and the submissions are assessed as a whole⁵. This is contrary to the task by task method, which views tasks of the submissions individually. It is unclear what method is more common as it is most likely reliant on one's preferences. The Inspera Assessment assessment support described further in [section 2.2](#) offers the possibility to customise this process oneself.

The assessment often has to be done within a specific time limit after the exams have been conducted. The marker has to evaluate whether the students have fulfilled the criteria. In the evaluation process, double marking, which is described in [subsection 2.1.6](#) is widely used and accepted as a part of the process. When all marking is finished and administrative procedures at the institution are in place, the grade is given back to the students. In some cases, students may ask for the reasoning behind their grades or send a complaint that needs to be handled.

⁴<https://www.cmu.edu/teaching/assessment/assesslearning/creatingexams.html>

⁵<https://methods.sagepub.com/reference/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i10438.xml>

2. Background Theory

2.1.3. Principles in Assessments

In educational assessment settings, principles one seeks to accommodate are fair and equal evaluations. They should be unbiased evaluations that consider the content of a submission. It includes that all students should be evaluated equally based on the same criteria and solely on the students' performance and not other factors⁶.

As assessments are usually high-stake evaluations, one desires to achieve the abovementioned principles. Having defined criteria increases the probability of fair and equal evaluations because it will guide markers while assessing. All of the students will be evaluated with the same criteria, meaning that they have the same opportunity to achieve specific scores or grades. This is given that markers actively follow and use the criteria during assessments. By doing so, markers will be able to ensure fairness regardless of when the students are assessed. If they happen to be the first submission, the last or after a long day, the students should be rewarded for the same things.

One of the reasons to have assessments is to drive students learning. Therefore, one should consider students' perceptions of assessments as they are intended for them. Findings show that students hold strong views about fairness in assessments, especially regarding how students are evaluated and their opportunities to demonstrate their knowledge of a subject. For instance, alternative assessments such as portfolios, oral presentations, peer assessments and more (Sambell et al., 1997) are perceived as fair by students as it rewards those who constantly make an effort to learn rather than those who pull an all-nighter or a last-minute effort (Sambell et al., 1997). In comparison, traditional final essay exams are considered more unfair, primarily when conducted within one day, because they can seem to rely on the daily form and luck. The lack of control over the evaluation process and feedback also contributes to the students' perceptions of traditional assessment as unfair. Hence, having equal opportunities and fair assessments are essential principles in assessments.

2.1.4. Referencing and Grading in Evaluation

When performing assessments, one needs to consider referencing and grading. According to McAlpine (2002), referencing is the basis of the judgment, and grading involves comparing the students' performance with a predefined set of criteria. There are three common referencing methods: norm-related, criterion, and ipsative. For grading, norm referenced grading and mastery learning are the most common types in use (McAlpine, 2002). Within grading, there are also scoring methods to assess writing where holistic scoring, analytical scoring and primary trait scoring are some to mention.

⁶<https://www.unl.edu/gradstudies/connections/grading-fairly-and-efficiently>

2. Background Theory

Referencing

Norm-related referencing is when the judgement is based on comparing individuals with their peers. It is typically useful for selective purposes. For classic norm referencing, a test is delivered to a group of representatives that one wishes to assess, and then norms are developed based on these results. These norms are then used to grade other groups. Cohort referencing is similar to norm referencing, but it takes the sub-groups of the assessed students as its' baseline. That means attaining a high grade can depend on one's performance and the peers who took the assessment simultaneously.

Criterion referencing is a comparison of the students' achievements with predefined criteria. When applying criterion referencing, the students must know that the requirements for success are based on their abilities to achieve the learning objectives rather than their performance against other peers.

Ipsative referencing is performed when the student is compared against themselves. The judgment compares the student's performance against their previous results. This allows the students to see their progress or whether or not they are taking advantage of prior feedback⁷.

Grading

As mentioned, norm referenced grading is when the student's performance is compared to others who have 'set the base'. Regarding cohort referenced grading, one knows that most educational institutions practise some form of it (McAlpine, 2002). Mastery learning is when the student's performance is compared with learning objectives. It is usually applied with criterion referencing because, within grading based on mastery learning, the student has to accomplish several criteria. The mark is either pass or fail as one has mastered the area or not (McAlpine, 2002).

Both of the grading types are improper for courses at higher educational institutions today. Although courses are often based on learning criteria and fit the mastery learning grading approach, this is rarely suitable for summative assessment. Final grading is usually of letter grades and seldom pass or fail. Hence, there are attempts to combine norm referenced grading and mastery learning. This is attempted by establishing levels of competence within criteria or by aggregation of criteria (McAlpine, 2002).

There exist methods for assessing writing. Each of these methods has its strengths and weaknesses. Holistic scoring is a method for evaluating a composition based on the overall quality. The marker considers the quality and decides on a single holistic score (Frey, 2018). Holistic scoring is often used in large-scale assessments as it is time and

⁷<https://www.questionmark.com/what-is-ipsative-assessment-and-why-would-i-use-it/>

2. Background Theory

cost-effective. It is because the markers do not provide any detailed comments or feedback on the student's work. The analytical scoring assigns separate scores for each different aspect of writing. It could be, for instance, content, organisation, vocabulary, grammar, and mechanics (Frey, 2018). This method is suitable for small-scale assessments with a smaller group. Analytical scoring is more time-consuming than holistic scoring as it provides detailed feedback. In the last scoring method, primary trait scoring, the evaluation is based on the performance on one or more specific aspects essential for the success of the task in the question. It is less time consuming than holistic and analytical scoring. This is due to only one score being assigned to the intended criteria for scoring.

2.1.5. Bias in Grading

The definition of bias is "The action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment."⁸. Concerning evaluation, there exist several biases that might affect the grading (Birkelund, 2015). Bias in this context is any factor that influences an assessment or a grade which is not the performance itself. It can be experienced as critical and unfair.

Types and Factors

Bias can occur both consciously (Malouff, 2008) and unconsciously (Malouff et al., 2014), and there are many possible types and factors for bias in grading. Some types are influenced by the personal well-being of the marker, and others are affected by external influences. For the latter, this can be factors regarding the student. Markers may assign higher grades to students based on any presumptions they have about the students. Nisbett (1977) experimented with evidence of bias in evaluations based on earlier impressions of the person being evaluated. This could be an impression of whether the students have performed well in class earlier, are interested, hardworking, eager in class, and speak and act like well-behaved people. Other factors could be looks, gender and ethnicity. Bernard (1979) did research that discovered that there are teachers with gender bias in oral exams. Any relationship towards the students, family-friend, child of a friend, and so may also lead to bias. These examples could typically lead to positive bias, but some factors could also lead to more negative bias. One might assign a lower grade if the student is not well-behaved or show less interest in class (Malouff, 2008).

Regarding bias concerning the markers' well-being and mental state, one must consider that markers are humans and that they are vulnerable to human needs like sleep, food, and general fatigue. Therefore they are not set to be 100% objective and unbiased like a computer could be. Danziger et al. (2011) discovered that in court decisions, "The

⁸<https://dictionary.cambridge.org/dictionary/english/bias>

2. Background Theory

percentage of favourable rulings drops gradually from 65% to nearly zero within each decision session and returns abruptly to 65% after a break.". An essential principle in jurisprudence is equality before the law, and many countries has it written in their constitution. [Danziger et al. \(2011\)](#)'s experiment shows that it might not be the case in real life settings. If jury members, who should be objective organs that value fair and just rulings, are vulnerable to human needs and bias, then it is not outrageous to draw parallels to markers in grading.

Markers' general fatigue could be caused by a huge workload or an inefficient assessment system that affects their mental state. Another cause of fatigue could be the evaluation process itself. As [Quinn \(1975\)](#) states, "The most efficient predictor of monotony and boredom was the number of times during a fixed period that the most often-repeated task was performed.". This can be seen in the light of markers assessing numbers of submissions, which can be considered as repetitive work. Repetitive work can cause boredom which might lead to fatigue and bias. Other causes of boredom could be the work environment, personal reasons and circumstances ([Fisherl, 1993](#)). According to [Klein \(2002\)](#), scoring procedures by sixty teachers indicated impairment of judgement, apparently resulting from mental fatigue. In addition, fatigue may influence the evaluation because markers make faster decisions, resulting in unfair assessments.

Another factor concerning the evaluation process is context switching which means changing tasks or focus. Context switching requires energy. Too much of it can be tiring and may lead to falling out of context, leading to a less efficient workflow. Multitasking, which regards the same issue, is a technique which is discussed whether it is profitable to do or if it decreases the total amount of work being done. According to [Dux et al. \(2006\)](#), one cannot execute two tasks at once, and switching tasks is also a bottleneck in the progress. [Katidioti et al. \(2014\)](#) state that the pupil dilation of a person who is about to switch tasks comes up several seconds before the actual task switching. This indicates that the decision to switch is costly and that the switch itself costs, but it is less than the actual decision to switch. One can draw parallels between these findings and the evaluation process in education. To change tasks to assess is costly. On the other hand, tasks in an exam are often related, meaning that the difference and change between two tasks may not be as prominent as in other settings.

Other than fatigue, the markers' background and knowledge of the subject might affect their evaluation. This is due to markers earning more experience for each submission they evaluate. Evaluations may therefore vary according to when they were performed. Another factor is that assessments might be affected by earlier assessments. [Kenrick and Gutierrez \(1980\)](#) did an experiment where they found out that the performance of the previous candidate might affect the grading of the following candidates. This is also supported by [Birkelund \(2015\)](#) as he states that "The markers may be viewing an average exam as less than average when grading it right after an outstanding exam.". This contrast effect might therefore contribute to bias in grading.

2. Background Theory

Measures to Prevent Bias

As the grading of exams are high-stake evaluations, it is important to find ways to reduce or uncover bias. Even though bias can result in positive consequences for certain students, it is not desirable because it can be seen as unfair and unreliable. Since mental fatigue can cause bias, one option is to look at ways to reduce mental fatigue. One can, for instance, make sure to take enough breaks, get a change of scenery, ensure healthy food and get enough rest and sleep. On the other hand, several of these measures require time and may reduce efficiency. A common measure is to have anonymous evaluations of exams. Archer and McCarthy (1988) recommends using this whenever it is possible. One way is to assign students a code number so the marker does not know the name of the students' submissions. Another measure is to have double marking described in subsection 2.1.6 to check for consistency and unveil any possible bias. Additionally, Nisbett and Ross (1983) did a study which showed that trying hard to be objective does not eliminate bias. This is not surprising as markers are humans and will naturally be influenced by factors even though one tries to avoid them. Hence, self-awareness is not sufficient to prevent bias and measures are needed.

2.1.6. Consistency in Assessments

One has to consider several factors when evaluating the quality of any work, and one of them is consistency. Consistency in assessment is essential because it helps to ensure that there is fairness for students across their peers⁹. There is also a requirement for confidence in the assessment process, meaning that one must be able to trust that the conducted assessment is valid and fair. As described in subsection 2.1.3, fairness is an essential principle in assessments.

There have been studies on the causes of inconsistency in evaluations. Coker et al. (1988) had a hypothesis that increased fatigue while evaluating a high workload results in inconsistency. Fatigue may affect decision-making, as mentioned earlier, making the assessments inconsistent. Another hypothesis is that markers do not utilise the prescribed criteria during the evaluation process. Klein (2002) stated that when the guide is not followed, markers may include inaccurate information in the evaluation process and then make inadequate decisions.

Several measures are applied to ensure consistency in the evaluation of exams, and one of them is double marking. In May 2021, the Norwegian parliament passed a requirement that double marking should be practised in higher education¹⁰. It states, among other things, that two markers must evaluate every grade given. Commonly in double marking

⁹<https://www.nzqa.govt.nz/assets/Studying-in-NZ/New-Zealand-Qualification-Framework/consistency-qual-outcomes.pdf>

¹⁰<https://www.uhr.no/temasider/karaktersystemet/veiledende-retningslinjer-for-sensur/>

2. Background Theory

is that the work is graded by two markers who agree on a final grade. Studies show that markers typically agree weakly with a correlation at about 50% or 60%. When enough time has passed, and the marker is asked to re-assess the same exam, the correlation is about 70% with the first rating (Page and Petersen, 1995).

By nature, it will not be possible to get two human markers to correlate 100%, as they will not necessarily have the same interpretations of the work, or they might disagree on how well some answers demonstrate what is being assessed. It is difficult to conclude whether or not the correlation mentioned earlier is sufficient or not, or just what the correlation should be.

Developing rubrics for assessments is also a measure to enhance the consistency of the assessment procedures and outcomes. Rubrics are grading guides that make up the evaluation criteria of any student work one seeks to evaluate¹¹. Rubrics strengthens consistency as they ensure that all markers use the same criteria to assess the performance of all students. They provide help for quick analysis to see patterns of strength and weakness in the students' work.

In addition, rubrics can also help perform fair assessments, as the work is compared towards a standard rather than against each other. If several markers are involved, and they all base their evaluation on the same rubric, it is believed that this ensures fairness. Faculty from Duquesne University report that their professors do grade more fairly and efficiently when using a rubric¹¹. Rubrics are time-saving as one does not have to keep writing and repeating comments. Many students tend to make similar mistakes, making it possible to give feedback from pre-written comments, which reduces time spent assessing. According to Kryder (2003) they contribute to time-saving by forcing the marker to be concise in feedback as there is limited space.

2.1.7. Efficiency in the Assessment Processes

There is a growing number of students in higher education and, consequently, a more significant number of exams. Following is a grading period where stretches with a pile of exams await the markers. Within this period, markers have to perform consistent and fair assessments. The high workload and maintenance of external requirements increase the need for assessing efficiently.

There are literature and tips shared by adjuncts, markers, and other professionals within the field of assessment on different strategies to make grading easier¹², more efficient¹³

¹¹<https://www.duq.edu/about/centers-and-institutes/center-for-teaching-excellence/getting-started-teaching-at-duquesne/grading-smarter-through-rubrics>

¹²<https://www.edutopia.org/article/7-strategies-make-grading-easier>

¹³<https://gsi.berkeley.edu/gsi-guide-contents/grading-intro/grading-efficiently/>

2. Background Theory

and on how to reduce time spent on grading¹⁴. Some mention preliminary work such as assignment design and state that clearly worded assignments and clear learning objectives will significantly improve grading efficiency. Having multiple-choice questions will, for instance, increase the efficiency of the evaluation process. Still, there are subjects where multiple-choice exams are not suitable. These will have other types of exam questions and be heavier to assess. Setting a limit on the length of the assignment and creating a rubric or a grading scale are other measures. During the evaluation process, tips are to mark in batches, create a comment bank and make a grading conversion chart. Other recommendations are related to body and health and concerns. This involves taking enough breaks, having a specific focus, and assessing when you are in a good mood.

In addition to the pressure on doing fair and consistent evaluations, markers also have a time limit. Many institutions have a deadline for the announcement of grades¹⁵. This deadline is independent of the number of students in the subject, meaning that if one is responsible for a subject with 100 or 900 students enrolled, both subjects still have the same deadline. Therefore, many subjects have several markers, which results in more time and work spent on coordination and communication. This delays the efficiency of the assessment processes. Additionally, external requirements such as the previously mentioned double marking will also require additional effort and time spent on coordination work.

2.2. Digital Assessment Tools

Technological tools have been integrated into a large part of our daily and working life, and there is no exception within education. There exist digital tools for different types of assessment, where some are better suited for specific purposes.

Inspira Assessment

Inspira Assessment is a cloud-based e-assessment platform covering the entire digital examination process, from planning and designing to secure and anonymous exam delivery and marking. All types of subjects can be examined in Inspira Assessment, ranging from Science, technology, engineering, and mathematics (STEM), social studies and human science. This means that there are many possibilities for question types. Inspira Assessment is used in more than 160 countries, with several world-class education institutions on the list. In 2020 there were in total more than 2.6 million submissions delivered¹⁶.

¹⁴<https://www.cultofpedagogy.com/cut-grading-time-in-half/>

¹⁵https://lovdata.no/dokument/NL/lov/2005-04-01-15/KAPITTEL_1-3#%C2%A73-9

¹⁶<https://www.inspera.com/assessment>

2. Background Theory

The EdTech entrepreneurs of Inpera Assessment claim that they provide an efficient and fulfilling assessment experience for both students and staff. There are mechanics for automatic marking of objective questions such as multiple-choice questions, drag and drop questions, selection tasks among predefined answers, fill-in questions, etc. In addition, it also offers collaborative tools for markers that contribute to speeding up the evaluation process. There is also support for double marking in the platform.

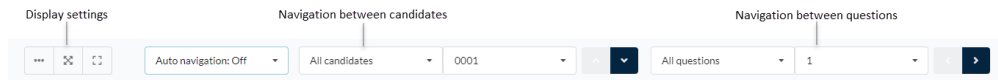
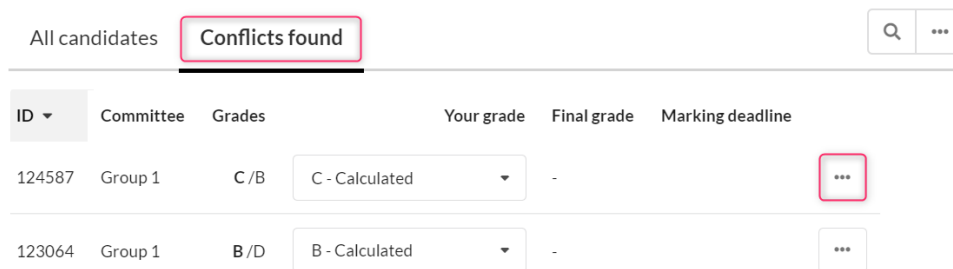


Figure 2.1.: Bottom navigation menu for evaluation from Inpera Help Center.

The platform offers much freedom to customize the evaluation process. Figure 2.1¹⁷, which is the bottom navigation menu, displays the options a marker can choose between.

A marker can filter on candidates that have been flagged or on candidates where the assessment is incomplete. There are also options for navigating between candidates if one prefers that in the dropdown menu with the candidate ids. Also, one can filter what questions to display by clicking on "All questions" and choosing between those that are manually assessed, automatically assessed and answered questions. Navigating between tasks can be done by clicking on the question list on the right.



ID	Committee	Grades	Your grade	Final grade	Marking deadline
124587	Group 1	C/B	C - Calculated	-	...
123064	Group 1	B/D	B - Calculated	-	...

Figure 2.2.: Detection of grade conflicts from Inpera Help Center.

Regarding the double marking for quality assurance, the system intercepts whenever there is a grade conflict. This functionality makes it easy for markers to discover conflicts, so they can review the assessment of the candidates it applies for once more and agree upon a grade. See Figure 2.2¹⁸ for an example of a grade conflict.

¹⁷https://support.inpera.com/hc/en-us/articles/360041446492-Footer#_01FS23D6E8B8J5ZJFRNGQ2QVTG

¹⁸<https://support.inpera.com/hc/en-us/articles/360041880811-Guide-Multiple-markers-with-committees>

2. Background Theory

Inspira Assessment offers the possibility to export results from an exam in JSON format¹⁹. The downloaded data is not processed, but it can be processed by outer software to retrieve more manageable data. As Inspira Assessment also offers the option to import CSV files²⁰, this opens up for flexibility and opportunities for markers to assess exams in other systems.

Word Processing Software

Depending on the type of exam and questions, not all exams require security, anonymity and more complex settings. Hence, word processing software like Microsoft Word, Google Drive Documents, and Apache OpenOffice Writer are sufficient for some exams. These software are fit for both shorter and longer text answers. One can also create multiple-choice tests or completion tests. Relying on how the exams are delivered, there are different assessment methods. If the file is provided as a PDF, the marker can assess the exams by reading through and evaluating them by hand. Another option is to assess the exams digitally. There exist software for reading PDFs, such as Adobe Acrobat PDF reader, and these software often give opportunities to highlight and comment on the PDFs.

2.3. Related Work

As global digitalisation extends, it opens new possibilities and improvements in the assessment process. Technology is already being used in evaluation settings, and different programs and techniques have been developed to support the assessments and assessment processes. Different attempts have been made to create technology support for grading, meaning there is a need for improvement within the area. To address RQ2, it is essential to study the ongoing work of others and what advantages or disadvantages there are to how they have addressed the challenges of the assessment process.

2.3.1. Computer-Assisted Grading Rubrics

Rubrics are claimed to be a tool that contributes to grading more consistently and efficiently. Though some believe that the standard grading rubrics help, there are criticisms of grading rubrics. One is that students may not perceive that they have been graded fairly. Students may lack the ability to self-assess the work and, therefore, struggle to see the context between the markers ratings and their work. Kryder (2003)

¹⁹<https://support.inspera.com/hc/en-us/articles/360029018731-Results-export>

²⁰<https://support.inspera.com/hc/en-us/articles/4406270769681-Import-a-CSV-file-into-Inspira-Assessment>

2. Background Theory

states that limited space is for some a benefit, others might see this as a problem as the feedback is not detailed enough or too generic. Popham (1997) means that it has to be somewhat generic for the assessment to fit in one category, or the rubric will break down.

To accommodate the criticism, people have been researching computer-assisted grading rubrics. Czaplewski (2009) created computer-assisted grading rubric built on the framework of the standard rubric. It utilises the power of database technology to store and quickly retrieve a set of pre-written comments and feedback. A database minimises the work of rewriting comments and feedback and allows them to be carefully constructed. The standard rubric does not provide any specific comments or feedback, but comments and semi-custom feedback are imported with the computer-assisted grading rubric. Letting students receive a scoring grid with comments and feedback to their work may increase the perception of fairness in grading. This is due to students only receiving what is relevant for their their submission, and not descriptions of other categories, which they get when receiving the rubric.

Anglin et al. (2008) have a study on the efficiency of computer-assisted grading rubrics compared to other grading methods. Efficiency was in this study measured by the time the professors used to grade the assignments. The study showed that computer-assisted grading was faster than both traditional hand grading without rubrics and hand grading with rubrics, with a score of 300% faster for the former and 350% faster for the latter. In addition, it was nearly 350% faster than typing the feedback into a Learning Content Management System.

2.3.2. Feedback System on Programming Assignments

With a starting point of Harvard University's CS50 course, a web-based utility was made to improve the feedback process on the assignments in the course. CS50 is an introductory course to computer science for majors and non-majors²¹. Each week, students handed in programming assignments that were assessed and received feedback from the staff in qualitative comments on PDFs. Staff had to download the assignments from a central repository and then create and annotate PDF documents of the students' source code. After this, the staff manually emailed the PDFs with feedback to the students. The process of downloading, annotating and sending PDFs was considered a bottleneck in the workflow. The staff expressed that much time was spent on this. An attempt to improve the efficiency of the process was made, so they could spend less time on logistics and more time on teaching and providing feedback.

MacWilliam and Malan (2013) created CS50 Submit, which is a web-based system that facilitates the collection of assignments and feedback on the source code. Students could submit their assignments using a command-line utility or a web browser, and the staff

²¹<https://pll.harvard.edu/course/cs50-introduction-computer-science?delta=0>

2. Background Theory

could attach inline comments to the source code in the form of sticky notes. There was also the possibility of leaving overall comments. The logistical bottleneck was removed by centralising the submission and evaluation process with a single web application. The staff reported that they spent 10% fewer hours on grading per week and a decrease of 13% in minutes per student while providing as much or more feedback. This led to more time on giving feedback, and it was reported that there was an increase in the amount of time spent answering questions from the students online.

2.3.3. Automated Essay Scoring Systems

In the work of improving the evaluation process, Automated Essay Scoring (AES) has been introduced. AES uses specialised computer programs to assign grades to essays written in an educational setting. The goal is to classify a large set of textual entities and then be able to categorise them before finding a corresponding grade²². The objective is that an AES system should score an essay such as a human marker.

The origin of AES is traced back to 1966 when Ellis Batten Page presented the possibility of scoring essays by computer (Page, 1966a). Later in 1968, Page published his work with a program called Project Essay Grade (PEG) (Page, 1966b), which was the first of the automated essay scorers. Back then, everyone was surprised to learn that a computer could do as well as a human marker.

