Norwegian Directorate for Education and Training

THE EXAM IN PRIMARY AND LOWER SECONDARY EDUCATION

The foundation of knowledge for evaluating the exam system in Norway

Report from the exam review group
February 2019

Table of contents

1 Introduction

The exam review group has been appointed by the Norwegian Ministry of Education and Research (cf. a letter of assignment from the ministry to the Norwegian Directorate for Education and Training (hereafter "Udir") dated 26 June 2018). The group shall assist Udir in its work to investigate an overall exam system for the subjects affected by the curriculum renewal. This document, entitled "The foundation of knowledge on exams: preliminary status and assessment", represents partial delivery 1 from the exam review group. The delivery will be further developed and expanded upon through two additional deliveries in 2019, and will be included in the exam review group's final report to the Norwegian Ministry of Education and Research, which shall be submitted in 2020.

1.1 Members of the exam review group and mandate

The exam review group
Sigrid Blömeke (group leader), the University of Oslo
Sissel Skillinghaug, the Norwegian Directorate for Education and Training
Marte Blikstad-Balas, the University of Oslo
Per-Odd Eggen, NTNU – the Norwegian University of Science and Technology
Henning Fjørtoft, NTNU – the Norwegian University of Science and Technology
Siv Therese Måseidvåg Gamlem, Volda University College
Tine Prøitz, the University of South-Eastern Norway
Sverre Tveit, the University of Agder
Rita Helgesen, the Norwegian Association of Graduate Teachers
Stig Johannessen, the Norwegian Association of School Leaders
Martin Minken, the Union of Education Norway
Agathe Waage, the School Student Union of Norway
Mette Johnsen Walker, the Norwegian Union of School Employees

The secretariat: Cathrine Hjulstad, Hilde Hultin, Trude Saltvedt, Øyvind Pedersen and Per Kristian Larsen-Evjen (Head of Department), the Norwegian Directorate for Education and Training

A reference group made up of professionals from the education sector was also appointed, for the purpose of providing input on the work with the deliveries. The reference group's contribution to the foundation of knowledge has been included in this document.

Reference group members:
Siri Halsan, the Norwegian Association of Local and Regional Authorities
Marianne Lindheim, the Norwegian Association of Local and Regional Authorities
Pål Georg Rødsten, Oslo kommune
Kjetil Stavø Høvig, County Governor of Vestland
Ragnhild Sperstad Lyng, County Governor of Trøndelag
Kajsa Kemi Gjerpe, the Centre for Sami Studies (UiT)
Karen-Inga Eira, the Sámi University of Applied Sciences (Kautokeino)

The exam review group's mandate is to establish a foundation of knowledge on exams, assess input from the curriculum review groups, and examine how significant an impact the curriculum renewal

and technological developments should have on the exam system. The exam review group can also suggest possible adjustments and new exam forms within the following framework:

1. Marks awarded for classwork and examination marks are to be retained as final assessment forms.
2. The final assessment will continue to be individual and academic.
3. The current range and distribution between examination marks and marks awarded for classwork on the students' academic record are mainly to be continued. Minor adjustments can be considered.
4. The exam system should be feasible with approximately the same resources that it currently has.
5. The exam will act as quality assurance for the individual pupil through external assessment and the right to appeal.
6. The exam will continue to be used as a tool for quality development and assurance for school owners, schools and teachers.
7. The exam system will continue to be used as a source of skills development for teachers (examiner training, common indicators of achievement, and participation in assessment communities).

1.2 Background to the mandate

The curriculum renewal will result in an education programme that is more relevant to the future. The goal is to strengthen the development of the students' in-depth learning and understanding, as well as to highlight clear priorities in the subjects. The curriculum renewal includes an expanded definition of competence and a new overarching part that emphasises the core value. The school will prioritise skills such as interdisciplinary working, critical thinking and reflection, and creativity in the students' learning. At the same time, a technological development is happening, and the majority of social arenas are undergoing digitalisation. The use of digital teaching materials in school has increased, and technological development is driving change when it comes to lesson content. This development is clearly reflected in the new curricula.