There are automatic systems that rely on different measures for evaluating essays. PEG relies on style analysis and proxy measures which are essay length, average word length, the number of prepositions and so on (Valenti et al., 2003). Therefore, it is based on writing quality and is not taking any content into account. On the contrary, the Intelligent Essay Assessor(IEA) focuses primarily more on the quality of the content of an essay. However, it still includes scoring and feedback on style, grammar, etc. Another system, E-rater, uses a combination of statistical and NLP techniques to score essays by extracting a set of features representing essential aspects of the writing quality for each essay.

Advantages and disadvantages of AES

AES systems have been researched and extended in decades since Page first introduced the topic. The systems mentioned above are just some systems within AES that exist. Even though there are systems that reduce the workload, one can question why these systems have not been more widespread and integrated into higher education than they are currently. This might be because doubts and a lack of trust in AES systems still exist, and there is an assumption that only humans can evaluate the quality of essays.

²²https://en.wikipedia.org/wiki/Automated_essay_scoring

2. Background Theory

One can conclude that the current AES systems have several benefits and challenges. First off, it is a time and money saver. While markers would spend lots of time and work hours to score and give feedback, an AES system can do this within seconds (Shermis and Burstein, 2003). In this way, the waiting time for students to receive feedback is also considerably reduced. In addition, these systems are flexible and can evaluate and score at any time. Another benefit of AES systems is that the systems can make decisions without being biased or judgemental.

Even though these AES systems improve the efficiency of the heavy evaluation process, some challenges still exist. A significant factor is the accuracy and reliability of the systems. Regarding the performance of the systems mentioned above, PEG has a multiple regression correlation of 0.87, the IEA has an agreement of 85%-91%, and the E-rater range from 87%-94% with human markers (Valenti et al., 2003). Whether these results are sufficient enough is debatable. On the other hand, the correlation between two human markers will most likely not match entirely. One can argue that computers lack the elemental human capacity to separate good and bad writing. Another challenge is that some people might trick the system by learning how it works. If the styling, language, and grammar weigh more, people can learn and use more rare or topic-specific words to get a higher score.

3. Design

A concept has been designed to address RQ2: "How can assessment support be designed to support efficiency and consistency?". It was developed with inspiration from two methodologies; Design and creation and Experimental research. Through this process, the Side-by-side comparison was created with other concepts to support it. As Figure 3.1 shows, a typical assessment process is to evaluate one answer at a time. However, with Side-by-side comparison illustrated in Figure 3.2, the marker is presented with multiple answers simultaneously and can use them to compare. It can help create a baseline when using norm-related referencing.

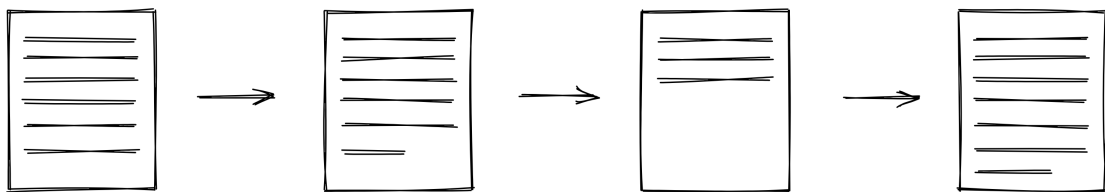


Figure 3.1.: The typical assessment process where the marker evaluates one candidate at a time. Each page represents a unique answer, resulting in four unique answers in this figure.

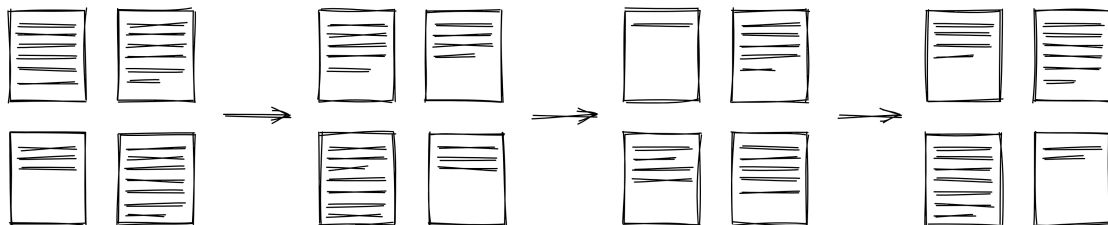


Figure 3.2.: The Side-by-side comparison presents four answers simultaneously, enabling comparison between the answers. Each page represents a unique answer, resulting in 16 unique answers in this figure.

The Side-by-side comparison and its' support concepts include the Task-by-task assessment method, Grouping and Consistency check. These four concepts aim to explore how they can affect efficiency and consistency in an assessment process. To arrive at the concepts, one can utilise different approaches. As described, two methodologies were used to design and develop the concepts.

3.1. Methodology

A research methodology is a strategy that describes the overall approach to answering the research questions (Oates, 2006). It contains all the steps used to identify, process and analyse information associated with the research questions.

This project is inspired by the strategy Design and creation research, with some inspirational aspects from the Experimental research strategy, both of which are described by Oates (2006). The steps of each strategy were used as guidelines and not followed rigidly in a strict sequence.

3.1.1. Design and Creation

The design and creation strategy focuses on developing new IT products, also called artefacts. This strategy is suitable when a system or parts of a system are produced to contribute to the research. Typically, the artefact represents an approach to solving a problem. When conducting design and creation as a strategy, Oates (2006) presents five steps that should be addressed. These are awareness, suggestion, development, evaluation and conclusion. These were followed and conducted as an iterative process, which Oates (2006) describe as learning via making. An example is that during the development and evaluation, one might discover new findings, or one might have to discard the current suggestion. This would then lead to revising and working further on the suggestion.

Awareness

This step is all about recognition and understanding the problem. Preliminary studies were conducted to build a knowledge base on the assessment process and previous work in the field. It is important to create a knowledge base to identify gaps in the current literature. This is to avoid reproducing any studies that have already been carried out and proven.

The preliminary studies provided insights on essential aspects within assessments, how evaluation is performed, what digital tools markers have available and what issues and expectations exist. The outcome is presented in [chapter 2](#).

The initial aim of this thesis was to explore if a digital Side-by-side comparison could support markers in the assessment process. The preliminary studies resulted in limited documentation on this method within evaluations and indicated that research is needed.

3. Design

Suggestion

The second step involves taking a creative leap about the problem and then creating an idea of how the problem might be addressed. In [chapter 3](#), a suggestion is presented.

To test the Side-by-side comparison, it would need to have some framework to showcase how the evaluation method could be applied in practice. As the method would be applied digitally, a software system was an obvious suggestion to showcase and test the evaluation method. Design suggestions for the software system started with paper sketches, which later were developed into more defined and concise digital mock-ups. Early suggestions were simple, revealing how assessments could be presented side by side.

Development

The development step is where the design suggestion from the second step is implemented. How this is implemented varies on the type of IT artefact. The assessment support system involved software development and the construction of algorithms. Early development iterations focused primarily on implementing the design and the graphical user interface (GUI) for a front-end system. Later development revolved around the back-end system's logic with data fetching and saving and the development of algorithms to support the Side-by-side comparison. The development in this thesis is described in [chapter 4](#).

Evaluation

The evaluation step aims to assess the worth of the developed artefact and the deviations from expectations. It is valuable to perform evaluations to get insights into working and non-working functionalities.

Throughout this project, several evaluations have been carried out. There was guerrilla testing conducted with Computer Science students to quickly and informally test ideas. There was also usability testing to uncover potential user experience problems.

The main evaluation was performed after developing a web application in the form of a minimal viable product(MVP) that could show how an assessment could be conducted with the Side-by-side comparison. Interviews with a demonstration or a walkthrough were held to evaluate the concept. The respondents were interviewed about their experiences and thoughts, which are covered in detail in [chapter 5](#).

Conclusion

The last step is about communicating the result of the process. This involves the knowledge gained and ideas for further work. The results of the evaluation are presented and discussed in [chapter 5](#).

3.1.2. Experimental Research

The experimental research strategy builds upon researchers that start developing a theory about their topic of interest, which leads to a statement based on the theory that can be tested empirically via an experiment ([Oates, 2006](#)).

This project draws inspiration from this research strategy. There is no exact solution to support efficiency and consistency in an evaluation process but rather an idea that needs experimenting. The idea is investigated and expressed in RQ2 and RQ3. To perform this experiment, the design and research strategy must be applied first. Hence, the steps explained previously were conducted. After this, one should test whether or not there is any difference in time spent evaluating exams with the Side-by-side comparison to their standard assessment practices. Markers could perform assessments, and one could do time tracking and measure their responses and actions.

An experiment requires a lot of preparation due to rules to comply with for the experiment to be approved. One must, for instance, ensure validity. The experiment must be reproducible, and there should be no external factors that may influence the experiment. One should have a correct and wide selection of respondents and more. Due to this project's scope and time limit, further work to prepare an experiment is not manageable.

3.2. Concepts

Several concepts have been thoroughly discussed and evolved in the work of exploring suggestions to support efficiency in the evaluation process. It could also be helpful if the concepts are easy to understand and apply so that users evaluating them can give comprehending insights and perspectives. The result ended in a digital Side-by-side comparison built on a Task-by-task assessment method foundation with the support concepts: Grouping and Consistency Check. Their connections can be seen in [Figure 3.3](#).

3. Design

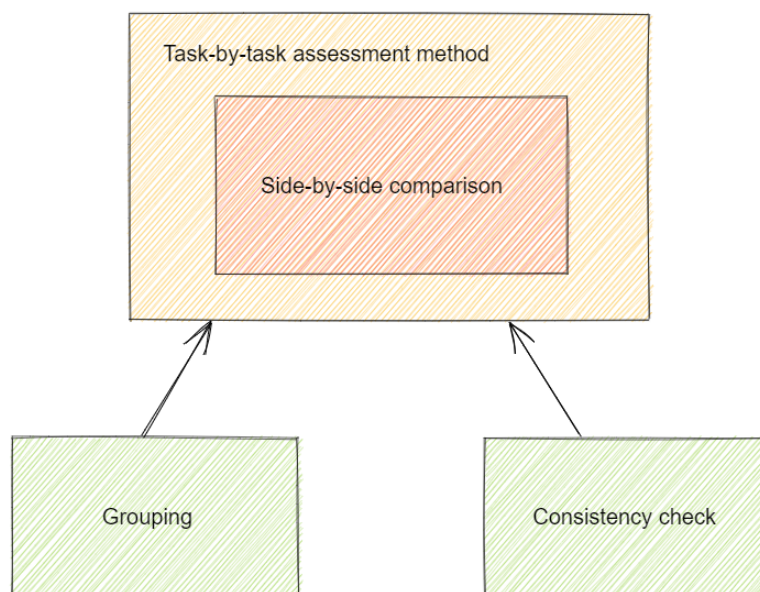


Figure 3.3.: The Side-by-side comparison and its' connections with the Task-by-task-assessment method, Grouping and Consistency check.

3.2.1. Side-by-side Comparison

Side-by-side comparison is a digital concept and evaluation method where multiple student answers are presented for assessment simultaneously. This method assumes that the exam has multiple tasks and the answers are from the same task but by different students. Then, the marker can use the simultaneously displayed student answers, the *batch* of answers, as a reference. Referencing is common practice that markers apply, and this method can be considered as a variant of norm-related referencing described previously in [subsection 2.1.4](#). The Side-by-side comparison should give the marker a quick overall impression of student answers, making it easier to perform assessments. This will again support efficiency in the process. In addition, the display of answers side by side can be a comfort for the markers, especially those that are not very experienced. It is essential to take into account *how many* answers should be displayed simultaneously. Too many answers may be confusing and result in cognitive overload, while too few answers may not give the wanted effect when it comes to efficiency. Displaying a batch of answers together reduces the number of clicks for loading more answers. In total, this may save time and contribute to a more efficient process.

One can argue that the Side-by-side comparison may invite more bias when assessing due to the enabling of comparison. An example is that if one has three excellent answers displayed with one mediocre submission, the latter may receive a minor score because it is relatively inadequate. It might have gotten a different score if it was in a batch with

3. Design

other answers. This implies that scores may vary for answers reliant on which batch it is a part of. Consequently, it can raise questions of the validity of the assessments. However, comparison between answers is also present in other assessment methods, and one can have faith in markers to assess fairly. Side-by-side comparison is introduced as a suggestion to support markers in their heavy evaluation process by giving them an easier overview of the answers to assist their assessments. Then, a baseline can more quickly be established. Additionally, it can reduce time and mental processing on switching between answers. This can then quicken the evaluation process.

3.2.2. The Task-by-task Assessment Method

The Task-by-task assessment method is a workflow where the marker assesses submissions by tasks. For instance, in an exam with five tasks, the marker will first evaluate all submissions for task 1. Then, the marker assesses all submissions for task 2 and so on. Therefore it presupposes that there are exams with multiple tasks. The method differs from the holistic scoring method, where markers perform assessments by candidate. This is to assess an exam from a candidate as a whole and then move on to a new candidate's exam submission.

The previously described Side-by-side comparison is applied with the Task-by-task assessment method as a foundation. This is because it is more valuable and convenient to display submissions from the same task simultaneously. It could be cluttered and less purposeful if different tasks for the same student were displayed together. It may be tiring to have much information on display with no clear common denominator.

Assessing task by task reduces the context switching with the candidate by candidate approach. With the Task-by-task assessment method, one only has to focus on the current task and its' topic. Markers do not need to adjust for new tasks and criteria frequently. This may help to diminish mental fatigue and thereby bias, as discussed earlier in [subsection 2.1.5](#). Additionally, when assessing the same task, one will read similar answers multiple times and have them fresh in mind. Therefore, the marker will become more experienced in assessing the task, which may increase efficiency in the process. Another perk is that the fragmented view of the exam might reduce bias. This is because the final grade is based on scores set on how well the students have demonstrated knowledge from the tasks individually. However, lack of context switching makes it more of a repetitive process which can increase boredom and, therefore, bias.

According to each marker's preferences, it varies whether or not they consider the Task-by-task assessment method preferable. The workflow facilitates focusing only on one topic at a time, which may increase efficiency. It might also amplify and substantiate that the assessment is set on the proper basis because one is in a mindset where the knowledge of the topic is fresh in memory. However, some markers might prefer to view exam submissions holistically, so they can see it as a whole and then set a grade.

3. Design

3.2.3. Grouping

There are different aspects to be considered when viewing answers side by side. They are compared to each other in a local capacity, which may lead to a slight and gradual change in evaluation criteria or biased assessments due to comparison. Therefore, which answers are displayed together in a batch is of high importance when considering consistency and bias. To impact what answers are displayed simultaneously, they can be grouped based on different qualities. The Grouping can be supported by using different sorting algorithms. It concerns what order the answers should be when assessing and, therefore, what answers are displayed in the same batch. The grouping should be done automatically to save time and energy for the user.

There are different ways to sort and group the answers. They can be similar, different or random, and one needs to consider in *what way* are the answers similar, different or random. Utilising these factors in different ways gives different results. E.g. grouping on answer length and grouping on candidate ids gives significantly different groups when assessing. Displaying answers in random order can give qualities like reducing bias and giving every candidate an equal basis in the assessments. However, the Grouping should give the user some added value, and since the answers are usually somewhat randomly ordered, there are not many extra advantages to be gained. Also, due to the qualities of Side-by-side comparison, having answers with no connection displayed next to one another may be confusing and of little value to the user. Some of the grouping factors can be candidate id, text length, and content.

Based on Candidate Id

Similarities between candidate ids mean, for example, increasing candidate ids. Candidate ids are somewhat random, and therefore the answers in the batch will be to some degree random. However, some markers might be used to assessing by candidate id, and by doing so when assessing task by task, it can be easier to recall earlier answers by the same candidate. This grouping might give a more holistic view, even when assessing task by task. However, it can lead to mistakes as remembering earlier answers is left to memory.

Based on Text Length

One example of similarity in text length is to sort the answers from longest to shortest. It is a straightforward solution which gives some correlation between the answers. However, answers may be alike in length, but the content can differ significantly. As it is the *content* of the answers the markers use to assess, this grouping might be confusing to them. If answers are in the same batch, one can assume that the answers have similar

3. Design

content and look for connections between them even though there are none other than the length of the answers. Although, it is more likely that two answers have similar content when the answers are of the same length than if they are randomly sorted. Consequently, there can be some added value when grouping on answer length. Therefore, this can be an easy-to-implement solution as a first step.

Based on Content

Moreover, one can compare actual text content to use for grouping. There are different ways to determine content similarity. Texts can contain exact keywords, assumably have the same score, or have similar sentence structure. Comparing text content can be challenging due to different writing styles, and there are different techniques to be used, such as term frequency-inverse document frequency¹, the Vector-Space Model² and the probabilistic retrieval model³. Another technique could be semantic similarity, where the relationship between texts or documents is scored using a defined metric in the area of NLP⁴. One challenge with text comparison is that texts might express the same subject while using different words, or they can use the exact words in a different way or context, resulting in different meanings.

In addition to content comparison, the answers can be compared to the marker scheme or each other. Comparing answers to each other can be challenging due to candidates' different writing styles. One can look at keywords, sentence structure, or semantic differences. Comparing candidates' contents to each other might help give a fair assessment because then similar answers can get the same score. It is easier to remember what score to give when they were given to similar answers recently. Also, it can help with uncovering plagiarism between candidates. The other form of comparison, comparing the answers with the marker scheme, can also be beneficial. Answers with high similarity will often be equally similar to the marker scheme. However, it can also apply to answers equally similar to the marker scheme, although not as similar to the other answers. By grouping on similarity to the marker scheme, it might be easier for the user to compare every answer to the same standard. If the marker has written some keywords that they are looking for in a particular task, grouping based on these keywords and possibly synonyms of the keyword might ease the process. When all the answers are grouped, it might be easy to add additional functionality, such as highlighting keywords or relevant text in the answers. This functionality might speed up the process, but it leaves room for missing significant and meaningful content in the answers if the system does not catch it.

¹<https://monkeylearn.com/blog/what-is-tf-idf/>

²https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_918

³https://aspoerri.com/info.rutgers.edu/InfoCrystal/Ch_2.html#2.3.2.1

⁴<https://towardsdatascience.com/semantic-similarity-using-transformers-8f3cb5bf66d6>

3. Design

Concluding Grouping Approach

To group answers, different sorting methods can be used. Generally, sorting answers based on content is more advanced than using candidate id and text length. Therefore, one needs to prioritise effort and gain. Using the correct form of grouping may reduce bias and increase consistency and efficiency.

3.2.4. Consistency Check

A consequence of the Side-by-side comparison described [subsection 3.2.1](#) is that markers may be unwillingly biased due to the comparison of student answers when presented in batches. Hence, a mechanic to unveil and prevent such bias was created. The Consistency check is based on having some student answers to be re-assessed. An interesting factor is that the marker is compared against themselves for consistency. The check also ensures safety in the quality of the assessments since multiple students' submissions are double-checked. It helps to increase the reliability of the evaluation and score set.

The Consistency check builds on algorithms that will regularly choose student answers to be re-assessed. When designing these algorithms, there were several factors to take into account. Questions that were discussed were:

- Q1: Which students' answers should be re-assessed?
- Q2: When are the student answers going to be re-assessed?
- Q3: How many student answers should be selected for re-assessment?
- Q4: What happens if conflicts in assessments appear?

Q1: Selecting Student Answers

Different mechanics were discussed when evaluating which answers to be chosen for re-assessment. Firstly, answers can be selected randomly. Secondly, they can be chosen based on the score given. Other options include choosing based on a correlation between answer score and answer length or selecting answers flagged by the markers. An alternative is also to have all answers re-assessed. It also had to be discussed whether the re-assessments should be chosen from the current batch or amongst the entire set of answers.

3. Design

Randomly

Selecting random answers may uncover inconsistencies that would otherwise not be discovered. On the other hand, not having a specific strategy for choosing answers can seem vague, which again will influence the confidence in the Consistency check. For this mechanic, one can select answers for re-assessments at any time.

On Outlier Score or Similar Score

When choosing answers based on score, the answers must be compared to a basis. One natural basis is the current batch with the student answers. Within the batch, there are several options. An option is to choose an outlier answer in the batch. An outlier is an answer that differs from the rest of the batch. For example, there are four student answers in the current batch. The maximum score for this specific task is 2 points. If 3 out of 4 answers get a score of 2 points, and the last one gets a different score, then the last answer is chosen for re-assessment. The intention is to detect situations where an answer may have gotten a lower score due to the batch it lies within. If there is an equal distribution of scores, meaning that, for instance, all answers are scored with the same points, there will be no outlier. For the latter, no answers will be chosen for re-assessment.

Another option is, to some extent, the opposite of the first option described above. This option is to choose an answer amongst those with the same score. Using the first case described above, this option will not select the outlier with a different score for re-assessment. For this case, the algorithm will pick one answer from the batch of the three answers that got the same score. Within this batch of 3 answers, the answer to re-assess is chosen randomly.

For these options, it seems suitable to choose re-assessments from the current batch, but it is possible to use it with several batches. One would then have to calculate and determine the basis of that larger batch. The advantage of using a larger batch is that one has a larger volume for comparison and gets a greater spread and nuance. This option will consequently find fewer outliers, but one can expect those to be more precise.

On Correlation between Answer Length and Score

The idea behind selecting answers based on the correlation between answer length and score is to explore if there is a recurring case that the longest student answer gets a high score and that markers may be biased and influenced by the word count. From [subsection 2.3.3](#) one knows that some AES systems emphasise proxy measures like length. It is reasonable to assume that human markers also value answer length and are affected by it when assessing. This correlation check is interesting with the Side-by-side comparison because when presenting answers in batches, the answer length is a measure one can quickly and visually see and compare consciously or unknowingly. Therefore, having this check is valuable to examine if answer length is a measure that may cause bias in assessments. To ensure consistency, the longest answer is therefore up for re-assessment if it also has a high score.

3. Design

For this correlation check, one can pick the answers that will be re-assessed from both the current batch and a larger batch. Choosing the current batch will result in more re-assessments to be checked. On the other hand, it may pick too many re-assessments that are actually "good" and would not necessarily be selected on a larger scale. Hence, a larger batch would be able to weed out these. A larger batch would not discover any new or different re-assessments as the batch is length-based, but it will choose fewer re-assessments.

On Flagged Answers

Regarding flagged answers, this is a feature where the markers can choose to flag (mark) answers. This feature is common in many web applications where the aim is to remind oneself to follow up or take action later. Hence it is natural that some flagged answers are re-assessed as there usually lies a thought behind flagging. For this option, one could select a proportion of the flagged answers randomly for re-assessment. It depends on the number of flagged answers. If the amount is under a reasonable limit, all flagged answers could be re-assessed, but if not, then a proportion should be selected.

All Answers

The last option is to have all student answers re-assessed. Then, the same marker assesses answers twice. The benefit of this option is that one can ensure that consistency and effort are put into the assessments since all answers are double-checked. On the contrary, it can be a high workload, and markers may not be motivated for it.

Q2: Choosing When to Re-Assess

There were two options considering when the answers should be re-assessed. The first option was to re-assess during the assessment of the current task. The re-assessments would appear with new, unassessed answers for the current task. The other option was to re-assess answers after having assessed all answers once.

The perks of re-assessing during the assessment of a task are that the re-assessments and the new assessments would be mixed up, making it less obvious to distinguish what type of assessment it is. This can be considered a benefit because an assumption is that markers may be affected if they know they are re-assessing an answer. One thought is that they may use less time and effort on assessing because they have already spent time evaluating and grading the answers. One must consider the chosen grouping and the frequency of the same answers for this option. It will be problematic to randomly mix up the re-assessments with the new assessments due to the different sorting algorithms from the grouping. Some of the algorithms are based on text length, making the possible re-assessments appear first in the next batch. With a short time between the appearance of each answer, the marker will most likely remember the first score and be influenced by it. There is a high probability that this will result in a correlation of 100%, which is unusual, and hence the Consistency check will be less reliable.

3. Design

To ensure that enough time has passed between the assessments of the same answers, the second option is to do re-assessments at last. Then, a marker will finish the evaluations of all answers and then assess all of the re-assessments. With this option, one will not have the problem with the sorting algorithms. A possible drawback with this option is that the assessments and re-assessments are relatively separated. With several of the group sorts except for the random sort, one can sense that the assessments are re-assessments. It is because the assessments are sorted by candidate id or by increasing or decreasing text length, which is visually easy to see. If one pays attention, one can see that this is all reset for the re-assessments. Because it is more evident for markers when they evaluate re-assessments, there is a danger that they will make faster decisions than what they would do initially. This may weaken the Consistency check, but on the other hand, it can detect if the marker's new and maybe faster evaluation is still aligned with the first evaluation.

Q3: Quantity of Assessments to Re-Assess

Considering how many assessments should be chosen for re-assessments, options were either all answers or a percentage of them. Choosing all answers will ensure that all assessments have been double-checked, which gives high reliability to the assessments. Simultaneously, performing re-assessments on all assessments is a high workload and time-consuming. As a result, markers might not support this and have a more negative attitude towards the Consistency check. One must be able to ensure consistency and reliability, but one has to consider the extent of which cost. The Consistency check will not be regarded as valuable if it, for instance, goes at the expense of the efficiency in the evaluation process.

Choosing a maximum percentage of the assessments to re-assess is convenient to compensate and accommodate consistency and efficiency. The percentage must balance the number of re-assessments, workload and efficiency. Regardless of the percentage, one can conclude that it is still valuable to have a Consistency check for quality assurance. Even though not all answers are double-checked, there will at least be some answers where one may uncover inconsistency.

Q4: Resolving Conflicts

The last question addresses the outcome of the Consistency check. If the original assessment and the re-assessment are aligned, meaning that both assessments resulted in the same score, then these assessments have passed the check. The contrary situation is when different scores have been set on an answer, which leads to the assessments not passing the check. The check detects that there is a conflict and inconsistency in the scores. To resolve this situation, an algorithm could calculate the average of the scores

3. Design

and then automatically set this new average score as the new score. It would be a quick and efficient solution, but it could result in an unfair score. Also, if the average number of the two scores lies between possible score options, it can be considered unfair to round the score automatically. Another option is to notice the marker about the different scores and then have the marker re-assess the student's answer a third time to fulfil the Consistency check. This option is similar to Inspira Assessment's function of resolving conflicts, which is described in [section 2.2](#).

The Chosen Solution

After discussing the possibilities, one of the decisions for the Consistency check was to have multiple options for choosing the re-assessments. It offers flexibility and opportunities for the marker. These options will allow the marker to personalise the Consistency check towards what they believe influence or bias them. Therefore, random selection, selection based on scores, and correlation between answer length and score are included. The maximum default percentage for re-assessments is set to 20% based on the assumption that this was a reasonable proportion markers would be willing to assess. However, flexibility is often desirable, and therefore a solution where the marker can adjust and decide what the percentage should be is designed.

Regarding when to re-assess, the decision was to have the re-assessments when all assessments have been finished. This is because this solution bypasses the conflicts with the length sorts, and it also assures that enough time has passed between assessing the answers. To resolve conflicts, the choice fell on having the marker re-assess the answer once more rather than automatic scoring to ensure no mistakes in the scoring.

4. Implementation

Implementing the concepts enables users to test a holistic product so that it can be evaluated in a more realistic setting. It also allows learning the domain and uncovering unknown hurdles which need to be considered. When implementing the concepts, there are several decisions to take into account. One has to find the suitable method for implementation, choose the technology stack, consider framework, programming style, and similar. Additionally, during implementation, it is essential to conduct user tests to get input on design and usability.

4.1. Method of Implementation

Choosing a method of implementation is helpful when organising, controlling, and planning the work process and workload. There are different options when choosing a methodology, and one must figure out which suits their project. The selected method for this project draws inspiration from the agile methods Scrum and Kanban.

Scrum is established around having a set of defined roles with artefacts to represent work or value¹. In a development context, the main roles are the product owner, the scrum master, and the developers. Common scrum artefacts are the product backlog and the sprint backlog. The method is based on splitting work into smaller iterations to shorten the feedback loop. These smaller iterations are referred to as sprints which consist of work items ordered in a backlog. Each sprint consists of several events, including a planning meeting at the start of each sprint, daily stand-up meetings, and a review and retrospective at the end of each sprint. A set of work items estimated to be completed within the sprint is selected during the sprint planning meeting. After this meeting, the sprint backlog is locked, and it is prohibited to add further work to the current sprint backlog.

Kanban is a methodology that aims to give team members just enough work, so the team is consistently working at capacity². It consists of managing work as Work In Progress and is usually performed by using a Kanban board with three columns which typically are

¹<https://www.scrum.org/resources/what-is-scrum>

²<https://www.atlassian.com/agile/kanban/kanplan>

4. Implementation

"to do", "doing", and "done"³. Kanban offers flexible planning, clearer focus, transparency and efficient communication. Work items on the board are the top priority, and the board helps communicate the status of the different work items. A core property in Kanban is to limit the number of work items in "doing" concurrently to prevent overloads and retain a smooth workflow⁴.

The adapted methodology includes artefacts from both of the abovementioned methodologies. It was decided to use an agile methodology rather than the more rigid Waterfall method, which requires specifications to be known ahead of implementation. It was known that several of the concepts were going to be evolved during implementation. Hence, an agile methodology was more suitable. Due to the number of people involved in this thesis, assigning roles was unnecessary. Scrum artefacts such as having a sprint, conducting sprint planning meetings and reviews were practised to help organise and plan the implementation of the web application. Regular updates which can remind of the daily stand-up meetings were conducted. The work items for each sprint were added to a Kanban board. The board with the three columns to define the status of the work items was used actively during each sprint. Regarding the feedback-loop, there were held two usability tests during the implementation phase, which are further described in [section 4.7](#). Additionally, there were continuous meetings with feedback from the supervisor.

4.2. Technologies

The development of the web application in this thesis focused on rapid development due to the project's scope. There was limited time, and implementing functionalities was prioritised to demonstrate how the Side-by-side comparison could be applied. Still, there were some techniques followed to ensure readable and clean code. The web application has a simple architecture that covers the purpose of this thesis. It uses the Next.js framework in the front-end and HTML5's localStorage for storage.

Framework

React is a JavaScript library for building user interfaces⁵. There were several motivations for choosing React. Firstly, familiarisation and experience with the framework lead to less time spent learning the framework, which gives more time for implementation. Secondly, React is free and open-source, making it flexible and cost-saving. Lastly, React is fast, scalable and simple to use to develop the web application. Using Vue instead of React was

³<https://kanbanize.com/blog/kanban-101-the-kanban-board/>

⁴<https://kanbanize.com/kanban-resources/getting-started/what-is-wip>

⁵<https://reactjs.org/docs/getting-started.html>

4. Implementation

considered because Vue is applicable for smaller projects, whilst React is more suitable for larger applications. For this project, React might therefore be considered excessive. However, due to the time limit and previous experience, React was chosen.

TypeScript is a typed programming language that builds on JavaScript, which results in better validation of code⁶. TypeScript was chosen as it is common to use with React development. The benefit of TypeScript is that it is easier to maintain code as it is easy to read and understand components with types. Additionally, it supports a tighter integration with the editor, making it possible to detect errors early.

Next.js is a React framework for developing single-page JavaScript applications. It brings benefits such as hybrid and static server-side rendering, TypeScript support, route pre-fetching, and more⁷. Using React and TypeScript was already decided, making it suitable to use Next.js. Next.js was chosen because it includes features such as an easy set-up, making it fast to build websites. Also, after designing the suggestion for the web application, it was known that it would include several pages and routing between them. Hence, the route pre-fetching offered in Next.js is especially profitable and helpful as it would save time in development. Additionally, Next.js was recommended by other developers. It was unexplored and unfamiliar, but it was still chosen due to the abovementioned advantages and great opportunity to acquire new knowledge and skills in web development.

Material UI is a React component library that implements Google's Material Design⁸. *Material Design* is a design language developed by Google to make good design⁹. By implementing Material Design, Material UI fulfils many criteria regarding Universal Design¹⁰ and Norman (2002)'s seven principles of interaction design which both indicate well-designed user interfaces. Material UI is a plug and play library, making it fast to set up and easy to use.

Storage

The data storage for this web application is HTML5's localStorage. With localStorage, the data is saved with key-value pairs in the browser with no expiration date¹¹. That means information is stored locally on a user's computer until it is deleted. The reason for using localStorage is that the storage is sufficient for the data volume of this project. There was no need for a heavier database as this would be superfluous.

⁶<https://www.typescriptlang.org/>

⁷<https://nextjs.org/>

⁸<https://www.freecodecamp.org/news/meet-your-material-ui-your-new-favorite-user-interface-library-6349a1c88a8c/>

⁹https://en.wikipedia.org/wiki/Material_Design

¹⁰<https://www.washington.edu/doit/what-universal-design-0>

¹¹https://www.w3schools.com/html/html5_webstorage.asp

4. Implementation

Programming Style

As for the programming style, it was desirable with a clean and readable code. Having clean code was also essential for the rapid development of the web application. One measure to ensure these qualities was to perform code reviews when integrating new code into the base. Code reviews were carried out to check and ensure that code conventions were followed, the code was understandable, and the code worked for its purpose. The disadvantage of code reviews is that it is a slow and manual process which requires time. However, it increases the code quality and encourages learning. Another measure to ensure a clean code was to follow certain coding standards. The Airbnb Style Guide¹² was followed and maintained by the Prettier¹³ extension in the IDE. The code was automatically formatted by using the extension, which ensured that the code style was of good quality and consistent.

Having reusable components was also a measure to ensure a clean code. Components were implemented to be as generic and flexible as possible. It would provide readable code and save time in development as the same code did not have to be re-written. The components had a base which were easy to build on and develop further.

A diligently used tool throughout the development phase was Git. Git is a free and open-source version control system designed to handle both small and very large projects with speed and efficiency¹⁴. It is suitable for collaboration projects as it allows for simple ways to compare versions between different developers. Hence, Git was chosen to ensure quality code and agile collaboration.

Pair programming is a technique where one developer writes codes, and the other observes and reviews code. This technique was helpful when dealing with more complex problems as it encourages developers to collaborate and exchange knowledge. Additionally, continuously checking the code contributes to higher code quality.

4.3. Processing the Data Set

Using data from actual exams allows testing the concept more realistically and helps uncover unknown challenges. For this reason, data from two exams in IT2810 Web development at NTNU was exported from Inpera Assessment and provided for this thesis. Therefore, the web application is developed based on the Inpera Assessment data structure and assumes that other subjects hold the same format. Figure 4.1 shows a condensed representation of the data.

¹²<https://airbnb.io/javascript/react/>

¹³<https://prettier.io/>

¹⁴<https://git-scm.com/>

4. Implementation

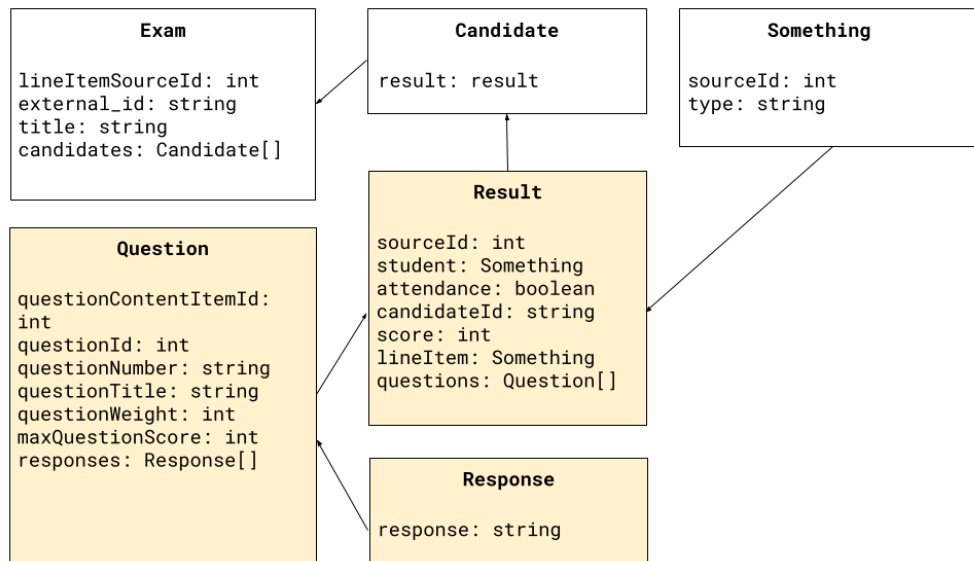


Figure 4.1.: A simplified visualisation of a data set from Inspera Assessment. The most relevant parts of the data set have yellow background.

The essential attributes for the concept is the following:

- **questionId**: The id of the question. It is used to find all the same answers to the same task.
- **taskId**: The id of the task.
- **questionTitle**: The title of the task.
- **maxQuestionScore**: The maximum points a student can get on the task.
- **candidateId**: The id of the student that submitted the answer.
- **response**: The student's answer to the task.

The JSON-file exported from Inspera Assessment is aggregated before being stored in the web application system, localStorage. The process changes the data structure by grouping the data on task number, and each task contains an object for each candidate consisting of the following data.

4. Implementation

- **assessmentId**: The unique id of the object generated by the NPM package UUID¹⁵.
- **answer**: The candidate's answer to the task.
- **candidateId**: The candidate's id.
- **maxPoints**: The maximum number of points one can get on the task.
- **taskNumber**: The task number of the current task.

Through the flow of the web application, other attributes are added and stored in `localStorage`.

- **score**: The score a user gives an answer.
- **isFlagged**: A boolean type that marks an assessment as flagged or not flagged.
- **inconsistentScores**: A list of the inconsistent scores an answer may have gotten after being assessed more than one time.

These attributes store the scores given to each student's answer and support the Consistency check. The processed data structures are shown in Figure 4.2. There, **Data** shows the initial data that is saved in `localStorage`. **Assessment** extends **Data** and contains the answers attached with a given score. **Approved** extends **Assessment** and adds **inconsistentScores** if there are any. **CandidateAndSum** is a fundamental data structure in which form the data can be exported.

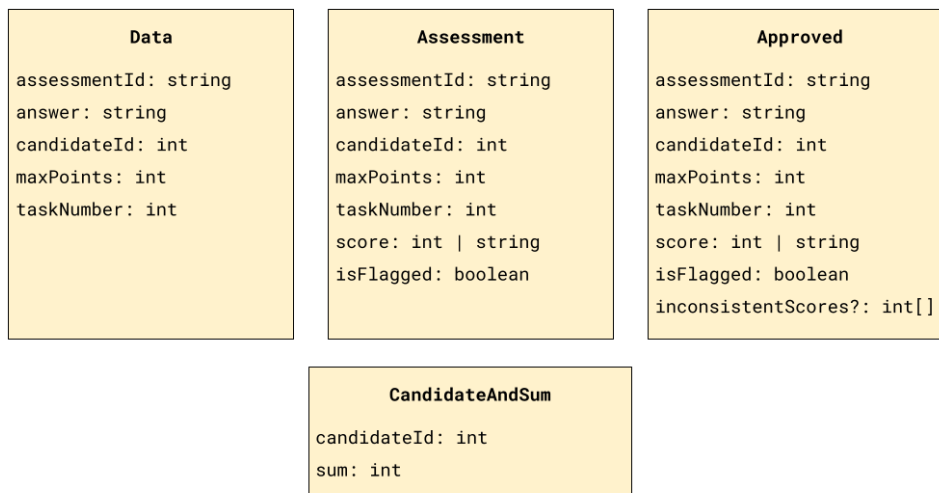


Figure 4.2.: The data structure of data stored in `localStorage`.

¹⁵<https://www.geeksforgeeks.org/node-js-npm-uuid/>

4.4. Enabling the Concept

There are different ways to develop a system that includes the mentioned concepts, technologies, and data sets. To enable the concept, it has resulted in three main parts, as can be seen in Figure 4.3. Through this flow, the exam is uploaded, assessed, confirmed, and exported, all in line with the concepts. The second part, Assessing, contains the core functionality regarding the concepts. Initialising the assessment and Finalising the assessment support the Assessing, so the concepts are set in a holistic system.

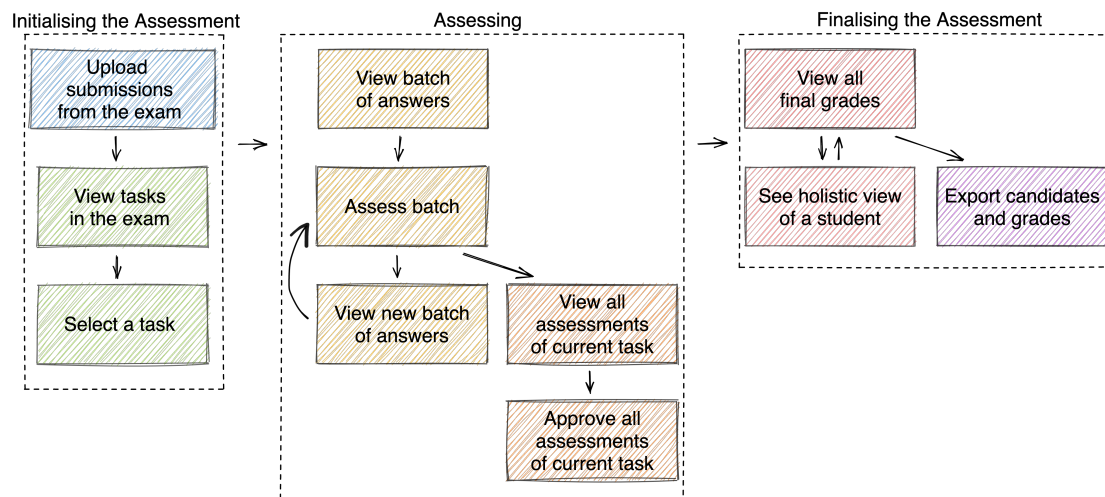


Figure 4.3.: The flow of a holistic system that enables the concepts and uses the chosen technologies and data set.

Initialising the Assessment

Before assessing an exam, considerable preparation needs to be done. There is coordination between academic staff regarding what exam and which candidates to assess, cooperation between markers, distribution of assessment responsibilities, and receiving the exam results to assess. Firstly, the user should be able to view the exam submissions, and, as this system is independent of other assessment systems, the user needs the functionality to upload their data set of exam submissions. Also, it could be useful if the user could upload multiple exams and have a user-friendly system to select which exams, candidates, or tasks to assess. Additionally, if the system could be synced between users assessing the same exam, it might ease the workload. When the exam results are in the system, they must be presented to the user. The exam results are presented task-wise rather than listing each candidate. The reason is to align with the concepts due to the task by task foundation of the Side-by-side comparison.

4. Implementation

Assessing

Assessing exam results includes the core concepts of the project. It covers the concepts previously mentioned in [section 3.2](#), such as the Side-by-side comparison, the Task-by-task assessment method, the Consistency check and Grouping. As a suggestion to RQ2, this thesis focuses on the Side-by-side comparison. It requires using the Task-by-task assessment method so that the answers displayed next to each other have a connection and are comparable. Also, to further the connection between the answers, a precise grouping is necessary. If the content of each answer is similar, they are more comparable, and comparison is a significant element in Side-by-side comparison. To evaluate and prevent bias in the assessment, the Consistency check is in place. It selects assessments that the system believes might need to be re-assessed to verify that the assessments are not biased.

Finalising the Assessment

After completing the assessment, the results must be conveyed in a useful matter. There should be an overview of every candidate with their corresponding grade. Also, a grade distribution can be of interest so that the user can know at what level the exam or candidates are. Then an adjustment of grades might be needed. It should also be possible to see the assessments for every candidate. It gives the user control of their assessment and the possibility to get a holistic view of a candidate. This functionality is not directly crucial for the concept, but it might be necessary for a real assessment system. Also, because the solution is independent of other assessment systems such as Inpera Assessment, the results need to be exported in a way that other systems can use. It could be exporting the candidate id and their corresponding grade. If the other grading system has functionality for uploading scores for each task or sub-task, that information can also be exported.

The Solution

To develop the concepts, a web application was made. It resulted in different pages containing components and functionalities. The four most significant pages are shown in [Figure 4.4](#), [Figure 4.5](#), [Figure 4.6](#), and [Figure 4.7](#). They hold most of the components and functionalities. The pages will be further explained after explaining each component.

4. Implementation

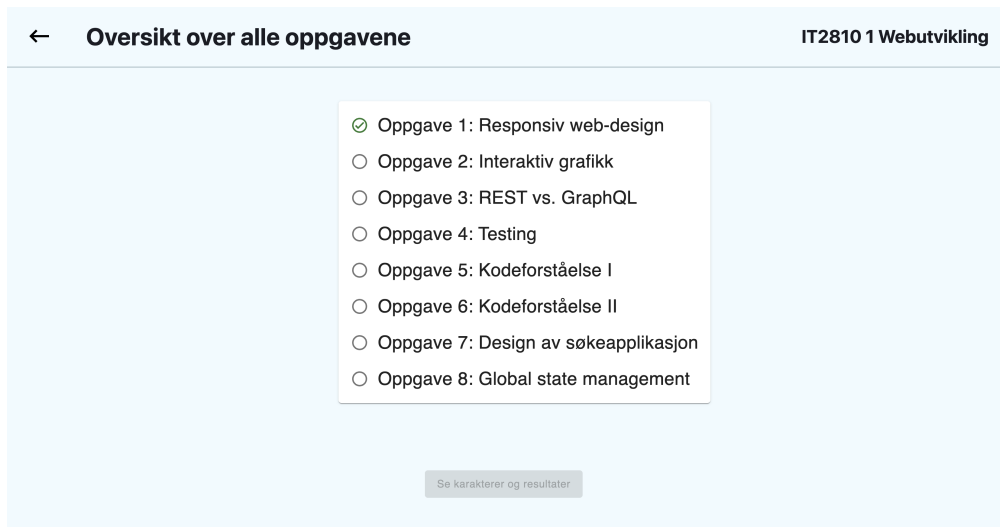


Figure 4.4.: The Task page contains an overview of the exam's tasks and is a part of initialising the assessment.

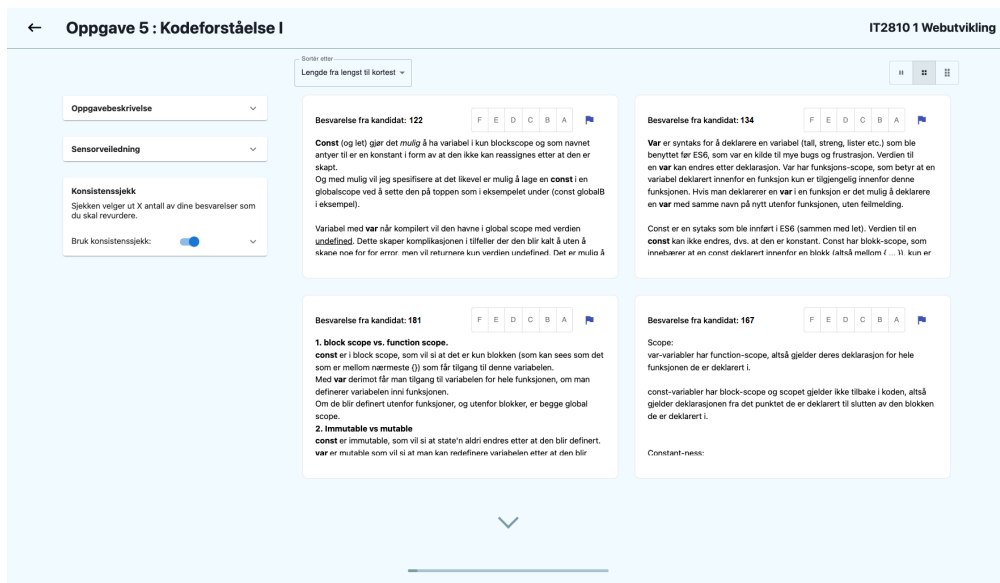


Figure 4.5.: The Assessment page holds the main and supporting concepts. The page has multiple components and functionalities and is the central part of assessing.