The role of exams as an important part of a student's final assessment is highly legitimate and accepted in society. However, when we adjust the curricula, it is necessary to also review the assessment systems to ensure that good cohesion is retained between the two. In *NOU 2015: 8*, it was therefore recommended that a selection of experts be appointed to review the final assessment system and investigate how classwork assessments and exams can together provide reliable and relevant information about the students' skills. The exam system has been relatively stable, and the procedures haven't changed much over recent decades. In light of the curriculum renewal and the development of digital technology, the Norwegian Ministry of Education and Research has therefore decided to review the exam system.

The framework below shows the definition of competence as given in the Norwegian Knowledge Promotion Reform LK06 (Kunnskapsløftet) and the revised definition in the LK20 curriculum renewal. Understanding and the ability to reflect and think critically have been retained in the revised definition, and the ability to use knowledge and skills in familiar and unfamiliar situations has been added. The new definition therefore sets high demands and presumes cognitive transmission, which can again be connected to in-depth learning (the Norwegian Ministry of Education and Research, 2016). In addition, the students will be made capable of acquiring knowledge and skills themselves.

DEFINITIONS OF COMPETENCY	
the Knowledge Promotion Reform LK06 (Kunnskapsløftet)	**Curriculum renewal LK20**
Competency is the ability to solve problems and to master complex challenges. The students demonstrate competency in specific situations by using knowledge and skills to solve problems.	Competence is the ability to acquire and apply knowledge and skills to master challenges and solve tasks in familiar and unfamiliar contexts and situations. Competence includes understanding and the ability to reflect and think critically.

In *White Paper No. 28 (2015–2016)*, it includes the assertion that assessment forms and the quality assurance system must support schooling that places a greater emphasis on in-depth learning and systematic progression. An important aspect of the exam review group's mandate is to examine how and to what degree the revised definition of competence in the curriculum renewal can be reflected in the exam papers. It must be a goal to ensure that the students feel that the curriculum, schooling and exam systems correspond meaningfully to one another. In part five of letter of assignment 03-17 (2017), the Norwegian Ministry of Education and Research commissioned Udir to

1) assess the status of and describe ongoing processes in the work to develop the quality of overall achievement grades and exams in light of the parliamentary request for papers XII in Recommendation no. 19 to the Storting (2016–2017)
2) evaluate the need for new measures that can contribute to increasing the quality of final assessments.

In their response dated 1 June 2017, Udir suggested measures that included conducting a review of the exam systems and strengthening the foundation of knowledge we have on exams.

The exam review group was appointed in the last half of September 2018 and will conclude its work according to schedule in January 2020. The work consists of four partial deliveries:

Partial delivery 1	Preliminary status and assessment of the foundation of knowledge on exams
Partial delivery 2	Assessment of and input on the curriculum review groups' suggestions for the subjects' exam systems
Partial delivery 3	Assessment of how significant an impact the curriculum renewal and technological developments should have on the exam system, as well as suggestions for adjustments to the exam system to be made accordingly
Partial delivery 4	Suggestions for any new forms of exams and further direction of work on final assessments.

1.3 Methodology and purpose of the report

The exam review group worked with partial delivery 1 on the foundation of knowledge up to the deadline of 17 December 2018 (preliminary version) and then refined the report up to its publication

on 27 February 2019 (final version). Since the kick-off meeting in October, three group meetings have been carried out in connection with the foundation of knowledge, and the education sector has submitted its feedback through the reference group. A peer review has also been carried out.