4. Implementation

← Godkjenning av oppgave 5: Kodeforståelse IT2810 1 Webutvikling

117
1. const kan ikke endres, var kan endres
2. const er definert i block scope, var i function scope

F E D C B A

142
Scope:
var-variabler har function-scope, altså gjelder deres deklarasjon for hele funksjonen de er deklart i.

const-variabler har block-scope og scopet gjelder ikke tilbake i koden, altså gjelder deklarasjonen fra det punktet de er deklart til slutten av den blokken de er deklart i.

Constant-ness:
var-variabler kan endres helt fritt til hvilken som helst verdi.

const-variabler er en constant-reference, altså kan de ikke endre hvilken verdi de refererer til, men dersom de refererer til et objekt så kan strukturen til dette objektet endres.

Konflikt! Det er blitt satt karakterene: B og C

159
Const (og let) gjør det mulig å ha variabel i kun blockscope og som navnet antyder til er en konstant i form av at den ikke kan reassignes etter at den er skapt.
Og med mulig vil jeg spesifisere at det likevel er mulig å lage en const i en globalscope ved å sette den nå innan en `!deklarert` under `!noet` `!deklarert`

F E D C B A

Figure 4.6.: The Approval page concerns verifying grades and resolving contradicting grades. It lies within the Assessing part of enabling the concept.

← Godkjenn karakterer IT2810 1 Webutvikling

Kandidat 153
F E D C B A

Kandidat 158
F E D C B A

Kandidat 168
F E D C B A

Kandidat 182
F E D C B A

Kandidat 235
F E D C B A

Figure 4.7.: The Completion page shows the total grade each student receives as a part of finalising the assessment.

4.5. Components

The components are the elements in the web application. They are created to give the web application the desired functionalities to enable the concept. The components are made concerning React and Typescript conventions. It is focused on re-usability, comprehensive design for all components, and object-oriented programming. When presenting screenshots of the components, the candidate ids have been anonymised.

Header

The Header is a component that always is on the top of the page using it. It provides information and navigation. The user is given an understanding of what kind of page they see through a headline in a large font. On the left side, there is an arrow pointing to the left, which takes the user back to the most recently visited page that makes sense for the user. For each usage, it can be programmed for a given page, making the component easier to reuse. On the right side of the Header, the current subject's name is assessing is displayed. The name is clickable and always leads the user back to the first page, the Uploading page.

Task List



Figure 4.8.: The task list component where three tasks have been assessed.

The Task list is a component that shows the task number and their accompanying task title on a list, which can be seen in [Figure 4.8](#). It gives the user an overview of the total

4. Implementation

amount of tasks. A symbol is placed on the left side of each task to indicate the task's status. If there is a circle outline, the user has not begun assessing the task. If it is a yellow circle with a dot, the user has begun assessing the task, but it is not completed. Lastly, if there is a green checked circle, the task is fully assessed. The tasks are clickable, leading the user to assess the chosen task. Depending on assessment status, it will lead to different pages within the assessment part of the web application.

In earlier stages of the web application development, the task list had a different appearance. It changed due to feedback from user testing, later described in [section 4.7](#), and the supervisor of this thesis. Originally, each task was represented by a button containing the task number. They were placed row-wise on the page and coloured according to their status (not begun, partially finished, or finished). As they did not contain the task title, it could be hard for the user to remember the topic of the tasks, and to improve this, the Task list was developed. However, the colour indication regarding status was kept due to feedback from the user tests.

Grade Component



Figure 4.9.: The Grade component where the grade is set to a "B".

The Grade component allows the user to set a grade ranging from letters "F" to "A". It can be seen in [Figure 4.9](#). The user only has to consider a known grading scale rather than possibly various points and max scores for different tasks. It might go on account of granularity, whereas there are only six levels, but it can be considered user friendly. For example, a similar score component is used in Inspera Assessment when assessing answers¹⁶. Therefore, many users may be familiar with that scoring function. For simplicity, the grades are displayed as letters to users but stored as numbers with a 0.2 interval for calculation purposes in the back-end.

The Grade component has evolved through user tests, which will be further described in [section 4.7](#), and feedback from the supervisor of the project. It was initially a dropdown menu where the user could give *points* in the range from zero to the maximum score of the task. The dropdown menu would have as many options as there were points to give. Feedback clarified that the dropdown menu was cumbersome and needed an easier-to-use, fewer-clicks solution. Then the Grade component was created based on Inspera Assessment's solution.

¹⁶https://wikihost.uib.no/sawiki/index.php/Grading_in_Inspira

4. Implementation

Text Boxes

In the web application, the answers from the candidates and their grades are displayed. Depending on the current page, selective information should be presented in a way that gives the user the most value. Therefore, different text boxes have been developed to serve each purpose.



Figure 4.10.: The Answer text box. The answer is within a scrollable window.

The Answer text box shows a candidate's answer and makes it possible to give each answer a grade, as seen in [Figure 4.10](#). The answers are retrieved from `localStorage`, and when each task is graded, the assessment is uploaded to `localStorage` for further use. Most of the space in the Answer text box is dedicated to displaying the answer, and there is a scrolling functionality that is enabled if the answers are longer than the Answer text box can display. Also, the text is formatted to improve readability. The score is set through the Grade component. Initially, there is no grade set, but the user can select an appropriate grade and change it. Within the Answer text box, the candidate id of the student submitting the answer also presented. Candidate ids are anonymous and therefore do not give any direct information about the answer or student. However, listing them gives the user control of the assessments and the knowledge that each user's answers are different and unique. Lastly, there is a flag in the top right corner of the Answer text box. Its purpose is to give an answer extra attention. The flag's status is saved with the grade and answer in `localStorage` to be used later.

4. Implementation



Figure 4.11.: The Approval text box. The answers have been assessed more than once and given different grades, so the box has an additional red outline and information on top of the box.

The Approval text box displays an answer with the related candidate id, grade and flag status. It is similar to the Answer text box, where the most significant difference is the text box length. In the Approval text box, the answer is displayed in total, meaning the box expands vertically in line with the length of the answer. Similar to the Answer text box, the answer text is formatted, and the top section contains the elements: candidate id, grade and flag status. However, here, an answer can be flagged initially, remembering its status from earlier. This can be used to locate specific tasks. Additionally, the given grade is automatically set, and if the user spots an error with one of the grades, they can change it instantly. Also, if the answer has been assessed twice due to the Consistency check explained in [subsection 3.2.4](#), and two unlike grades have been set, a warning and red outline occur to indicate inconsistency. When this is the case, no grade is initially set. This can be seen in [Figure 4.11](#). The warning persists until the user sets a new grade.

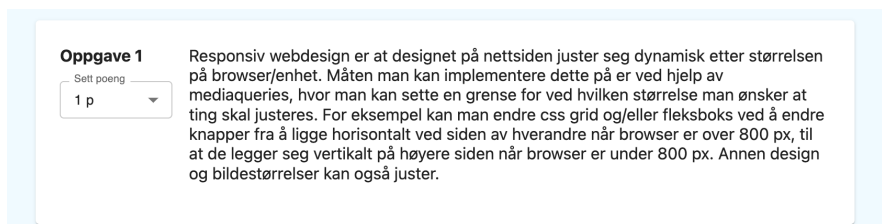


Figure 4.12.: The Candidate text box in the web application using an old iteration of the Grade component.

4. Implementation

The Candidate text box is highly similar to an earlier iteration of the Approval text box. As displayed in Figure 4.12, the Candidate text box has similar information as the Approval text box, but it is placed on the left side of the text box. Also, the earlier iteration of the Grade component, containing points rather than grades, is used. It was preserved for test purposes, so that one can ask test subjects their perceptions of giving points rather than grades in a user test. In the Candidate text box, the task number of the current task replaces the candidate id. That is because the Candidate text box is being used on a candidate page where the candidate id is given in the Header component.



- (a) The Dynamic grade text box contains a candidate id and the Grade component initially marked with their calculated grade. It is clickable, leading to the candidate page.
- (b) The Static grade text box contains a candidate id and their final calculated grade.

Figure 4.13.: The text boxes which contain candidate id and their total grade.

The Dynamic grade text box, which can be seen in Figure 4.13a, is a simplified version of the Answer and Approval text box. It contains a candidate id and the Grading component, where the grades are set initially. Also, the Dynamic grade text box has a grey hover effect on the boxes to indicate its clickability. Clicking the box leads the user to a page viewing all of the candidate's answers and corresponding grades, enabling a more holistic view.

The Static grade text box in Figure 4.13b is a simplified version of the Dynamic grade text box, where it displays a candidate id and the candidate's grade as text. Its sole purpose is to give information. A user cannot change anything in this text box.

Consistency Check

The Consistency check component is a card that gives information about the Consistency check, has an on and off switch, and an expandable panel with different options. These elements are presented to show the opportunities a user has and can thereby give insight during user tests. The component can be seen in Figure 4.14. It is meant to give the user control over the assessment process and the option to adjust the Consistency check to their own need. As mentioned, the web application is developed bearing in mind Norman (2002)'s principles. In addition to the principles covered by Material UI¹⁷, the seventh principle, constraints, is also in use. The Consistency check is a relatively advanced concept that is not directly linked to the assessment. That is why there is an expandable panel that hides the detailed options, and the panel is initially collapsed until

¹⁷<https://mui.com/>

4. Implementation

the user clicks the arrow on the bottom right side. In the expandable panel, a slider sets the percentage threshold of the answers the user is willing to re-assess. There are also multiple radio buttons where the user can choose how the answers that will be re-assess are selected.

Konsistenssjekk
Sjekken velger ut X antall av dine besvarelser som du skal revurdere.

Bruk konsistenssjekk: ^

Andel besvarelser som skal revurderes:

0% 20% 40% 60% 80% 100%

Velg type sjekk som ønskes:

Tilfeldig Like scores Ulike scores

Korrelasjon

Figure 4.14.: The Consistency check component expanded.

The Consistency check algorithms are described in [subsection 3.2.4](#). The algorithms that were implemented were based on selecting answers in the following ways:

- Randomly
- Based on similar grades
- Based on outlier grades
- Based on correlation between answer length and given grade

These were the algorithms that could give the most value in the shortest time. The default algorithm in the web application is selecting answers based on the correlation between answer length and grade. Within its batch, an answer is selected if it is the longest answer with a high grade. It was chosen because it is straightforward to understand, manageable to test, and deterministic given the right conditions.

4. Implementation

View Button

The View button lets the user choose from three different view options. The options represent how many of the Answers text boxes are displayed at once. It is possible to show two, four and six Answer text boxes simultaneously. Changing the number of Answer text boxes displayed simultaneously depends on personal preference, task types, screen size, and answer length. Displaying many answers at once can increase efficiency but may lead to cognitive overflow.

Grouping Box

The Grouping box is a dropdown menu that allows users to choose which answers are displayed together. In that way, the answers are grouped with somewhat similar answers. There are various sorting algorithms that can be used for grouping. They can value assorted factors of the answer, which places the answers various in orders. The different algorithms discussed are explained in [subsection 3.2.3](#), and the following were implemented:

- Random
- Candidate id
- Text length, long to short
- Text length, short to long

Progress Bar

The Progress bar is a line that shows how close the user is to complete the assessment of all answers in a task. A count goes up and down depending on the user's progress. The light background is filled with a darker fill stepwise as the progress proceeds. The darker fill represents count, which is the percentage of finished assessments, and the light background represents 100%. Both the count, meaning the darker fill, and the total count, meaning the light background, are dynamic.

4.6. The Web Application

The web application is the product created to enable the concepts. It puts them in a holistic evaluation process, from importing and viewing the results to assessing and exporting them. Then, it is easier for markers to get a realistic view of the concept in a test setting, making it easier to get more comprehensive opinions to address RQ3.

The web application consists of three main parts, which reflect the parts described in [section 4.4](#). The parts are Initialising the assessment, Assessing and Finalising the assessment. [Figure 4.15](#) shows how the functionalities within the web application are connected. The first, top, left elements are covered by Initialising the assessment. The elements within the dotted squares "Perform assessment" and "Approve assessment" lies within the Assessing part. The downmost elements outside the dotted boxes are covered by Finalising the assessment. To cover the functionalities, the web application and its parts consist of different pages and functionalities using the technologies presented in [section 4.2](#). The pages are connected to each main part, and the flow between the pages can be seen in [Figure 4.16](#). Within the pages, there are components. Some of the critical components are the Text boxes, the Consistency check, and the Grouping box. All of them are located on the Assessment page, which contains the core of the concept; the Side-by-side comparison.

A well-designed implementation is essential for interviewees to focus on the critical parts of the product when evaluating the web application. It gives the interviewee fewer distractions and makes it easier to focus on the purpose of this project; to address the research questions in [section 1.2](#). To improve the front-end look, the web application is developed bearing [Norman \(2002\)](#) in mind, with a focus on the seven principles of interaction design. The Material UI library presented in [section 4.2](#) is, for example, used to give a continuous look and well thought out components. There are instantly signifiers, affordance and discoverability ([Norman, 2002](#)) due to Material UI's component design. Also, to ensure usability while developing at a high pace, a decision was made to focus purely on screen sizes larger than a small computer. Although the mobile-first design is getting increasingly significant¹⁸, the Side-by-side comparison requires a certain screen size to view multiple answers simultaneously, and the computer is a more common assessment tool than a smartphone. It was therefore seen as beneficial to prioritise other functionalities.

¹⁸<https://xd.adobe.com/ideas/process/ui-design/what-is-mobile-first-design/>

4. Implementation

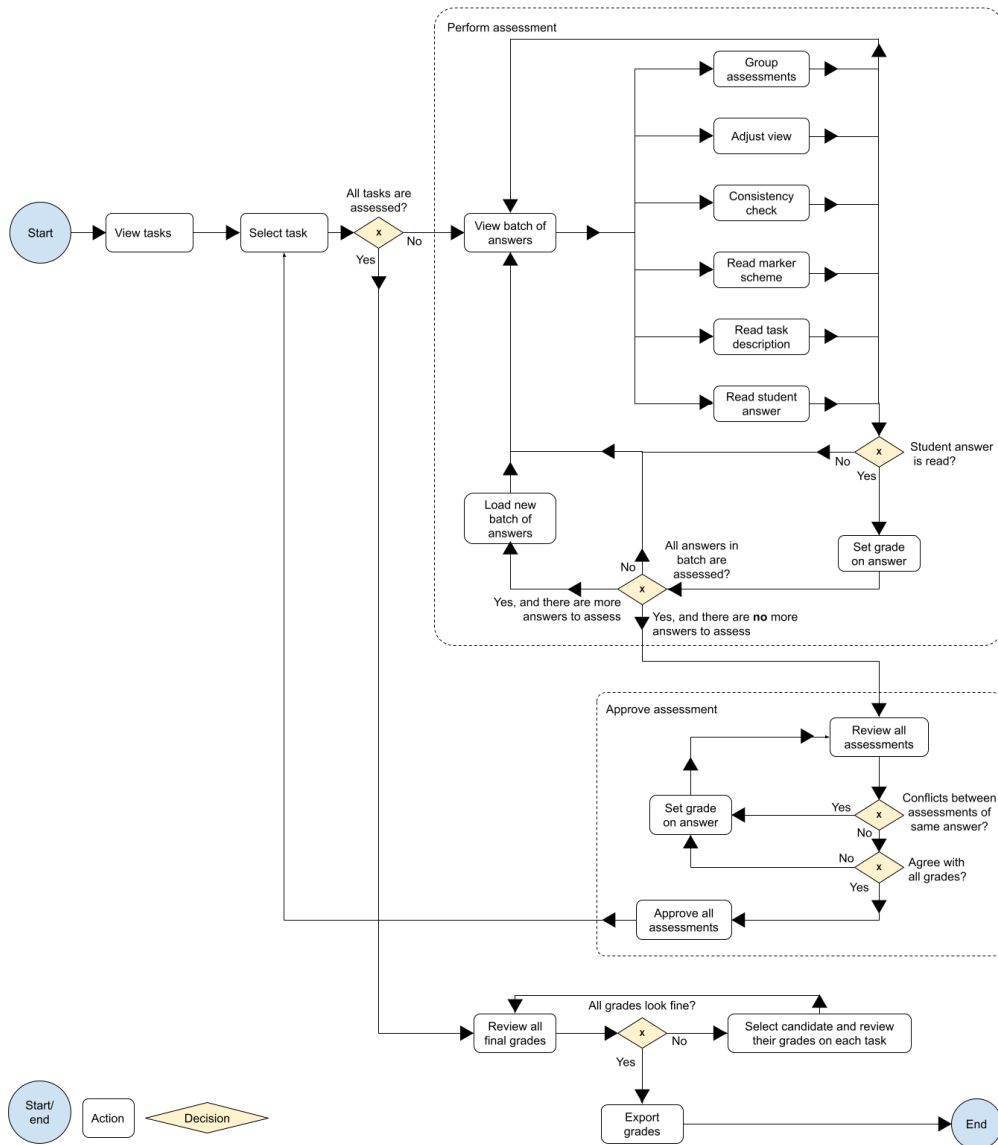


Figure 4.15.: The process diagram of the web application. It explains the different functionalities a user is presented with throughout the web application and how they are connected.

4. Implementation

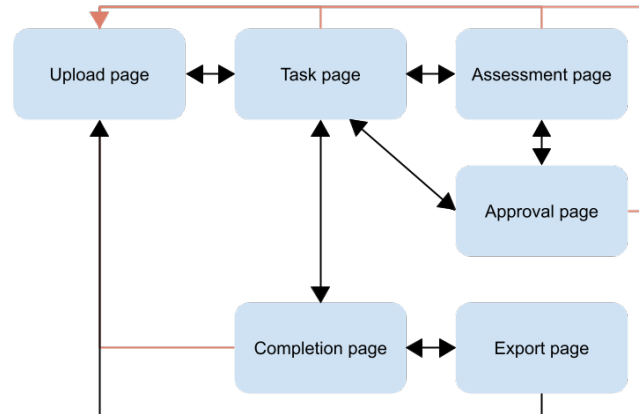


Figure 4.16.: The routing between pages in the web application. The black arrows represent routing between pages specific to the current page. The red arrows show every page's standard "home" link.

4.6.1. Initialising the Assessment

The first part of the web application is Initialising the assessment. It focuses on importing the results from an exam, processing them, saving them, and presenting the exam tasks to the user. Initialising the assessment consists of much work that is not directly visible to the user. Therefore, this part will only reflect the three first elements, start, View tasks, and Select task, in the process diagram in [Figure 4.15](#). There are two pages to fulfil these functionalities: the Upload and the Task page.

The Upload Page

The first page concentrates on uploading the data set to `localStorage` and processing the data for later usage. The system is designed for data exported from Inespera Assessment, so a prerequisite to using the system is having exam results in their format. Currently, the data set is retrieved from the project, and it is possible to see one set of exam results. The data is imported and processed by extracting essential information and uploaded to `localStorage`. After the initial "get method", the data is saved in `localStorage`, and then the web application communicates solely with the `localStorage`. The system is programmed to be ready for further upload features such as uploading multiple data sets, which the design from [section 1.4](#) shows.

4. Implementation

The Task Page

The second page focuses on displaying the exam tasks and giving the user an overview of the work ahead. The page can be seen in [Figure 4.17](#). It shows the Header component, the Task list component, and a button that can lead to the final pages. The Task list gives an overview of the total number of tasks and their progress in assessing them. When all of the tasks are assessed, the button at the bottom of the page routes the user to the Completion page and thereby views all grades. If some tasks have some answers left to be assessed, the button is grey and disabled.

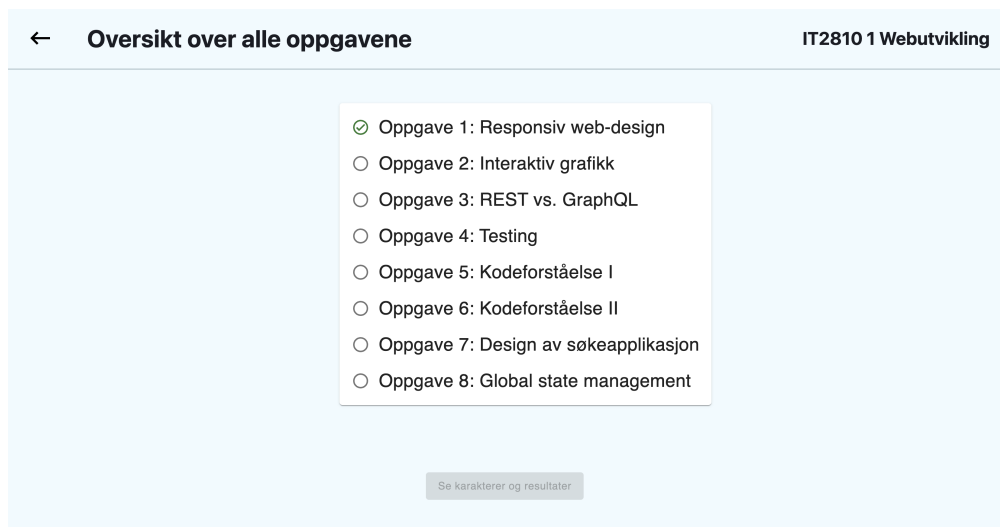


Figure 4.17.: The Task page. It shows the different tasks of the current exam.

4.6.2. Assessing

The assessment part of the web application is where the concepts come forward and the actual assessment of answers happens. It consists of the Assessment page and the Approval page, which combined has the purpose of assessing students' answers in an efficient, consistent and reliable way. As this thesis focuses on evaluation of high-stake tests, the user should be able to trust the system. The Side-by-side comparison is visible on the Assessment page, where multiple answers are displayed side by side for assessment. Also, the Consistency check is used on both the Assessment page and the Approval page.

4. Implementation

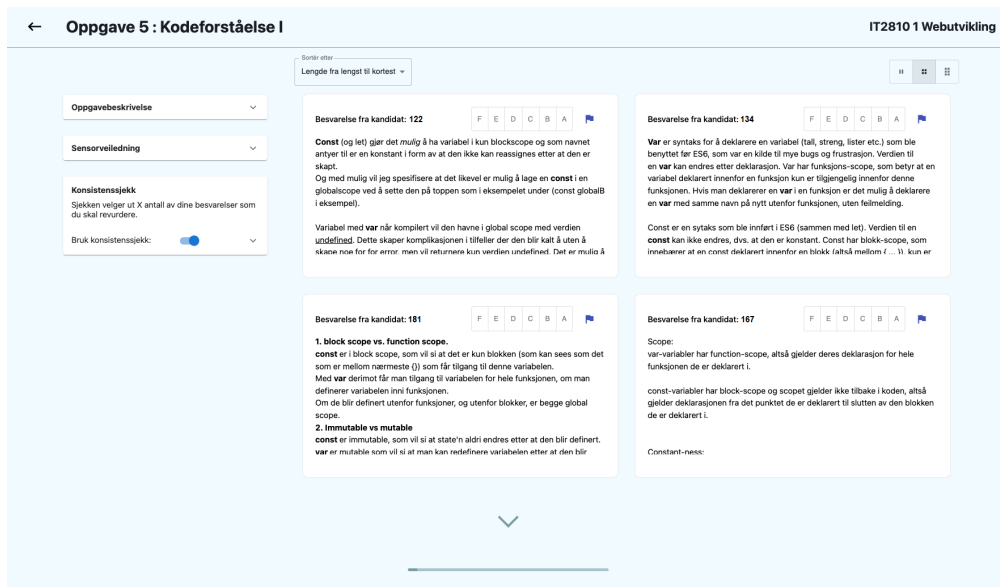


Figure 4.18.: The Assessment page. The user can assess candidates' answers side by side with a Consistency check.

The Assessment Page

After selecting a task on the Task page, the user is led to the Assessment page of the chosen task, which can be found in [Figure 4.18](#). The Assessment page consists of multiple components; the Answer text box, the Consistency check, the Grouping box, the View button, the Header, and the Progress bar. Also, there are expansion boxes containing the task description and the marker scheme, and on the lower part of the page, there are navigation buttons.

On the left side of the page, there are two smaller cards that expand. They have a headline and can expand to show more information on the subject, but the component is initially collapsed. The first expansion box contains the task description, and it has a header that reflects it. A similar functionality applies to the next expansion box, which contains the marker scheme. It enables mastery learning, which is explained in [subsection 2.1.4](#), where the user easily can use the marker scheme as a reference when assessing the answers. The expansion boxes keep their state when moving on to the next batch of answers to reduce unnecessary repeating clicks.

Underneath the expansion boxes is a Consistency check component. It provides the user with an understanding of the Consistency check and the option to modify the parameters. When the user hits the "next" button, they are presented with a new batch of answers, and in the background, an algorithm from the Consistency check is triggered. When triggered, the algorithm potentially extracts one answer to be re-assessed. It depends on

4. Implementation

if the answer meets the criteria and if the percentage of re-assessments is not exceeded. When the user has assessed every answer once, the answers extracted for a re-assessed are presented. Then the user can assess them a second time in another batch and with more experience. Additionally, the extraction of answers to re-assess depends on the selected grouping and the View button option. Both of these features affect the content of the current batch, which is the basis of the Consistency check algorithms. The View button option and the percentage threshold in the Consistency check affect the *number* of extracted answers. The chosen grouping and Consistency check algorithm affect *which* answers that are extracted.

On the centre of the page, the submissions from the candidates are displayed. To regulate the display of the submissions, one can click on the View-button and choose between displaying two, four or six submissions side by side. Additionally, one can vary which type of submissions are shown together by clicking the Grouping box placed to the left above the submissions. Then the answers can be grouped based on length, candidate id, or randomised.

At the bottom of the page, there is a Progress bar component, which shows the progress of the assessment on the current task. It was implemented as a result of one of the usability tests, which will be described in [section 4.7](#) where one finding was that they lacked a sense of holistic overview of the assessment process. The suggestion was to have a count or progress bar showing how many of the answers were assessed. Due to the Consistency check, the total number of answers to assess may increase. One way to solve this challenge is to adjust the total count throughout the assessment. However, the Progress bar would fill slower as more answers are extracted to be re-assessed, which might demotivate the user. Another way, which is implemented in the web application, is to assume that the percentage threshold set in the Consistency check will be the total number of answers to assess. That means the user can be finished before the Progress bar is filled, but it can give a strong enough indication for the user to be valuable.

When on the last batch, the "next" button is replaced with a button that says the user can finish the assessment of the current task. The button is disabled until all answers are assessed. When all answers are assessed, the assessments are saved, and the user moves on to the Approval page.

4. Implementation

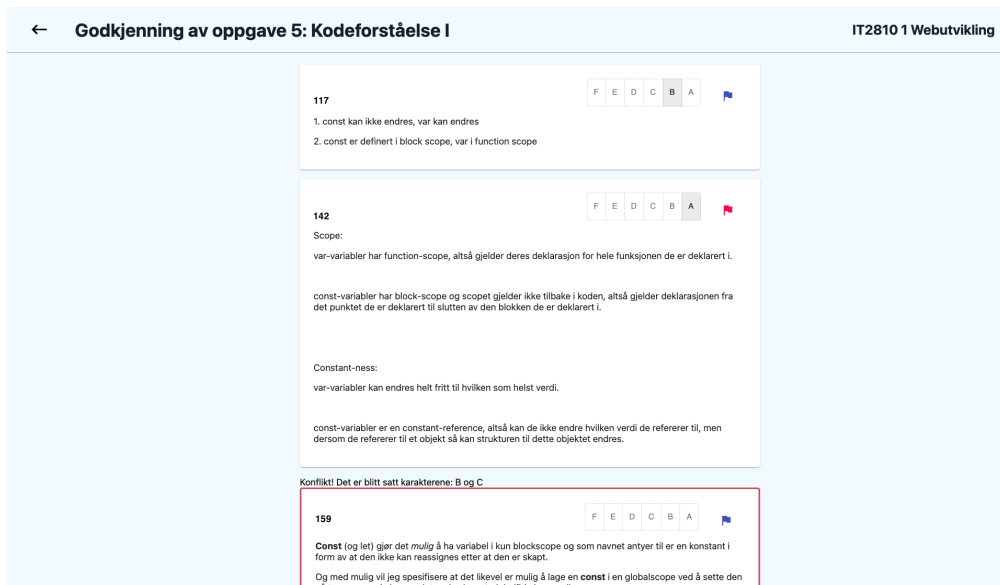


Figure 4.19.: The Approval page. The user can review and edit given grades on each answer and resolve potential conflicts where different grades were set to an answer.

The Approval Page

The Approval page's purpose is to ensure that the grades were set as intended and fix any contradicting grades. Every answer is displayed with their given grade in an Approval text box. As can be seen in Figure 4.19, the user can verify that the given grade is correct. The grades are stored in `localStorage` from the Assessment page, and re-assessed answers are saved as different assessment objects. Figure 4.20 shows a simplified example of the back-end data, where the answer of candidates 103 and 101 have been re-assessed. Candidate 103 got the same grade twice, meaning the grade is consistent. However, candidate 101's answer have gotten inconsistent assessments, which results in a conflict. If there is a conflict between two contradicting grades, the user must resolve the conflict by setting a grade. They can not move on from the Approval page before all conflicts are resolved. When going through the assessments, the user can choose how carefully they want to review and verify the grades. The flags and the conflict warnings can indicate where to look. These indications can minimise extra search for anything ambiguous and make the approval process more efficient, especially considering that the user already has gone through and assessed all of the answers. On the bottom part of the page, a button lets the user approve all of the assessments. The button becomes clickable when all answers have a corresponding grade. Clicking the button completes the assessment of the current task, triggers the green checked circle in the Task list, and returns the user to the Task page.

4. Implementation

```
[
  {assessmentId: 4801, candidateId: 109, score: A},
  {assessmentId: 2649, candidateId: 102, score: D},
  {assessmentId: 5492, candidateId: 103, score: C},
  {assessmentId: 6254, candidateId: 103, score: C},
  {assessmentId: 8572, candidateId: 107, score: E},
  {assessmentId: 3985, candidateId: 108, score: B},
  {assessmentId: 7141, candidateId: 101, score: D},
  {assessmentId: 5492, candidateId: 101, score: B},
  {assessmentId: 9348, candidateId: 104, score: B},
  ...
]
```

Figure 4.20.: Stripped example data stored in localStorage where all assessments are from the same task, and thereby have similar `answer-`, `taskNumber-`, and `maxPoints-` values. The coloured lines indicate the status of re-assessed answers as consistent (green) or inconsistent (red).

4.6.3. Finalising the Assessment

When all tasks are assessed, the web application allows the user to enter the third and final part of the concept. A button on the task page is made clickable, and the user can continue to the Completion page. There, they can see and edit the given grades for each student, see through all of one student's assessments simultaneously, and confirm the grades. The user can review and reassess tasks if needed. When completed, the user is routed to the Exportation page, where they can view and download the grades given to each student and finish and exit the assessment of the current exam, which routes back to the Upload page. The Completion page and the Exportation page have similar qualities besides the fact that the Completion page has editable grades with the option to see a holistic view of a student, and the Exportation page has static grades and an export function.

The Completion Page

The Completion page gives the user an overview of the grades, the chance to change the grades, and the option to see and edit grades on all the assessed tasks from the perspective of one candidate. The page, which can be seen in [Figure 4.21](#), consists of a Header component, Dynamic grade text box components, and an approval button. A screenshot of the Completion page can be seen in [Figure 4.21](#). There are Dynamic grade text boxes displaying candidates' ids and corresponding grades. The initial state of the Grade component is the candidate's calculated grade. The grades are calculated based on the assessments of every answer. For every student, their total grade is calculated based on each answer's given grades and possible maximum scores. The user can override

4. Implementation

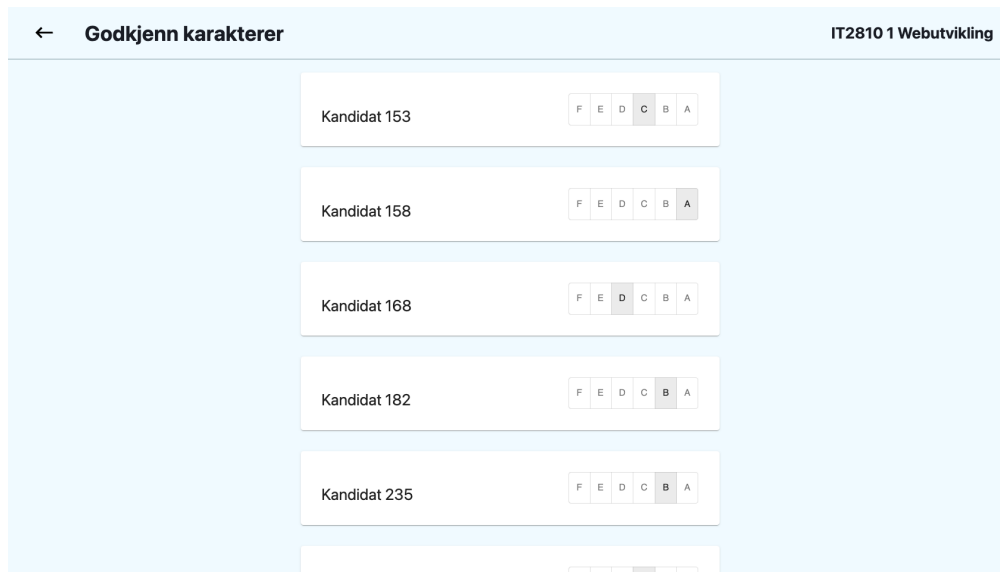


Figure 4.21.: The Completion page. The user can view and edit the final grades of every candidate as well as select one candidate and review all their assessments.

the grade generated from the sum of the grades given on each task if they think it makes sense regarding the holistic view. Additionally, the user can click on a Dynamic grade text box to view the assessment of every answer the student has delivered. These assessments are displayed in Candidate text boxes. It can give a holistic view of the student, which might change the impression of the grade. The grades given on the answers are mutable and will thereby influence the grades shown in the Completion page overview. Lastly, the user can confirm all final grades by clicking the button at the bottom of the Completion page and be routed to the Exportation page.

The Exportation Page

When the user is finished reviewing grades, they are sent to the Exportation page, which is similar to the Completion page. One difference is that the cards are Static grade text boxes. As the grades are immutable, the user can get a final look at the grades before exporting the results. In the web application, actual exportation functionalities are not implemented, but the button gives feedback to the user saying that the grades are downloaded. This feedback is in place to give the user a more extensive user experience and lets the user focus on the more essential parts of the web application, meaning where the concept is enabled.

4.7. User Testing

An essential part of the development of artefacts is to test. Testing is used to uncover errors and defects of the system to ensure that the artefacts are working as expected. It is also done to improve the system's quality and performance to develop an artefact that accommodates the demands of the target group.

Guerilla Testing

User testing was performed during the implementation iterations. One type of test was guerilla testing. Guerilla testing is a fast and informal procedure to test ideas, get feedback, and potentially detect user experience problems. The test can be carried out anywhere and usually lasts around ten minutes¹⁹.

The guerilla testing conducted was in association with the design of the Grading component. It was uncertain how the scoring should take place. The first suggestion was to use a dropdown menu with all the possible scores for the task from zero to maximum score. This was presented during a guidance and feedback session with the supervisor. The supervisor suggested that a different solution with fewer clicks could be more suitable. The dropdown menu required two clicks to set a score; one to open the menu and one to select a score. After this feedback, new design suggestions for the grading components were created, evaluated and reformed. These suggestions were then tested with guerilla testing.

The test was carried out on three Computer Science students who were recruited through a network. There were two paper sketches of the component. One was the original dropdown menu described above. Another solution was a slider with the same range as the dropdown menu. The test subjects were then presented with each option one by one, and then they were asked questions about what they liked or disliked with the option. At last, when all options had been presented, they were asked to answer which option they preferred and why.

No clear and distinct solution was preferred over the other amongst the test subjects. Hence, the design resulted in a component inspired by Inopera Assessment's default scoring system. This is due to it being a simple component with few clicks and familiar for markers. The latter was desirable so that the component is neutral and does not take any undesirable focus from the concepts of this thesis.

¹⁹<https://xd.adobe.com/ideas/process/user-testing/hallway-usability-test-guerrilla-testing/>

4. Implementation

Usability Testing

Usability testing of the web application was also performed further out in the development process. This was when components of the system could be considered sufficiently usable and functional for evaluation or to reveal defects that could weaken the user experience.

Early in the planning phase, paper prototypes and a Figma prototype were created. The Figma prototype was created after having design workshops where designs were formed, discussed and evaluated. It was sporadically presented to the supervisor for evaluation and testing. The supervisor is a part of the target group, and it was therefore believed that any received feedback from the supervisor regarding design decisions could give additional insight.

Two usability tests were conducted during the implementation phase. The purpose was to discover any errors or minor bugs in the system that could confuse the user and distract them from concepts. If such errors are present, one may not get to test what one actually wants to test and evaluate. The tests were also conducted to retrieve feedback and confirmation that the components and interface worked as desired. It was regarded as essential to have a professional and well-functioning system for the demonstration in the interviews.

The First Iteration

The first tests were carried out when the Assessment page was considered functional enough to illustrate the Side-by-side comparison. [Figure 4.22](#) shows a screenshot of the page at that time. Two test subjects were recruited through a network by personal invitation. There were no specific criteria as the purpose at this stage was to weed out GUI errors. However, it was believed that test subjects who knew about the project and concept might be influenced and biased. Hence, the two subjects recruited were subjects who were unfamiliar with the concept.

During the execution, one person guided the test, and the other was an observer responsible for taking notes. The test execution started with initially giving an introduction to the project, the purpose of the test, and the user context. The subjects were asked to imagine that they were a marker who assess exams. They were also informed that the answers displayed were test answers, and it was not expected that they should actually assess them. The subjects were told to think out loud and that they could ask questions whenever they wanted. Then they were presented with a set of tasks. After finishing the tasks, some more questions were asked, and the subjects could speak freely and add any comments.

4. Implementation

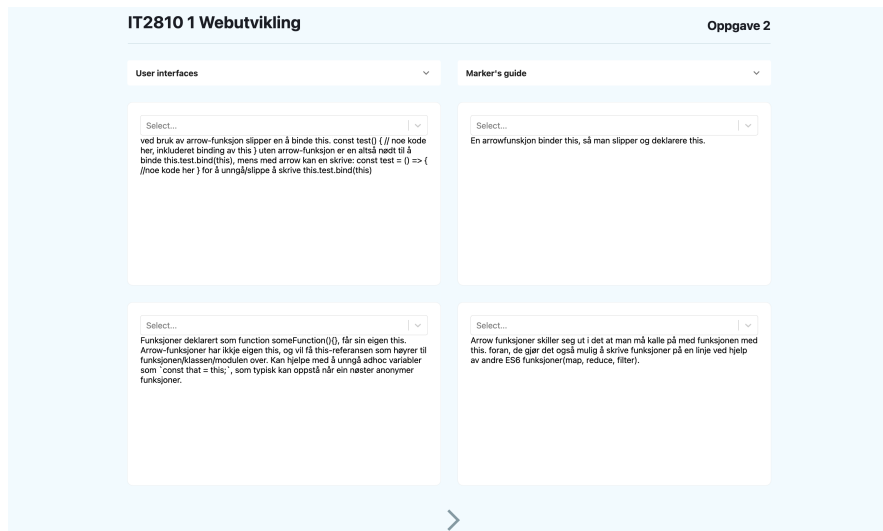


Figure 4.22.: A screenshot of the implementation of the Assessment page at the stage of the first usability tests.

The key takeaways from the first usability tests were:

- It must be clear that there are submissions from different students
- There is some uncertainty about what the dropdown menu is (referring to an early iteration of the Grade component)
- The task title, task description and marker scheme must differ more from the answers
- It was challenging to know the progression of the assessment
- The function of the arrow (the next-button) was unclear

A workshop was held to discuss the feedback and how to accommodate it. To clarify that there is presentation of several student answers, it was decided to include the student's candidate id within the box of their answer. A solution for the dropdown menu was to change the placeholder text to something more informative like "Set points", which is the action performed when clicking that component. Adjusting the placement of the task title, task description, and marker scheme was considered sufficient when separating the components from the students' submissions. The task description and marker scheme would be placed to the left of the students' answers. Also, the task title would switch place with the course title in the Header component and include more information than just the task number. Additionally, it was decided to change the size of the cards with this information to distinguish them more from the student submissions. Regarding the

4. Implementation

progression, a progress bar was suggested, as having, for instance, $x/100$ submissions would be troubling with the intended Consistency check described in subsection 3.2.4. For the arrow representing "next batch", it was decided to change the direction of the arrow downwards to see if that could have any effect. A suggestion was also to have an explaining text saying something like "More Answers". These suggestions for improvements resulted in new issues to create for the project on GitHub. After this, a new prioritisation of issues for further implementation was carried out.

The Second Iteration

The purpose, context and set-up for the second round of usability tests were similar to the first round. The aim was mainly to uncover usability errors that would weaken the user experience. Accordingly, it was not necessary to get test subjects representing the target group. There were two new test subjects recruited. An announcement was sent out to a group of students at the Department of Computer Science at the Faculty of Information Technology and Electrical Engineering at NTNU²⁰. Two students volunteered as test subjects.

At this stage, the implemented pages were the Upload page, the Task page and further development of the Assessment page. The Consistency check and the View button had been added to the Assessment page. See Figure 4.23 and Figure 4.24 for screenshots of the Task and Assessment page.



Figure 4.23.: A screenshot of the implementation of the Task page at the stage of the second usability tests.

²⁰<https://www.ntnu.edu/idi>

4. Implementation

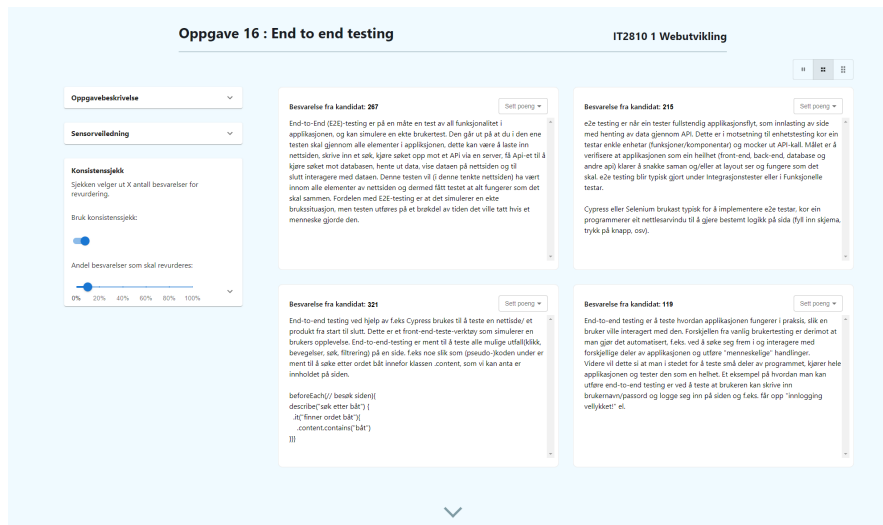


Figure 4.24.: A screenshot of the implementation of the Assessment page at the stage of the second usability tests.

The setup for the second iteration of usability tests was similar to the first iteration. The outcome of the second iteration resulted in these key takeaways:

- The grey coloured buttons on the Task page indicated that they were unavailable or disabled
- The purpose of the Consistency check component was uncertain
- A type of indicator that shows the progression is needed
- More navigation and button options are desirable
- When changing the view with the View button, the size of the text boxes should be adjusted

One thing to notice is that most of the feedback from the first iteration of tests was not mentioned this time. That means that the developed improvements were successful. This shows that usability testing is valuable and gives insights.

After this round of tests, a workshop was also held to work out the feedback mentioned above. Regarding the first comment, the idea with the grey buttons was to express that a marker had finished assessing those tasks. They were grey, so the marker did not have to focus and deal with those as the assessments were complete. In retrospect, it is understandable that the test subjects experienced the grey buttons as disabled buttons. This is because, in design, disabled buttons are often with lower contrast and coloured

4. Implementation

grey²¹. The buttons did not indicate completion as intended but purely unavailability. A suggestion from a user was to colour them green, but it is believed that green buttons would not necessarily mean that something is finished. Though, the conclusion after this workshop was to use the suggestion from the user as no other option came to mind at that stage of time. However, the final result presented in [section 4.5](#) differs from this suggestion after a revision with the supervisor.

For the second comment concerning the Consistency check, it was not surprising that the test subjects would struggle with understanding the component's purpose. The test subjects have no experience in marking exams. However, one can assume that a marker would not have the same impressions of the Consistency check. Still, some minor design adjustments were suggested. This was to change the information text to be more precise and descriptive. The other change was to move the switch to be aligned with the "Bruk konsistenssjekk" text, as feedback revealed that it was unclear what the switch was related to.

Regarding progression, this was feedback that was mentioned in the first usability tests. The proposed solution with a progress bar existed but was unfortunately not manageable to implement in time for the second iteration of usability tests. Also, the following comment regarding navigation was expected. The current views of the pages offer navigation to the Upload page or the next page if one is on the assessment page. It was natural that the users wanted more buttons and options for navigating back and forth through the different pages. Hence, there were created issues on GitHub for back arrows for routing between pages and between batches of answers on the Assessment page.

The last comment regarding the View button is meaningful. It is not surprising that one would expect the size of the text boxes to adjust when changing how many student submissions to view. A user would want to change the view due to the size of the content of the submission. This means that a user would probably change from a view of 4 submissions to 2 submissions when some of them are too long to be displayed four at once. This would require much scrolling within the text boxes, which can be experienced as inconvenient.

As with the first usability tests, all of these analysed and reviewed suggestions were added as issues on GitHub. Here, they were compared and prioritised with current issues.

²¹<https://www.smashingmagazine.com/2021/08/frustrating-design-patterns-disabled-buttons/>

4.8. Limitations and Workarounds

During the implementation of the web application, there were some limitations, and hence some workarounds had to be made. The data set retrieved from Inspira Assessment only consisted of the student submissions and details around them. In the web application, it was desirable to have the task description and marker scheme on display. As the task description is available in Inspira Assessment during exams, it was reasonable to assume it could be downloaded and made available. After contacting NTNU IT help desk, it was found out that the task descriptions could be exported as Question Test Interoperability(QTI). Though, when this information was received, a workaround was already made. The task description was available on external PDFs and had been manually written to JSON format to fit the web application. The same workaround was applied to the marker scheme. As the marker scheme is often an external document, it was expected early that it was not in the same format as the Inspira Assessment data set and that some solution had to be made to take this into account. However, due to the project's scope, it was not prioritised.

Another limitation was that the Side-by-side comparison presupposes short text answer tasks in the submissions. Hence, the design decisions and the development of the web application were made based on that premise. The design is not suitable for longer text answers, but a quick solution was to add scrolling in the text boxes. Thus, it will be possible to have submissions with longer text answers, but it is not optimal in relation to design and the comparison of submissions. For answers that are not text – some answers may, for instance, consist of tables and figures – the solution was also to use the scrolling functionality.

The data set from Inspira Assessment includes formatting from the submissions in HTML code. When developing the display of these answers, the given HTML code was used to keep the students' composition in the answers. This was to ensure readability and get the presentation as similar as possible to the student's layout when submitting their exam tasks. On the other hand, this also meant that any written HTML code in the text answer would be formatted. However, this was considered rare and would only apply to some code subjects. Therefore a fix or a workaround was not made.

5. Evaluation

5.1. Interview Preparations and Processing

The purpose of having interviews is to get feedback from users and gain insight regarding the research topic. The goal is to get people in the target group's opinions, thoughts, and perspectives on the concept. It involves understanding how they usually assess, what they would like to improve, how they think it would work to utilise the concept, and which criteria are needed for the most practical usage.

There is a lot to consider when conducting interviews. They can be executed in different ways. In this project, they were qualitative, semi-structured interviews with people from the target group: markers in higher education who assess digital exams where the tasks preferably require short answers. Also, preparation is needed to ensure valuable and efficient interviews that are pleasant and worthwhile for the respondents. It includes planning how to approach the respondents, the interview flow and questions, and the setup routine before the interviews occur. Lastly, the results need to be processed and analysed for evaluation. The interviews would be conducted concerning RQ3, evaluating the Side-by-side comparison and the supportive concepts.

5.1.1. Method

The interviews were semi-structured, allowing a more natural flow of conversation with the interview subject. The questions were created before the interviews, and they should all be asked, but the interviewer can change the order. As [Oates \(2006\)](#) explains, "The interviewees are able to speak with more detail on the issues you raise, and introduce issues of their own.". This way, the responses could expand beyond the pre-setup questions, resulting in a broader understanding and feedback. Because the purpose of the interview is mostly "discovery", it fits with semi-structured interviews. It also fits that the research is qualitative to go even further in-depth on each interview subject, get a greater understanding, and let the interview subjects share information in their own words. However, as field notes were taken during the interviews, they might be coloured by subjective perception and, thereby, influence the research process. Therefore, in addition to field notes, the interviews were also recorded so that information could be fact-checked later.

5.1.2. Preperation

To conduct the interviews, preparation is needed. It includes creating an interview guide, preparing the web application for demonstration and testing, creating a plan for the walkthrough of the application, applying for permission from the Norwegian centre for research data (NSD) to store data gathered from the interviews, create a consent form for the interviewees to approve, mail people in the target group asking for an interview, and arrange time and rooms for the interviews to take place. The interview guide contains the timeline and overview of the interview. It includes the introduction of the project and the questions that need answering, which will be described in depth later. Preparing the web application for a demonstration was done, bearing in mind the walkthrough that would happen during the interviews. The walkthrough of the web application was constructed considering the research questions of this thesis. Additionally, a consent form regarding personal data needs to be in place to perform the interviews. NSD has a straightforward guide on how to make one and what to consider. An important principle when gathering data is to avoid unnecessary personal data. Therefore, all places to store data from interviews, devices that handle data, and questions regarding the interview objects were considered to minimise the amount of unnecessary data stored in an insecure way. The application in [Appendix C](#) regarding data gathering in research was sent to NSD and approved. The main points were that the interviewees have to accept that they are participating in an interview, the interviews would be audio recorded, and that anonymised data could be stored until the end of the project. The voice recordings of the interviews were made using Nettskjema Diktafon-app¹ which is an app that records audio, cryptates it and sends it to Nettskjema for safe storage. It is recommended by NTNU for data gathering². As this thesis is a part of NTNU's study program, and NTNU has a data processor agreement with the online cloud storage OneDrive³, documents were stored on NTNU's OneDrive space.

Then, people in the target group were contacted. It was done through mail, where the purpose of the interview was concisely explained. The interview subjects would preferably be markers who assess short answer exams digitally and preferably from different fields of academia. They were found by asking friends who study other subjects and systematically searching through schools' internet pages. After that, they were compared to find the most suitable candidates. By inviting a smaller, more targeted group, fewer markers were prone to spam, and there was a higher likelihood of acceptance. Due to a narrow range of contacts, all of the candidates who accepted the interview were markers at NTNU, and most of them were from the Department of Computer Science. It could give a limited view and feedback on the concept, but as people have different assessment routines regardless of the school department, it was seen as valuable nevertheless. When the interviewees had accepted, the time and place of the interview were arranged. Also,

¹<https://www.uio.no/tjenester/it/adm-app/nettskjema/hjelp/diktafon.html>

²<https://i.ntnu.no/wiki/-/wiki/Norsk/datainnsamling>

³<https://www.microsoft.com/nb-no/microsoft-365/onedrive/online-cloud-storage>

5. Evaluation

the consent form was sent to give the interviewees a heads up about the data gathering and spend less time on it during the interview.

5.1.3. Setup

Right before every interview, there was a preparation routine. It helped with getting into the right mindset and making sure everything was ready so that the interview could go smoothly. The questions and tasks were revisited, the web application was prepared, everything was charged, and roles were assigned, meaning determining who would hold the interview and who would take field notes. To the interview, an extra computer and phones were brought as well as a copy of the consent form that all respondents had to sign. As Oates (2006) states, "You should thoroughly check your equipment before starting", which confirms the routine.

As the interviews began, an effort was made to be clear and professional with the interviewees and make them feel comfortable and ready. The equipment was set up quickly, and meanwhile, an attempt to small talk was made, and the consent form they were previously given via mail was integrated into the conversation. It was to avoid an off-putting atmosphere and purposefully use the limited interview time. The seating arrangement was planned so that the interviewer and the interviewee were sitting 90 degrees toward each other to allow for comfortable interaction, as Oates (2006) suggests. It was also advantageous because both could easily see the screen when presenting the web application. Depending on the room, the person taking field notes sat next to the interviewer so that it was easier to be engaged and ask potential questions that could come up but were still in the "back seat". Then, the respondent could see them both, which is more relaxing for the respondent.

5.1.4. Interview Questions and Walkthrough

To conduct the interviews, a set of questions and a plan for a demonstration of the web application were created. The interview flow was split into different sections, which would give the wanted insight. Conducting interviews was done to evaluate the concept in conjunction with RQ3. Also, it was possible to receive additional background information which could supply RQ1.

Introduction

At the beginning of an interview, the respondent was asked about the consent form, if they had signed it and if they had any questions. The purpose was to let the interviewees know that they are in control of their own data and underline that the interview is

5. Evaluation

voluntary. They were again asked if it was okay that the interview was recorded, and after confirmation, the recording began. Thereafter, an outline of the interview was drawn to present the expectations of the interview regarding length and content and then get feedback from the interviewee to know if there is any consideration that should be taken.

Background

As Oates (2006) suggests, the interview began with two easy-to-answer questions to warm up the respondent. They covered the interview subject's background by asking if they had created or assessed exams and, if yes, what kind of exams. They were also asked what tools they used for assessment. The answers would give insight into the interviewees' experience and enlighten new and unknown territory regarding assessment types and tools.

The second subject of the background questions regarded the interviewees' initial thoughts on the concepts. It included their opinions on assessing answers more than one time, whether or not they had done it before, and on what scale they would be willing to do it. Additionally, there were questions concerning their experience with side by side assessment, task by task assessment, and how they prevent bias when grading students. The purpose of these questions was to uncover their experience with assessment techniques related to the concept and if they have any preliminary thoughts on the concept.

Walkthrough of the Web Application

The walkthrough tested the web application's functionalities by using the 2019 exam of IT2810 Web development as test data to assess. As none of the interview subjects was expected to have any prior knowledge of the subject, they were asked to imagine that they knew what score to give students as the purpose was to unveil unknown factors of the concept. Specific questions regarding the components were asked during the walkthrough, and the larger, more overall questions were saved until after finishing the walkthrough. The web application could be used for a demonstration or a use-case test for the interviewees, depending on the time available in the interview. Having a demonstration takes less time than having use-cases, but letting the interviewees try the web application can give more detailed and in-depth feedback. Regardless of having a demonstration or use-cases, the respondent sees the same functionalities and is taken through the entire web application.

The walkthrough of the web application began at the Upload page on the first part of the app. See [section 4.6](#) for a description of the web application. Then it moved to the Task page, where the user was given a task to assess. In this case, the user could be the

5. Evaluation

interview subject or the interviewer, depending on the type of walkthrough. The user selected the task and was asked to test different functionalities during the assessment of the current and following tasks. The tested functionalities were previewing the marker scheme, changing the grouping, trying to flag an answer, changing the display view, and understanding the Consistency check. The user also tested how to handle conflicts if an answer was assessed twice with two different grades. These functionalities, as described in [section 4.5](#), are the crucial parts of the web application as it directly reflects the concepts described in [section 3.2](#). That is why they were essential to visit during the walkthrough so that the interviewees could get a more detailed look, and it could give them more ammunition to tell their thoughts on the concept. Lastly, the user was asked to enter the final part of the web application, look through each student's grades, and export them. It is possible to select a candidate to review their whole set of answers, but it was optional for the user as it is not directly linked to the concept. The final step in the walkthrough completes the assessment process to give the user a natural ending to the walkthrough.

Perception of the Web Application

Directly after the walkthrough, the user was asked for their impression of the concept and how it may or may not align with their expectations. The first question after the walkthrough was relatively open because of the desire to get an unbiased opinion and give the interviewees freedom to tell with their own words what made the different impressions. As it could be unknown elements, it was important not to narrow their mindset. It allowed the respondent to lead the interview order and a chance to get a more in-depth perspective and further details.

Feedback on the Concepts

The respondents were asked about their perception of Side-by-side comparison after having seen it in an assessment tool. The purpose was to figure out if they were familiar with the concept, what their thoughts were, and to understand their perspectives and viewpoints of the approach. They were also asked if they thought Side-by-side comparison could lead to more or less bias and how it would affect efficiency in grading. Although it could be hard for the respondents to answer when not having tested the assessment tool in a realistic assessment situation, they could have valuable input due to their experience and reasoning. As the Side-by-side comparison is the main concept and related to RQ2 and RQ3, it was questioned directly and openly and supplemented with how it would affect consistency and efficiency. The other concepts, the Task-by-task assessment method, the Consistency check, and Grouping were also questioned. Lastly, the interviewees were asked what they thought about the web application as a whole in regards to consistency and efficiency and how it compared to their usual methods. It could give a holistic evaluation of the web application and how the components coincide.

5. Evaluation

As the interview came to an end, the respondents were asked if they had any other comments. Although it can feel like the other questions cover everything needed to evaluate the concept, the target group, which has much experience and knowledge on the subject, can have additional information that is difficult to foresee. This question opens up for unknown details and information regarding the concept and assessment in general.

5.1.5. Processing the Interviews

During the interviews, field notes were taken. As the interviews were recorded, there was no need to write everything exactly as they were said. Therefore, they contained condensed feedback from the interviewees and some timestamps so that it would be easier to find in the recordings later. Directly after having an interview, all key takeaways were written down to easier recall essential points when analysing the interviews later. It often contained a brief and concise version of the interviewee's perspective on the concept and potentially other relevant sidenotes.

When analysing the material from the interviews, the field notes and recordings were used. The recordings were not transcribed in their entirety as it is time-consuming and costly, and there was no need to have everything written for later analysis. As the goal of the interviews was to get different people in the target group's perspectives and thoughts regarding the concept, it was more meaningful to transcribe selective parts of the interviews. Therefore, the interviews were processed by reading the key takeaways, the field notes, and listening to the recording, either the entire recording or parts where the field notes recommend it. Then specific sentences or parts were transcribed to be used for quotes or comparison in the analysis.

5.2. Interview Results and Discussion

There were, in total, six respondents for the interviews. All of them were professors at NTNU, and more than half were from the Department of Computer Science. The interviews were conducted in Norwegian, but all quotations from the interviews are translated in the following results. Moving forward, all respondents will be referred to as males to preserve anonymity and textual flow.

The interviews were conducted to gain knowledge regarding markers' experience and their thoughts on the concepts. Different topics emerged through the interviews, and the respondents brought up different information to enlighten their experiences and perspectives on the matter. They had the opportunity to test the concepts through a test case, or they could watch a demonstration of the web application. Their thoughts are categorised by different topics emerging throughout the interviews.

5.2.1. Side-by-side Comparison

Most of the respondents had little to no experience assessing answers displayed side by side. All but one respondent had tried an assessment method close to the Side-by-side comparison. However, during the course of the interviews, the respondents expressed interest in the solution, and many were excited to explore further possibilities.

A Similar, Self-Made Solution

One respondent mentioned that he used a self-made program to improve the assessment process in one of his programming subjects. The program simultaneously presents the marker scheme, the task description, and the code submitted by the candidate. *"The markers say that it is the most important tool. Both regarding saving time and keeping track in the assessment process."* It indicates that the assessment is more efficient when the marker quickly can compare the answer with the marker scheme and the handed out task. Additionally, it implies that the current tools are not sufficient. The respondent also stated that it can be problematic to compare tasks if they do not have a strict setup. With freely written answers, there can be many variations on solving the code problem, and it is, therefore, harder to compare the solutions with the marker scheme as it only represents *one* solution. Additionally, the fact that they found it valuable to create a program shows that the assessment program they are required to use, Inspera Assessment, does not have the desired functionality for this kind of code task. It indicates a need for improvement.

Formal Claims

One interview subject said that a prerequisite for using the Side-by-side comparison is that it has to be for a lower-level subject and that one has to create exam tasks with formal claims that suit the method. It can concern how the answers should be structured, the answer length, or how to write the answer. Then, it is easier to compare the results, and one can quickly extract valuable information. Setting formal claims and designing the tasks to fit the assessment method is more time-consuming when creating the exam, but it can reduce time when assessing the answers later. This balancing concerns efficiency, which is further described later. Several stated that the method is suitable for independent tasks, but sometimes an exam needs to contain dependent tasks. They can be hard to compare when assessing as they can have different starting points and conditions for each candidate. Two respondents mentioned the challenge. One suggested having tasks that tell the candidate to assume that the previous sub-task gave specific results they then can use further. That way, every candidate has the same starting point in each sub-task, removing some of the dependency. The other proposed

5. Evaluation

showing the sub-task of one candidate side by side rather than the standard setup from the concept where different candidates answer the same task. This way, it can be easier to see the entire flow of a task, and one removes the problem of comparing answers with different conditions. The latter idea shows another way to utilise the concept, which can be an idea for further development.

Display Content

Another requirement to using Side-by-side comparison was said to be that the simultaneously displayed answers need some relation. A respondent said that the method's usage depended on which answers were set side by side. This resonates with the Grouping concept, which decides the answers that are displayed together. Some respondents mentioned the advantage of it being easier to give similar answers the same score. However, if there is not enough similarity, the marker may use unnecessary time trying to find a connection between unrelated answers.

One requirement that two interview subjects noted was that the Side-by-side comparison requires wide or multiple screens. It is necessary to have enough space to view multiple answers simultaneously, especially if the answers are longer than a sentence. Therefore, it should balance screen size, answer length, and the number of answers displayed. When having short answers, there is more space on the screen to view more of the answers, and then it is easier to get an overview. Also, it is less information to comprehend and keep in mind at once because the answers are short and precise. This can reduce cognitive overload. Another respondent mentioned cognitive overload, saying that it can be too much to see four answers at once in addition to the marker scheme. He thought it might be enough to see just one answer and the marker scheme side by side. Another interviewee said he had to look at one answer at a time, regardless of how many answers were displayed simultaneously. However, he also thought that it was nice to have the possibility to look back at previously assessed answers quickly. He expressed that it gets more efficient when more answers are displayed simultaneously, indicating little concern with cognitive overload.

Referencing

One respondent was mainly critical of the concept. He had concerns regarding the referencing and comparison when assessing using the Side-by-side comparison. He explained that *"I feel sceptic to viewing many [answers] simultaneously because I think that you tend to rank them. Therefore, it might affect what kind of grade you set. I do not know, but I have a feeling about it."* Two submissions may be on the same level, but they can possibly be given different scores due to comparing and ranking. As described in [subsection 2.1.3](#), students value fairness in evaluations and, therefore, the comparison

5. Evaluation

might be questionable. As the Side-by-side comparison tries to explore how it can uphold and support the consistency in grading, this point needs attention. On the other hand, another interview subject saw the ranking of answers as a possibility. The different approaches show that enabling comparison by simultaneously showing multiple answers can lead to biased assessments, but it can also open new possibilities.

According to an interviewee, an advantage of the Side-by-side comparison is that it gives the marker a reference when assessing the first task. It can make it easier to get an overview of the level the candidates are at and hopefully make it less likely that the scores need adjusting. Another respondent points out that *"You will read one answer at the time, but you'll have the chance to compare it with the last one you looked at also, or the two last answers. It is an easy way to compare."* Both respondents communicate that by using the Side-by-side comparison, one can use other answers to realign expectations in an answer.

Interest in the concept

Three respondents were positive about the Side-by-side comparison. One of them had experience with assessing side by side and reasoned that he actively created exam tasks that could be answered on maximum one page, so that he could perform a side by side comparison. He argued that this would ensure consistency. Several said that the method looks efficient, reasonable, and helpful. One respondent said that *"I think it has value, you know, regarding fairness."*, and another said that *"It looks smart to have multiple [answers] viewed at once, so you can do it more efficiently because it takes very much time to assess."* However, they also said that the method needed to be further tested. They saw varying degrees of potential in the method and liked that there is an effort to support the assessment process, but as they had not tested it, they could not say anything with certainty. As one interview subject said, *"Would Inspira maybe be interested in doing something similar themselves? [the Side-by-side comparison]"*.

5.2.2. Comparison of Candidates

A topic that occurred in multiple interviews was the comparison of candidates. One respondent stated that *"Comparison is inevitable. You can not get around the fact that it is in relation to population."* It means that if a class is homogeneous, for instance, if all students are outstanding and get an "A" as a grade, one still has to adjust and spread the grades because everyone is not exceptional. Another respondent shared that *"Statistically speaking, there are always some out of 1000 students who get an "A"."* The same respondent expressed that it is more likely that the difficulty level on the exam is wrong than none of the 1000 students getting an "A". A situation like this can be reason to adjust the grade scale, and is also an explanation for some respondents seeing

5. Evaluation

comparison as essential in the assessment process. Also, in regards to fairness, one can view comparison as a technique to ensure fairness through consistency. When comparing, one can assure that those candidates who have, for instance, similar mistakes get the same amount of deduction, which again results in consistency and fairness.

However, there was some scepticism regarding comparison. As mentioned previously, one respondent meant that it would encourage ranking and referencing, which he regarded as a disadvantage. Another interview subject reflected on the outcome and significance of which submissions were presented together or consecutively. He wondered what effect it could have assessing an answer after a much stronger or weaker answer. It shows that some markers are concerned with the possibility of biased assessments when comparing.

Compatible Fields of Study

One respondent expressed that the main function of the web application is comparison. He argued that comparison is dependent on the field of the course. The respondent stated that *"It is fairer to compare technical subjects. That's what's fine with technical subjects."* In STEM, things are often objective. There is usually something right or wrong. When this is the case, it makes sense to compare because one can break things down into true or false. In social studies and similar, the answers are not necessarily comparable because concrete facts may not be the criteria.

5.2.3. Task-by-task Assessment Method

Several of the respondents had experience assessing task by task. Their currently used assessment system, Inpera Assessment, supports this approach, which some respondents considered functional. Some interviewees shared that they used to assess task by task before digital assessment support systems were introduced. The evaluation approach has both advantages and disadvantages, where the respondents brought up efficiency, consistency, and requirements for designing tasks.

The Effects of the Method

From the interviews, one knows that assessing exams task by task is actively used by markers today. There are benefits to using this method that is valuable for both the markers and the students. Several respondents shared that they valued efficiency. They meant that it is more efficient than assessing one candidate at a time because one considers the same task consecutively rather than a variety of tasks. One claimed that it is easier because one remembers the solution for the task. Then, one only has to focus on one topic at a time and does not have to scroll back and forth between different topics and

5. Evaluation

solutions. This results in less context switching and will then quicken the process. On the other hand, one respondent mentioned that it was boring to perform assessments task wise. A consequence of boredom can be that one gets tired, leading to less efficiency and more biased assessments. The theory described in [subsection 2.1.5](#) substantiates this. Hence one has to find a balance. Markers have different preferences, and they have to consider which work method suits them.

For students, fairness in the assessments can be considered the most remarkable outcome of the task by task method. When multiple markers are involved in the assessment process, and the markers assess all answers to a given task, they only have to agree with themselves. When they have set a basis for the assessment, they can continue to follow it throughout the assessment, making the assessments consistent. One respondent stated that assessing task by task *"Is very useful because it gives consistency."* Another interview subject agreed with this perspective, highlighting fairness and consistency as the advantages of this approach. By assessing task by task, the same marker somewhat assesses every student. If one has a strict and less strict marker, everyone gets the same marker for the same tasks, which is considered fair. If the markers assess by candidate, they have to coordinate between themselves to ensure that they emphasise elements in the answers equally. An interviewee mentioned that markers can differ in average points, which means all students are not assessed on the same basis. It can therefore be considered unfair when candidates get different markers. Additionally, one respondent reasoned that assessing task by task was done to compare the submissions to set the level for assessment. This was important to ensure consistency in the evaluations made.

External Factors

There could also be external factors that decide the chosen assessment approach. Some markers may get paid for each candidate they assess. Therefore, for practical reasons, it can be more convenient to distribute a batch of candidates for which each marker is responsible. Each marker could choose to assess task by task within their batch, but one does not get the gain where the same marker assesses all students.

5.2.4. Consistency in Assessments

Preparations

Two respondents emphasised preparatory work as essential to ensure consistency. One described that he tried to create a precise marker scheme. This applies especially to subjects with multiple markers due to different strictness amongst markers. Thus, a precise marker scheme will help achieve consistency across the markers. However, there is a limit on how long a marker scheme can be, and there may be outcomes one cannot foresee

5. Evaluation

in advance. The other respondent shared that he sets precise and formal requirements for the submissions. This concerns how the task is formulated and how the student's answer is formalised. As mentioned previously, a formal claim was for the answer length to be a maximum of one page, so he could compare the submissions side by side because he meant that this resulted in consistency.

The Consistency Check

One respondent was fond of re-assessment as a consistency check but mentioned that by experience, he had not corrected anything twice other than maybe the first 20 submissions. Another respondent also supported this. He expressed that if anything was going to be re-assessed, it was evident that this had to be the first assessments because these may have had a different baseline when assessed. Regarding the type of consistency checks, several mentioned that other correlations could be interesting. One said that "*The usual thing is that we like concise answers, so the shorter the answer, the higher the probability of a better grade.*". Therefore, a correlation between shorter answers and high scores could be relevant.

One interview subject mentioned that the Consistency check could be more helpful after discussing the grade limit. Then, it would be relevant to re-assess the submissions that are close to the limit. Several respondents mentioned that this is something they do as a procedure to ensure consistency. They check if a candidate should get one more point to pass or if candidates with similar knowledge level receive the same grade. One mentioned that it would be helpful to get a list of everyone very close to the grade limit. However, the distance between candidates near the grade limit can not be too great because there would be too many submissions to check for consistency.

A critique of the current Consistency check is that it only gains a small proportion of the students, and the consequences should be more significant. One respondent valued that one is asked to re-assess and check if one still agrees or disagrees with the given score. However, since the Consistency check only chooses several submissions, only the students of the selected submissions will be affected. A respondent mentioned that "*It [the Consistency check] should have larger consequences.*". However, one can not have all submissions on consistency check because of time limits and workload. It would not be possible to re-assess all submissions as a consistency check, and getting markers to do this could be challenging. An idea for more significant consequences could be to have some kinship. If one changes the score on one submission, one could get suggestions on similar submissions and ask if one would like to re-assess these.

5. Evaluation

Performing Re-Assessments

Several of the respondents stated that they often assess a proportion of the tasks twice and that it is common to do a warm-up round. The most common method was to take a new look at the first batch of assessments when having finished the assessments. One interviewee mentioned that it takes around twenty assessments to set a foundation, like rules for points deduction. Subsequent assessments will have the same baseline, but this does not necessarily apply to the first ones. Hence, one sometimes has to go back to check for compliance and possibly adjust the first assessments. Another interviewee also supported that after a number of assessments, one can discover *drifting* in the expectations of the answers. One may have had too high or too low expectations for the submissions, and through the assessments, the expectations are adjusted. Instead of re-assessing, this interviewee kept the same baseline for that task and accepted some skew before adjusting all submissions. It was considered easier to scale it rather than assess some submissions twice.

Other interview subjects expressed that conducting re-assessments was not something they did. There is already enough trouble with finishing the assessments within the deadline. Afterwards, there are requests for justifications and complaints. They already have a massive workload within a period, so performing re-assessments can not be prioritised.

Some interviewees shared that they did not actively perform re-assessments but that it sometimes occurs. Many courses have several markers that perform assessments. Usually, one can see the other markers' grading on the submissions. Accordingly, one can compare one's evaluation with other markers' evaluations. If one detects a large variety in scoring, for instance, if one marker has a different average of scores than another, it could be relevant to re-assess.

Re-Assessments by Different Markers

Another interview subject stated that they did variants of re-assessments. They did not assess the identical submissions more than once, but the submissions themselves could be evaluated multiple times by different markers. This applies to complaints cases. A new marker will assess a formerly marked submission. The interview subject also explained that it was common to practice double marking when there were new markers in a course. It was carried out by having the new markers perform assessments first, and then the experienced markers assessed the same tasks. After this, they compared their assessments before drawing some conclusions and adjusting their marks with each other.

Suitable Situations to Re-Assess Answers

Based on the statements from the interviews, one can see that it is common to conduct a form of re-assessment. The motivation behind re-assessments is mainly to ensure fairness and consistency. The markers want the assessments to be correct, but a "correct assessment" can vary between markers. One can conclude that markers only perform re-assessments on answers they are somewhat uncertain about. None of the interview subjects mentioned that they assessed all submissions twice – there had to be a case for them to do the re-assessments. Also, one knows that there is a lack of motivation and high workload regarding assessments. These factors might be reasons to not perform re-assessments or only re-assess a smaller portion of the assessments.

5.2.5. Grouping

Most of the respondents had an opinion regarding how to group the answers on the Assessment page. Because the sorting algorithms decide which answers are displayed together, it is the main influence regarding the Grouping and can impact the assessment. The implemented sorting algorithms are random, based on candidate id, and based on length. All of the respondents were quick to explain that answer length does not implicate the quality of content, but the opinions were split on whether they preferred grouping answers based on their quality. Also, some respondents had no immediate or significant interest in how the grouping was performed.

One respondent meant that grouping answers by candidate id would not add value to the concept. Having nearly random assessments would not give the best conditions for utilising Side-by-side comparison as the answers probably are not comparable. Then, the concept would, in reality, be similar to assessing answers one at a time.

Answers with Similar Content

The majority of the respondents said they preferred a grouping that is based on *similar content* in the answers. That way, one can easily get an overview of the answers and faster set a baseline of expectations. It is a form of norm-related referencing, which is explained in [subsection 2.1.4](#). One respondent claimed that presenting similar answers was an absolute necessity in order to use Side-by-side comparison. He questioned how one could compare answers with no immediate relation. Another respondent said "*The more similarity, the better. Maybe.*". Others also said they wanted at least some similarity between the answers. However, another respondent said that displaying similar answers together would lead to bias.

5. Evaluation

Although most respondents wanted grouping based on similar answers, they proposed different approaches to achieve it. One suggestion was to look at *semantic similarity*. Semantic similarity is a metric that measures how similar two pieces of texts are in terms of their meaning. Multiple resources can be used for this method; however, not many resources have expanded to the Norwegian language, which can pose a challenge. Another respondent wanted to look at sentence structures, as he admitted being biased in terms of negative influence when candidates display flawed writing abilities.

Answers Similar to the Marker Scheme

While some of the respondents meant that the answers should be compared to each other, most of them suggested they should be compared with the marker scheme. Conclusively, as the marker scheme is a correct answer to the task, answers would be sorted based on their quality. This is a type of mastery learning, which is a common type of grading described earlier in [subsection 2.1.4](#). One idea, which two respondents mentioned, was to use keywords for comparison. Then, when creating the marker scheme, a set of keywords is also made. Based on those keywords, the answers would be grouped depending on how many of the keywords the answer contains. This grouping demands some extra pre-work, but implementing the solution is quite trivial, and it can give a quick indication of the answer quality. However, some candidates may use all the keywords incorrectly, thereby showing inadequate competence, but be grouped with stronger answers. On the other hand, other candidates may convey the same meaning as the keywords, but as they do not use the exact words, they are categorised as weaker answers. After considering this, one can see that the suggestion can improve the grouping, but it cannot guarantee that answers of similar quality are displayed together. However, if the markers are aware of the grouping method, the knowledge can help them understand the reason for why certain answers are grouped together and thereby perhaps reduce confusion or bias.

Additionally, one of the respondents wanted to combine the grouping method based on keywords by comparing the absolute length of the marker scheme with the answer length. Answers containing many keywords and having similar lengths to the marker scheme should come first. Then, those answers that are not as similar can gradually follow. The interview subject meant that one can often see certain thresholds on what is included in an "A-answer", "B-answer", "C-answer", and so on. With this kind of sorting algorithm, one can easier see to which grade an answer belongs. Also, by evaluating answers that most likely are the strongest, one can re-adjust the expectations from an "A-student" earlier in the assessment process. Although the length of an answer is not decisive for knowing if the content is relevant, it can give added value when combined with the keywords. As multiple interview subjects have mentioned, they like concise answers, so using the answer length as a factor can improve the sorting algorithm.

Compatible Subjects

In general, using automatic digital tools to determine the quality of answers gets more challenging as the tasks demand increasingly complex answers. Therefore, two respondents pointed out that it works best in technical subjects on a low level. As mentioned, those subjects often test if the students know certain curriculum elements rather than discussing in-depth problems. Then, grouping answers of similar quality can be more straightforward to implement.

5.2.6. Setting Scores on Answers

All of the respondents commented on the Grade component, and they did prefer giving numerical points rather than a letter grade. From previous experience, they were not familiar with setting grades on tasks. They were unsure about the specific percentage limits in the grade scale, which they disapproved of in a high-stake assessment situation. As there are percentage numbers behind each grade, some respondents pointed out that it is easier to relate directly to the numbers than to grades. Also, they disliked that the grade scale is not linear, that a grade is a range, and that the nuances within the grades were lost. However, one respondent argued that it could possibly be user-friendly in other, non-technical fields, where there are few, large tasks, although, for his own sake, he preferred giving points.

None of the interview subjects had a clear solution regarding the granularity, but they had some preferences. Some meant that 6, the same as the grade scale, was a sufficient level. Others wanted to set scores from 0 to 10. However, most mentioned that they preferred having a familiar and stable number as granularity, so it is easy to calculate straightforward fractions such as 50% or two thirds. Therefore, having a dynamic granularity based on the tasks' max points was not requested. However, a repeated point was that the granularity depended on the task and how much the task was weighted. They, therefore, suggested that it should be up to the user.

Using grades to set scores has recently become the standard for Inspera Assessment. However, all respondents avoided the feature, whereas some did not know it existed. The goal with the component was to have it neutral so the respondents could focus on other aspects of the web application. It had the opposite effect. The Grade component, although recommended by the supervisor of this project, worked against its purpose and took up unnecessary attention.

5. Evaluation

5.2.7. Bias in Assessments

From [subsection 2.1.5](#), one knows that a concern within the evaluation process is bias. It applies to both positive and negative biases in favour of the students. Many respondents shared that both external and internal factors can lead to bias.

Causes of Bias

Several respondents said they were aware that bias in the evaluation process occurs. One stated that he could be affected by what he described as good answers. *"When you read through, and then you see that there are very good answers, then maybe it makes you believe that everything is good."* Thus, one is in a positive mindset meaning that positive bias can arise, and the candidates may receive a more excellent score than otherwise. Another respondent expressed that he was affected by sentence structures, language mistakes or common mistakes in the tasks that the candidates make. *"If many have made the same mistake, then I get more and more annoyed, so I may have to go back and adjust."* This results in the opposite of the former case, where one holds a more negative attitude. The respondent admitted being biased in terms of negative influence, which may result in candidates getting poorer assessments. However, he said that by being self-aware about his bias and pointing it out, he reckoned that it could reduce his bias.

Multiple interviewees said that time and place have an impact on the assessments. They believed that assessing before and after a meal would be different, which supports the theory described in [subsection 2.1.5](#) that humans are sensitive to human needs. One interviewee said that *"When you consider something, it's a bit random what you end up with. If I had marked the same assignment ten minutes later, I might have ended up with a different score."* This indicates that coincidences or noise in the measurements may affect the assessments.

The evaluation process itself can lead to bias. A few respondents mentioned that it is not their favourite job, but it is necessary and needs to be done. Assessing many similar answers is tedious work, and as described in [subsection 2.1.5](#), it may cause fatigue and bias. Consequently, it could result in improper work. An initiative to prevent this type of bias could be to limit how many assessments one should assess. It requires hiring of more markers to spread the work. However, several markers expressed that it is challenging to recruit markers as low wages and repetitive work are not attractive. One respondent said *"I think many people agree that if you do repetitive work, you will get bored in the end."* Boredom might result in bias as discovered in [subsection 2.1.5](#). Hence, distributing the workload might not be sufficient alone to reduce bias.

5. Evaluation

One interview subject said that expectations influenced his evaluations. He said that "*Mentally, one adapts to the answer one expects.*", and called this *drifting*. It means that the first candidate does not necessarily receive the same foundations for grading as candidate 30. The expectations can be adjusted when learning the candidates' knowledge level throughout the evaluation process. Consequently, the last batch of assessments can have a different grading foundation due to the evolved expectations. A solution to drifting could be to assess in different orders for the tasks. That means evaluating candidates 1 to 30 for the first task and then candidates 30 to 1 for the next task. It is uncertain if this has any effects. However, it is reasonable to assume that drifting exists and that measures taken to reduce bias are valuable.

Preventive Measures

Measures to prevent bias are similar to efforts to ensure consistency due to the kinship of bias and consistency. One action is to have a new look at the first batch of assessments after all assessments are finished. Another action is to create a precise marker scheme. Several interviewees shared that they made a coding system for scoring points with common mistakes and their associated points deduction. One also shared that it was common to categorise answers. Another action concerns how one designs the tasks for the exam. One can try to construct tasks that can be assessed objectively. There also exist external measures that, for instance, the institution decides. A respondent shared that a quarantine period is required before becoming an examiner after graduating from university. It is to ensure that one does not assess someone one knows. However, many exam submissions are anonymous, but Master's theses are, for instance, not⁴.

Regarding bias in the web application, a respondent stated that the Side-by-side comparison provides an opportunity to compare submissions in different batches in different tasks. It was argued helpful because one will not always start assessing the same candidate for each task and learn the order. This will contribute to preventing a marker from being biased towards any candidate. Additionally, the respondent expressed that it is good to compare several submissions as this itself is a measure to uncover if any bias has occurred.

⁴<https://i.ntnu.no/wiki/-/wiki/Norsk/Karaktergivende+vurderinger+i+eksamens-+vs+e-1%C3%A6ringssystem>

5. Evaluation

5.2.8. Efficiency in the Assessment Process

Since the concept seeks to support efficiency, it was essential to get the respondents' experiences and thoughts on efficiency concerning the evaluation process. As described earlier, many respondents expressed that there are other factors than the large pile of assessments that is time-consuming. One of the challenges was all the arrangement around process concerning getting enough markers, doing sufficient preparatory work, and the framework's usability.

Measures to Support Efficiency

Several interviewees mentioned that they had created their own systems for grading as they did not feel that Inspira Assessment accommodates their needs. Many stated that navigation issues made it confusing and, consequently, less efficient. As mentioned, one respondent made his own comparison program for code correction to make the evaluation process more efficient. Additionally, it is common to have several markers for a course, which requires time spent on marker recruitment and coordination work. Several interview subjects expressed that communication and coordination possibilities were a motivational factor for creating their own grading systems where they could distribute responsibilities and easily get an overview of the progress of the assessments. Most of the respondents said that communication between markers can be a huge challenge and that it is *the* bottleneck in the assessment process. One of them said he prefers to assess all of the submissions himself because communicating the marker scheme and making sure the markers weights answers similarly is too time-consuming. Hence communication and coordination are extra work for markers in addition to assessing submissions.

A measure to reduce time when assessing is to set decisions and requirements for the submissions. It could be to limit the length of the answer or type of question asked, such as drag and drop, fill-in or multiple-choice questions. Having these requirements can help to make the evaluation process more efficient as one knows what to expect when assessing. However, creating and designing tasks for an exam with such formal claims requires much time. Hence, one must balance the preparatory work and the assessing time. As several respondents explained, multiple-choice takes more time to create, but it can be assessed automatically and is, therefore, a way to quicken the assessment process. It is especially valuable considering the limited assessment time. However, multiple-choice questions may not cover everything a student needs to show they know. Reflection and thought process is often lost. One respondent shared a piece of advice for creating systems to support evaluation. He stated that *"All measures that can quicken the assessment process at the same time as the assessment is useful with measuring learning objectives. That's a point."* Therefore, markers must find a balance between assessing time and the measurements of the students' knowledge.

5. Evaluation

Double Marking

Several respondents talked about the double assessment requirement mentioned in subsection 2.1.6. The respondents who brought up the resolution explained that it would double the workload. There is a struggle in recruiting markers, and due to the marker shortage in their subjects, they would have to change their assessment types. This is because they would have to create assessments that can be assessed more efficiently due to the increased workload.

Effectiveness of the Concepts

Feedback to the Side-by-side comparison was that it had potential in regards to efficiency. Several stated that it depended on which submissions were displayed together where similar answers would be beneficial. Several respondents pointed out that having the Task-by-task assessment method as the foundation for the main concept, there is less context switching, which can increase efficiency. Other feedback was that efficiency might disappear due to the scrolling of submissions within the Answer text box. It depends on the length of the answers, which again is related to the subject area and level.

In general, many respondents seemed favourable to how the concepts could support efficiency when assessing. However, more development is needed for the concepts to be optimal. The current assessment process should be more efficient, and the respondents appreciated the focus this research project sets on the issue. That might be one of the reasons for their openness to the Side-by-side comparison. They saw advantages in exploring alternative assessment methods.

5.2.9. System Flow

All of the respondents had experience with the digital assessment platform Inspira Assessment because it is required by the university. They have used it in different ways, and several respondents said that it can be hard to navigate, but it is okay if one is familiar with it. However, there are indications from the interviews that Inspira Assessment has room for improvement. Two respondents have created their own systems to improve the assessment process, which shows a shortcoming in Inspira Assessments functions in their situations. Another interviewee stated that *"There are many [markers] who are annoyed with Inspira Assessment. I think it is fine when it works. It [the program] is a bit hard to navigate, but when you know where things are, it works. And it is going to be like this with this system [the web application] as well, you need to get used to it."* One can see that Inspira Assessment has fundamental tools in place to enable different evaluation methods. However, Inspira Assessment does not accommodate every markers' need, and a further upgrade is desired.

5. Evaluation

The interviewee also shared that he liked using a structured Excel sheet when assessing subjects that involve other markers. Several other interview subjects also explained their liking for Excel as an efficient way of communicating with other markers when setting scores. As discussed earlier, communication is a bottleneck for markers in the evaluation process. Excel is valued because it gives an overview where markers can easily see what baseline other markers have on each task, as it is likely that all of the markers have candidates at different levels. With Excel, it is easy to adjust scores if one sees a gap between baselines. Therefore, there is a demand for an improved communication system for markers internally. Additionally, Excel gives a logical, understandable, and clean system that allows the users to get an overview customised to their needs. It is a program many are familiar with, so it requires less effort to start using.

User-friendliness of the Web Application

Regarding the flow of the web application, multiple interviewees expressed the importance of an *user-friendly* app. However, they had different views on what would make it user-friendly. One respondent wanted the option to quickly tab through the page and set scores so that the user do not need to lift a finger from the keyboard. Another respondent suggested using a keyboard shortcut to score an answer. Additionally, there was a suggestion from another respondent who wanted an introduction when beginning to use the web application. That way, one could quickly learn how it works and what possibilities the system holds. He also wanted it to be intuitive and without multiple ways to navigate the same pages, which can be confusing. Other respondents wanted the system to be more flexible and have more possibilities to integrate self-made solutions with the assessment tool. Due to the respondents' different preferences, subjects to assess, types of tasks, and skill sets, the system must be flexible while remaining user-friendly. One interviewee informed that it should be possible to use the Side-by-side comparison on *some* tasks while other tasks could be assessed in another way. The balance between having many features and a user-friendly, straightforward web application can be hard to determine. Therefore, it is meaningful to involve users from the target group to test the web application in future development.

5.2.10. Future Ideas

Throughout the interviews, respondents shared many ideas and improvements to the current concepts. Several respondents were engaged in learning technology and had ideas for other digital assessment tools that they gladly shared.

Manual Grouping

Two respondents emphasised grouping similar answers to score similar answers easily. One suggestion was to have an additional button that gives all answers the same score. This feature reduces clicks when setting grades by speeding and giving a more fluent process. However, giving students similar scores can become too simple as some answers displayed together might deserve different scores. It may lead to inaccurate scores, but one should be able to trust markers to prioritise giving students correct scores.

Moving forward on the suggestion, one respondent proposed grouping answers *manually* and giving each group a score. This should be done by "flagging" each answer as a "type of assessment", for example, "Type 1". Then, if another answer with similar correct or wrong parts, the marker adds it to "Type 1". Additionally, the respondent continued suggesting that if the system spots an answer that looks similar to the answers in a group, the user should be presented with a tip that suggests they put the answer there. This method could make it easier to perform targeted adjustments in scores if the marker notices a drift in the scores given. However, to make it efficient, it could be helpful to have some automatic sorting before beginning the grouping. As mentioned earlier, some respondents said they would adjust the scores of all candidates up or down if there is a change in baseline through the assessment process. Giving groups scores could prevent this kind of adjustment as it enables a more targeted correction. However, this is unknown territory, and the method would need further research and investigation.

Ranking System

One interviewee suggested that having a comparison and ranking system could help ensure consistency. An algorithm could ask if a submission is better than a batch of other submissions. Instead of scoring points, one could compare and ask which of them is better. He meant that submissions would then be likely consistent with each other. He suggested that it *"Can be both yes or no, this is better or a type of ranking."* He believed that it is easier to determine if something is better than something else than setting a score and also that it is easier for markers to agree with each other with ranking. At last, one will get all the answers sorted by ranked level. Then, one can set the number scores and determine where the weaker submissions in the scale start, how many points they should have, and where the submissions with one more point should start.

5. Evaluation

Consistency Check Reminder

One more idea involved some re-assessment. When one has finished assessing a set of submissions and then sets a grade, there will be earlier submissions with the same grade presented to the marker. The purpose is to compare the submissions and check if they are on the same level. It can work as a reminder to check if one is consistent and is, therefore, some form of consistency check.

Interest in Exploring the Assessment Process

Considering the feedback and ideas received in the interviews, one can assume that markers are interested in the field. They do see the potential for improvements and support in the evaluation process. The new ideas show that they are receptive to new digital solutions that can help them ensure consistency and quicken their work with assessing exams.

5.2.11. Summary of the Interview Results

The respondents have different experiences, assessment methods, and preferences. Some of them have assessed exams using the Task-by-task assessment method, and one has even tried a form of Side-by-side comparison. Several respondents use or have used Excel in addition to the required assessment system and see it as a beneficial communication system between markers. Also, there is a general understanding that the assessment work was not something they preferred doing but was instead described as something that has to be done. The lack of enjoyment can be a motivation to make the assessment process more efficient.

There were divided opinions regarding the concepts. Most respondents saw them as a possible improvement to make scoring answers easier by earlier establishing similar baselines. Some of them also saw the comparison of candidates as necessary to utilise the Side-by-side comparison. Others were worried about how the comparing could lead to markers wanting to differentiate and rank the answers even though they might deserve similar scores. Then again, other respondents saw the ranking as a possibility for a new assessment method. The respondents usually valued the Consistency check or Grouping when evaluating the web application. Improvements for both of these elements were suggested, often related to the interviewee's situation. Regarding Grouping, the respondents often wanted similar answers displayed together. Considering the Consistency check, some respondents saw benefits with re-assessing answers but re-assessing more than the first 20 answers would be more time-consuming than valuable. Also, as fairness and objectivity were considered valued qualities in assessment, a measure to prevent bias and preserve consistency was appreciated. This might explain the openness to the

5. Evaluation

Grouping and Consistency check. Also, having a user-friendly and flexible system was valued by many respondents. Moreover, when creating an exam, adapting it to the concept was considered important to ensure consistency.

Overall, the respondents shared their thoughts and opinions, but as none tested the system in a realistic situation, they had limited possibilities to express an in-depth evaluation. Also, a desire to see the concept in a more realistic assessment system was addressed. The results from the interviews indicate that the concept has potential for further development. The concepts were not clearly rejected, suggesting that there are interests and susceptibility to the proposed concepts. Though, as emerged in the interviews, there is room for improvement.

5.3. Limitations in the Study

This study has a few limitations which are worth noting. First, a lack of expertise and experience within the evaluation process leads to some presumptions in the design process of the concept. For example, choosing to put more effort into designing and developing the Consistency check than Grouping and its' sorting algorithms proved to be a misjudgment. It was assumed that the Consistency check was valued and needed to prevent bias and inconsistency. However, interviews showed that the respondents were more interested in the sorting algorithms and how they could contribute to ensuring consistency. Hence, an amount of time could have been spent developing the sorting algorithms instead and getting further insights and feedback on this concept. Also, the grading component was redesigned after input from the supervisor and guerilla testing. However, none of the respondents was familiar with scoring tasks with letter grades and preferred numerical points, which was the original design. Thus, time spent redesigning and discussing the component consumed time that could have been spent on the core concepts.

Secondly, all of the respondents were markers in higher education, and all of them, in addition, were recruited from NTNU. Most of them were also men, and many belonged to the Department of Computer Science. An attempt was made to recruit markers from Handelshøyskolen BI in Trondheim, but unfortunately, no one could participate. markers from other departments were contacted, but only a couple could participate. It would have been valuable to get a greater spread in the data collection and insights from markers of different schools and at different levels. As the interview period was held late in the project, and the priority was to first recruit markers in higher education through a network, it was not enough time to start recruiting teachers at high schools and middle schools. Additionally, the number of respondents also means that the results are indications and do not apply to the population.

5. *Evaluation*

Finally, time constraint was a limit of this project. Several planned implementations had to be put on hold and were not implemented for the walkthrough in the interviews. Also, as mentioned, more time could have been spent recruiting a broader group of respondents. The interview period took place for a week, and most of the interviews were planned and conducted within a couple of days. If there was more time between the planning and conducting of the interviews, it could have been possible to get exam data sets from the respondents and use this in the walkthrough. The walkthrough would then have been more personalised and familiar. It would most likely be easier for the respondents to envision how the concepts could be applied in practice.

6. Conclusion and Future Work

The overall goal of this thesis is to explore how one can support assessment processes in regards to efficiency and consistency. Through the research questions, different parts of the goal is addressed and discussed by uncovering challenges markers have when assessing, designing a Side-by-side comparison with supporting concepts and evaluating the concepts. The concepts were brought to life through a web application. It can be found on GitHub¹.

6.1. Findings for the Research Questions

RQ1: What Are the Challenges Encountered When Assessing?

As explained in [chapter 2](#), both external and internal factors can affect the evaluation and induce challenges when assessing. One of the external factors is the given time limit to assess the exam submissions. Also, the exam submissions can be assessed by one or multiple markers, depending on the number of students enrolled in the subject, the assessment type, and the preparations made. A respondent expressed that communicating the criteria and baseline to other markers can be time-consuming because different markers have unlike perceptions on how to weight the points in a task, and that it was easier to assess all of the submissions oneself. With the different weighting, one might need to adjust the scores after the assessments to ensure a fair evaluation. In addition, as explained in [section 1.1](#), the number of enrolled students in higher education is ever-increasing, and therefore, the already substantial workload will grow.

The markers can also have internal factors causing challenges within the assessment process. Bias can occur due to mental fatigue, which can come from context switching, high pressure, lack of sleep, hunger, and monotone work. Interview respondents said that the assessment work is not something they genuinely enjoy doing; other aspects of their professions appeal to them more. The lack of motivation can be challenging.

These challenges give an incentive to explore an alternative concept that can make the assessment process more efficient and consistent.

¹<https://github.com/amalieeh/assessment-support>

6. Conclusion and Future Work

RQ2: How Can Assessment Support Be Designed to Support Efficiency and Consistency?

In [chapter 3](#), concepts were designed to explore support for the assessment process. The concepts narrow towards digital, summative assessment in higher education of written, short answer exams. The main concept is the Side-by-side comparison with the supporting concepts Task-by-task assessment method, Consistency check, and Grouping. From the preliminary studies conducted, a finding was that it exists several AES systems with the aim of making the assessment process more efficient. However, automated systems are not sufficient for all types of exams, and there is still a barrier for trusting them fully to score high-stake assessments. Therefore, the Side-by-side comparison became a viable proposal as there are still humans behind the evaluation while it supports the assessment process.

The Side-by-side comparison has the Task-by-task assessment method as a foundation. By having multiple student submissions from the same task displayed simultaneously, one can be able to assess more efficiently. Firstly, focusing on the same task can reduce context switching and heavy mental processing. Secondly, displaying several tasks side by side can reduce time spent on getting an overview and loading and switching tasks. Finally, by assessing the same task for a more extended period, one will be skilled and proficient with that task, meaning that assessing will be easier and more efficient.

Additional concepts to support the Side-by-side comparison were Consistency check and Grouping. These concepts were considered necessary to strengthen the Side-by-side comparison. A common denominator was also considered beneficial when displaying anything side by side. Having the same task was one denominator, but additional options could be relevant. Hence, Grouping with its' sorting algorithms was designed with text lengths and randomisation as options. Additionally, an assumption was that bias could occur due to the enabling of comparison, so designing a Consistency check for support seemed convenient. Having a Consistency check that chooses a proportion of assessments to be re-assessed based on measures compared to their current batch can unveil any occurring bias or inconsistency in the assessments.

RQ3: What Are the Perceptions of Using Side-by-side Comparison?

As presented in [chapter 5](#), interviews were conducted to gather insights into respondents' experiences with assessing and thoughts on the concepts. A web application was implemented and demonstrated during the interviews to showcase how the Side-by-side comparison could be applied in practice.

The interview subjects had different impressions regarding the concepts. Some found it interesting and could see the Side-by-side comparison as relevant or having potential. Others were unsure whether they would like it and highlighted the pitfall of ranking students amongst themselves and saw this as a disadvantage.

6. Conclusion and Future Work

Feedback retrieved was that the Side-by-side comparison required that one did preparatory work when creating exams. This is because not all types of questions would fit the concept, and one would have to develop suitable questions. Additionally, respondents expressed that it depended on the subject. It had to be subjects with questions that do not require much elaboration because longer answers would not fit for a side by side view. The subjects would also need to have comparable topics for it to be fair to use the concept in assessments. This is because it enables comparison.

Several respondents practised the Task-by-task assessment method as they meant that it was efficient and would lead to consistency in the assessments. One only has to focus on one topic at a time, meaning one would be able to make decisions quicker and assess more efficient. Also, consistency would be maintained as the same marker indirectly assesses all the submissions with this approach. This also results in another gain, fairness.

There were comments concerning Grouping that it was essential to consider which submissions were displayed together. This is because it was believed that it could affect the scoring and lead to bias. The weighting of a submission may vary depending on the batch the submission is a part of. Hence, many stated that the grouping of submissions greatly impacted the weighting and the potential of the Side-by-side comparison. Some respondents mentioned that by grouping similar answers together, it would be easier to weigh the similar submissions equally. This would then result in consistency in the assessments. Others meant that this would make it easier to differentiate submissions that should be weighted equally and get the same score. This means that one could be biased and rank submissions when one should not.

Regarding the Consistency check, some stated that the idea of re-assessing a proportion of submissions was smart. Still, many respondents shared that they usually do not have time to re-assess submissions twice. They already struggle to finish within the deadline. Hence, this Consistency check would lead to an increased workload. Though, many shared that the sometimes assess the first submissions twice to ensure that they correspond to the level of the submissions. Many set the level or baseline throughout the evaluation process, meaning that the first set of submissions may have been assessed with a different baseline. Therefore, to ensure consistency and fairness, many checked these submissions twice. Thus, the designed Consistency check is not as relevant and valuable due to the time constraint. However, some saw potential in the Consistency check and suggested that it could be helpful once the grade limit was set.

Concerning efficiency, there is no dominant conclusion. It was difficult for most respondents to get an impression of whether assessing with a Side-by-side comparison contributes to efficiency in the process. This is not surprising as they were only presented and demonstrated with a test case. However, many were interested in the research and the Side-by-side comparison, but to evaluate efficiency, it has to be tested in a real setting.

6.2. **Future Work**

Throughout this master's thesis, several improvements and ideas have arisen in conjunction with assessment support. First and foremost, the grouping options for the Side-by-side comparison need some modification and adjustments. The suggestion retrieved from the interviews was that grouping could be based on similarity. Therefore, further work could be to implement grouping based on the similarity between the submissions, meaning that similar submissions are grouped and displayed together in the same batch. This similarity measure could be on text analysis of the student submissions, where submissions with similar terms are grouped together. Another measure could be a similarity to the marker scheme or a set of terms markers decide.

The Consistency check needs further work as many stated that they barely have time to assess all the submissions. With the current solution, the Consistency check would increase their workload. It would therefore have a counteracting effect on efficiency. However, the purpose of the Consistency check is beneficial, and one could consequently evolve it. Many respondents stated that they perform a consistency check on submissions close to the grade point limit. Since they did this manually, an implementation of a digital solution could aid and support the efficiency of the process.

One of the aims of the Side-by-side comparison was to explore it could affect efficiency in an assessment process. Therefore, future work would be to perform an experiment where this can be tested and measured. In order to conduct this experiment, some improvements have to be carried out first. Markers must be able to complete an entire assessment process meaning that the web application must be further developed or it has to be integrated with an existing system such as Inspera Assessment. For this to be possible, markers must actually assess a set of real submissions as they would usually do. This means that one must get hold of data such as the submissions, task descriptions and the marker scheme. Whether one chooses to continue the work on the web application or to integrate the Side-by-side comparison with a system, end-users with different backgrounds should be included as part of the development process to gather feedback and insights on the implementation. Once all of this is in place, one will be able to perform an experiment and test if the Side-by-side comparison impacts efficiency in the assessment process.

Bibliography

- Anglin, L., Anglin, K., Schumann, P. L., and Kaliski, J. A. (2008). Improving the efficiency and effectiveness of grading through the use of computer-assisted grading rubrics. *Decision Sciences Journal of Innovative Education*, 6(1):51–73.
- Archer, J. and McCarthy, B. (1988). Personal biases in student assessment. *Educational Research*, 30(2):142–145.
- Bernard, M. E. (1979). Does sex role behavior influence the way teachers evaluate students? *Journal of Educational Psychology*, 71(4):553.
- Birkelund, J. (2015). The lunch effect. Can it result in biased grading at universities? Master's thesis, UiT The Arctic University of Norway.
- Coker, D. R., Kolstad, R. K., and Sosa, A. H. (1988). Improving essay tests: Structuring the items and scoring responses. *The Clearing House*, 61(6):253–255.
- Czaplewski, A. J. (2009). Computer-assisted grading rubrics: Automating the process of providing comments and student feedback. *Marketing Education Review*, 19(1):29–36.
- Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.
- Dux, P. E., Ivanoff, J., Asplund, C. L., and Marois, R. (2006). Isolation of a central bottleneck of information processing with time-resolved fmri. *Neuron*, 52(6):1109–1120.
- Fisherl, C. D. (1993). Boredom at work: A neglected concept. *Human Relations*, 46(3):395–417.
- Frey, B. B. (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*.
- Katidioti, I., Borst, J. P., and Taatgen, N. A. (2014). What happens when we switch tasks: Pupil dilation in multitasking. *Journal of Experimental Psychology: Applied*, 20(4):380–396.
- Kenrick, D. T. and Gutierrez, S. E. (1980). Contrast effects and judgments of physical attractiveness: When beauty becomes a social problem. *Journal of Personality and Social Psychology*, 38(1):131–140.

Bibliography

- Klein, J. (2002). The failure of a decision support system: inconsistency in test grading by teachers. *Teaching and Teacher Education*, 18(8):1023–1033.
- Kryder, L. G. (2003). Grading for speed, consistency, and accuracy. *Business Communication Quarterly*, 66(1):90–93.
- MacWilliam, T. and Malan, D. J. (2013). Streamlining grading toward better feedback. In *Proceedings of the 18th ACM conference on innovation and technology in computer science education*, pages 147–152.
- Malouff, J. (2008). Bias in grading. *College Teaching*, 56(3):191–192.
- Malouff, J. M., Stein, S. J., Bothma, L. N., Coulter, K., and Emmerton, A. J. (2014). Preventing halo bias in grading the work of university students. *Cogent Psychology*, 1(1).
- McAlpine, M. (2002). *Principles of Assessment*. Citeseer.
- Nisbett, R. E., . W. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4):250–256.
- Nisbett, R. and Ross, L. (1983). Human inference: Strategies and shortcomings of social judgment. *The Philosophical Review*, 92(3):462–465.
- Norman, D. A. (2002). *The Design of Everyday Things*. Basic Books, Inc., USA.
- Oates, B. J. (2006). *Researching Information systems and Computing*. SAGE Publications, London, Thousand Oakes, New Dehli, 1 edition.
- Page, E. B. (1966a). The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(6):238–243.
- Page, E. B. (1966b). The use of the computer in analyzing student essays. *International Review of Education*, 14(3):253–263.
- Page, E. B. and Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *The Phi Delta Kappan*, 76(7):561.
- Popham, W. J. (1997). What’s wrong—and what’s right—with rubrics. *Educational Leadership*, 55(2):72–75.
- Quinn, R. P. (1975). What makes jobs monotonous and boring? *Paper presented at the annual meeting of the American Psychological Association, Chicago, IL*.
- Sambell, K., McDowell, L., and Brown, S. (1997). "But is it fair?" : An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, 23(4):349–371.
- Shermis, M. D. and Burstein, J. C. (2003). *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Routledge.

Bibliography

- Valenti, S., Neri, F., and Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1):319–330.

A. Interview guide

Interview Guide

Introduction

In association with our master thesis, we have developed a minimal viable product web application that facilitates a digital Side-by-side comparison for assessing exams. The method is based on exams with multiple tasks and displays several student answers for the same task next to each other. The project is conducted to uncover unknown factors which affect the assessment process. It will examine if an alternative evaluation method can contribute to an efficient and consistent process of assessing exams. The web application is a suggestion we have created to illustrate how the method can be applied in practice.

All information that is shared in this interview will be anonymised.

The interview consists of three main parts. First, there will be some questions to get a baseline. Then there will be a walkthrough of the web application, and lastly, there will be more questions.

Before walkthrough

1. Have you created or assessed exams? If so, what type of answers did the exam have?
2. What or which tools do you use for assessing exams? // Hold det relatert til tema
 - a. Do you think that they have any potential for improvements? If so, can you give us a description of which?
3. What are your thoughts on re-assessing exams?
 - a. What do you think about reassessing exams versus using a marker for consistency check?
 - b. How often do you re-assess an exam?
 - c. How large of a proportion of the exams would you be willing to re-assess?
4. Do you have any experience in assessing side by side? If so, will you share your thoughts on this method?
5. Do you have any experience with assessing task by task? If so, will you share your thoughts on this method?
6. When assessing exams, how do you prevent bias in the assessments?

Walkthrough

Follow the given instructions.

After walkthrough

1. What is your impression of the concepts? Did it align with your expectations? Please explain both what did and did not align.
2. How would the Side-by-side comparison impact your assessment? Bias?

3. In what way would the Task-by-task assessment method affect your assessment?
Bias?
4. What was your impression of the Consistency check?
 - a. What do you think would be a good way to check for inconsistency and bias?
5. Regarding both of the previously mentioned assessment methods and the Consistency check, do you think bias will be of any concern? If so, in what way?
6. How was the sense of control over the general progress? Was the flow intuitive?
Please explain.
7. How do you think this method impacts efficiency and consistency compared to your usual methods?
8. In what settings would the concept be useful? Do you have any concerns?
9. Do you have any other comments?

B. Walkthrough guide

Testoppgaver

Backstory:

Du skal rette et eksamenssett. Du har akkurat lastet opp datasettet med eksamensbesvarelsene på denne siden.

Dette er eksamensoppgaver fra IT2810 Webutvikling. Vi forventer ikke at du skal faktisk rette det, det er ment som et eksempel.

Husk på:

- Tenke høyt
- Si hva du ser
- Begrunn alle valg du tar
- Fortell hva du er usikker på

Oppgaver

1. Du skal rette og bekrefte retting av oppgave 1: Responsiv web-design i eksamenssettet. Hvordan ville du løst dette? Si ifra når du tror du er ferdig.
 - a. *Hvor effektivt synes du det var?*
 - b. *I hvilken grad tror du det var en riktig og rettferdig vurdering*
 - c. *Ser du poenget med å bekrefte retting?,*
 - d. *Poeng-knappen*
2. Du ønsker å rette Oppgave 2: Interaktiv grafikk. Hvordan går du frem?
 - a. Du er usikker på løsningsforslaget og sjekker det.
 - b. Sorter oppgavene på en annen måte
 - i. *Hvilken velger du? Hvorfor? Ser du noen nytte i denne funksjonen?*
 - c. Flagg en oppgave
 - i. *Ser du noen nytte i denne funksjonen? Hva ville du brukte den til?*
3. Rett oppgave 3: REST vs. GraphQL
 - a. Du ønsker å justere antall besvarelses som vises samtidig halvveis i rettingen. Hvordan løser du dette?
 - i. *Hvordan tenker du det påvirker rettingen*
 - b. Kan du justere konsistenssjekken, og forklare hva det betyr?

Nå har vi for test purposes gjort at den anses som ferdig med å rette alle oppgaver når man har rettet tre. Så anta du er ferdig å rette alle oppgavene. Hva ville du gjort nå?

4. Godkjenn og eksportér karakterene, og fullfør vurderingsprosessen.

C. Application for NSD

NSD NORSK SENTER FOR FORSKNINGSDATA

Vurdering

Referansenummer

260634

Prosjekttittel

Masteroppgave om effektivisering av manuell digital eksamensretting

Behandlingsansvarlig institusjon

Norges teknisk-naturvitenskapelige universitet / Fakultet for informasjonsteknologi og elektroteknikk (IE) / Institutt for datateknologi og informatikk

Prosjektansvarlig (vitenskapelig ansatt/veileder eller stipendiat)

Trond Aalberg, trond.aalberg@ntnu.no, tlf: 73597952

Type prosjekt

Studentprosjekt, masterstudium

Kontaktinformasjon, student

Sylvi Phuong Huynh, sylviph@stud.ntnu.no, tlf: 96514486

Prosjektperiode

25.04.2022 - 31.07.2022

Vurdering (1)

21.04.2022 - Vurdert

Det er vår vurdering at behandlingen av personopplysninger i prosjektet vil være i samsvar med personvernlovgivningen så fremt den gjennomføres i tråd med det som er dokumentert i meldeskjemaet med vedlegg, og eventuelt i meldingsdialogen mellom innmelder og Personverntjenester. Behandlingen kan starte.

TYPE OPPLYSNINGER OG VARIGHET

Prosjektet vil behandle alminnelige kategorier av personopplysninger frem til den datoen som er oppgitt i meldeskjemaet.

LOVLIG GRUNNLAG

Prosjektet vil innhente samtykke fra de registrerte til behandlingen av personopplysninger. Vår vurdering er at prosjektet legger opp til et samtykke i samsvar med kravene i art. 4 og 7, ved at det er en frivillig, spesifikk, informert og utvetydig bekreftelse som kan dokumenteres, og som den registrerte kan trekke tilbake.

Lovlig grunnlag for behandlingen vil dermed være den registrertes samtykke, jf. personvernforordningen art. 6 nr. 1 bokstav a.

PERSONVERNPRINSIPPER

Personverntjenester vurderer at den planlagte behandlingen av personopplysninger vil følge prinsippene i personvernforordningen om:

- lovlighet, rettferdighet og åpenhet (art. 5.1 a), ved at de registrerte får tilfredsstillende informasjon om og samtykker til behandlingen
- formålsbegrensning (art. 5.1 b), ved at personopplysninger samles inn for spesifikke, uttrykkelig angitte og berettigede formål, og ikke behandles til nye, uforenlige formål
- dataminimering (art. 5.1 c), ved at det kun behandles opplysninger som er adekvate, relevante og nødvendige for formålet med prosjektet
- lagringsbegrensning (art. 5.1 e), ved at personopplysningene ikke lagres lengre enn nødvendig for å oppfylle formålet

DE REGISTRERTES RETTIGHETER

Så lenge de registrerte kan identifiseres i datamaterialet vil de ha følgende rettigheter: innsyn (art. 15), retting (art. 16), sletting (art. 17), begrensning (art. 18), og dataportabilitet (art. 20).

Personverntjenester vurderer at informasjonen om behandlingen som de registrerte vil motta oppfyller lovens krav til form og innhold, jf. art. 12.1 og art. 13.

Vi minner om at hvis en registrert tar kontakt om sine rettigheter, har behandlingsansvarlig institusjon plikt til å svare innen en måned.

FØLG DIN INSTITUSJONS RETNINGSLINJER

Personverntjenester legger til grunn at behandlingen oppfyller kravene i personvernforordningen om riktighet (art. 5.1 d), integritet og konfidensialitet (art. 5.1. f) og sikkerhet (art. 32).

OneDrive er databehandler i prosjektet. Personverntjenester legger til grunn at behandlingen oppfyller kravene til bruk av databehandler, jf. art 28 og 29.

For å forsikre dere om at kravene oppfylles, må dere følge interne retningslinjer og/eller rådføre dere med behandlingsansvarlig institusjon.

MELD VESENTLIGE ENDRINGER

Dersom det skjer vesentlige endringer i behandlingen av personopplysninger, kan det være nødvendig å melde dette til oss ved å oppdatere meldeskjemaet. Før du melder inn en endring, oppfordrer vi deg til å lese om hvilke type endringer det er nødvendig å melde: <https://www.nsd.no/personverntjenester/fylle-ut-meldeskjema-for-personopplysninger/melde-endringer-i-meldeskjema>

Du må vente på svar fra oss før endringen gjennomføres.

OPPFØLGING AV PROSJEKTET

Personverntjenester vil følge opp ved planlagt avslutning for å avklare om behandlingen av personopplysningene er avsluttet.

Lykke til med prosjektet!

D. Consent form

Vil du delta i intervju til masteroppgave omhandlende ”Effektivisering av digital vurderingsprosess”?

Dette er et spørsmål til deg om å delta i et intervju til en masteroppgave hvor formålet er å gjøre digital vurderingsprosess av eksamensoppgaver ved høyere utdanning mer effektiv og konsistent. I dette skrivet gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg.

Formål

Formålet med prosjektet er å avdekke faktorer angående digital vurdering av eksamensoppgaver ved høyskoler og universiteter og få tilbakemeldinger på et konsept som skal gjøre denne prosessen mer effektiv og konsistent enn tradisjonelle metoder. Konseptet er realisert i form av en web-applikasjon, og denne vil bli gjennomgått i intervjuene.

Vi ønsker å se på hvilken innvirkning det har å vise flere besvarelser samtidig når man vurderer. I tillegg ønsker vi å se på andre faktorer som påvirker digital vurdering.

Det er til en masteroppgave.

Hvem er ansvarlig for forskningsprosjektet?

Norges teknisk-naturvitenskapelige universitet / Fakultet for informasjonsteknologi og elektroteknikk (IE) / Institutt for datateknologi og informatikk ved Trond Aalberg gjennom Amalie Eline Henni og Sylvi Phuong Huynh er ansvarlige for prosjektet.

Hvorfor får du spørsmål om å delta?

Utvalget er trukket blant professorer ved høyere utdanning med erfaring i digital retting av eksamensoppgaver som kan bidra med kunnskap og tilbakemeldinger på forskningsprosjektet. Totalt vil om lag 20 få henvendelse om å delta.

Rekrutteringen av utvalget foregår gjennom eget nettverk. Utvalget blir kontaktet via e-post funnet på den offentlige ansattprofilen hos utdanningsinstitusjonene.

Hva innebærer det for deg å delta?

Hvis du velger å delta i prosjektet, innebærer det at du deltar på et intervju som vil vare i opptil 1 time. Det vil bestå av to spørsmålsrunder og en gjennomgang av web-applikasjonen som er utviklet i forbindelse med prosjektet. Spørsmålsrundene tar for seg erfaringer rundt eksamensretting og tilbakemeldinger på prosjektet.

Det vil bli utført notatskriving og lydopptak av intervjuet. All informasjon fra intervjuet vil bli anonymisert.

Det er frivillig å delta

Det er frivillig å delta i prosjektet. Hvis du velger å delta, kan du når som helst trekke samtykket tilbake uten å oppgi noen grunn. Alle dine personopplysninger vil da bli slettet. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg.

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger

Vi vil bare bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket.

- Trond Aalberg, Amalie Eline Henni og Sylvi Phuong Huynh vil ha tilgang til dataene.

- Personidentifiserende opplysninger og data er passordbeskyttet.
- Personidentifiserende opplysninger vil ikke ligge sammen med data.

Deltakerne vil ikke kunne gjenkjennes i publikasjon da dataen kommer til å anonymiseres.

Hva skjer med opplysningene dine når vi avslutter forskningsprosjektet?

Opplysningene anonymiseres når prosjektet avsluttes/oppgaven er godkjent, noe som etter planen er slutten av juli 2022. Ved prosjektslutt slettes alle personidentifiserende opplysninger og all data er anonymisert. Formålet ved å beholde anonymisert data handler om etterprøvarhet. De lagres på en passordbeskyttet skyløsning der kun Trond Aalberg, Amalie Eline Henni og Sylvi Phuong Huynh vil ha tilgang.

Dine rettigheter

Så lenge du kan identifiseres i datamaterialet, har du rett til:

- innsyn i hvilke personopplysninger som er registrert om deg, og å få utlevert en kopi av opplysningene,
- å få rettet personopplysninger om deg,
- å få slettet personopplysninger om deg, og
- å sende klage til Datatilsynet om behandlingen av dine personopplysninger.

Hva gir oss rett til å behandle personopplysninger om deg?

Vi behandler opplysninger om deg basert på ditt samtykke.

På oppdrag fra Norges teknisk-naturvitenskapelige universitet, Fakultet for informasjonsteknologi og elektroteknikk (IE), Institutt for datateknologi og informatikk har NSD – Norsk senter for forskningsdata AS vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

Hvor kan jeg finne ut mer?

Hvis du har spørsmål til studien, eller ønsker å benytte deg av dine rettigheter, ta kontakt med:

- Norges teknisk-naturvitenskapelige universitet / Fakultet for informasjonsteknologi og elektroteknikk (IE) / Institutt for datateknologi og informatikk ved
 - Trond Aalberg (veileder), trond.aalberg@ntnu.no, tlf: 73597952
 - Amalie Eline Henni (student), amalieeh@stud.ntnu.no, tlf: 45859125
 - Sylvi Phuong Huynh (student), sylviph@stud.ntnu.no, tlf: 96514486
- Vårt personvernombud: Thomas Helgesen, thomas.helgesen@ntnu.no, tlf: 93079038

Hvis du har spørsmål knyttet til NSD sin vurdering av prosjektet, kan du ta kontakt med:

- NSD – Norsk senter for forskningsdata AS på e-post (personverntjenester@nsd.no) eller på telefon: 55 58 21 17.

Med vennlig hilsen

Trond Aalberg

(Veileder)

Amalie Eline Henni

(Student)

Sylvi Phuong Huynh

(Student)

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om prosjektet “*Effektivisering av digital vurderingsprosess*”, og har fått anledning til å stille spørsmål. Jeg samtykker til:

- å delta i intervju
- at mine anonymiserte data publiseres i en masteroppgave
- at mine anonymiserte opplysninger lagres etter prosjektslutt, til etterprøvbarehet

Jeg samtykker til at mine opplysninger behandles frem til prosjektet er avsluttet

(Signert av prosjektdeltaker, dato)