This delivery should be seen as preliminary documentation of the foundation of knowledge and will be expanded upon in the subsequent partial deliveries. Advice on the subjects' exam systems will be delivered by the curriculum review groups by March 2019, and recommendations for adjustments to the exam systems due to the curriculum renewal and technological developments will be submitted for a final decision on 15 May 2019. The exam review group's final report to the Norwegian Ministry of Education and Research, including suggestions for new forms of exam and the further direction of work on final assessments, is due in 2020.

Generally, exam papers and matters such as use of study aids are debated equally in the public arena, but the exam system and quality have not been the subject of the same kind of research as national tests or the large international surveys, with the exception of a few small studies and questionnaires. The current basis for being able to evaluate whether classwork assessments and exams together provide reliable and relevant information on the student's competency is therefore limited and not very systematic.

As far as we are aware, there are only two previous reports that have affected exams, although this was not their main focus. One report, from Sjaastad, Carlsen and Wollscheid (2016), has examined the extent to which the number of teaching hours for subjects taught in upper secondary school are completed. The authors concluded that the exam period was the main cause of most class absences, and based on this report, the exam period in the school year will be changed. This report also highlighted other exam-related challenges, particularly the exam lottery (*trekkordning*). The Lied Committee (Lied-utvalget) has also investigated exams as part of its mandate. The partial recommendation describes the exam system in detail, and the committee has announced its intention to explore the system further in its main review, but that it wants to await the recommendations from the exam review group first (*NOU 2018: 15*).

There is currently no one definitive source of information about the exam system. Instead, knowledge is widely distributed throughout various types of sources. The purpose of this report is to collect everything we know about exam procedures, quality and results in a structured way in order to lay a better foundation for political decisions. We have selected both a theoretical-descriptive approach and an empirical-analytical approach, as both perspectives produce important information. In line with the exam review group's mandate, we are concentrating on the exam system in its entirety, while the subject-specific content in the exam (concerning its changes over time, see for example Nygård Arntzen, 2015; Smestad and Fossum, 2019) will be reviewed by the different curriculum review groups as part of the curriculum renewal.

The report is based on various types of knowledge relating to both advantages and disadvantages of the current system. Theoretically and empirically derived knowledge, such as measurement theory and validity research, is able to give us more concrete information on the exam system as long as it exists. Since the scope of this type of knowledge is extremely limited, we have, to a large degree, also incorporated experience-based knowledge, even though this type of knowledge is of varying quality and scope. However, while we are aware of its limitations, user experience can convey a lot of insight into the exam *in practice*. This experience-based knowledge includes information on the various forms of documented user insight, results from queries and technical reports. This type of foundation of

knowledge involves a certain degree of uncertainty, and we can only draw very tentative conclusions from it.

The exam review group has placed its primary emphasis on summarising the technical and organisational reports drawn up on behalf of Udir, as well as research on the Norwegian exam system. International research has been included where it is appropriate, but there is a lack of research on exams in this area as well. The studies that exist have largely been conducted from two different perspectives: a measurement perspective that emphasises studies on classic quality criteria, such as construct validity and examiner reliability, or a school perspective that emphasises studies on how exams can affect the school in practice. From a modern measurement perspective, the latter could be defined as consequential validity (AEA Europe, 2017). Our aim is to include both perspectives equally because we believe they contribute important aspects of our foundation of knowledge. We also want to be consistent in our concept use, even though this creates challenges where the concepts are complex and may be understood differently.

The exam review group is coordinated with the work on the curriculum renewal to ensure that final assessment systems are in place when the curricula is ready. This report will act as a foundation of knowledge for further work by the exam review group and the curriculum review groups as part of the curriculum renewal. The exam review group's mandate is related to the curricula in the curriculum renewal, which covers subjects in primary and lower secondary school, the core subjects in upper secondary education and training, and individual programme subjects[1]. Adjustments to the exam system for core subjects in upper secondary education and training will also impact students on vocational education programmes. This report does not mention exam systems in vocational educational programmes, but that does not exclude them from being taken into account when the subject of this report is the exam system in its entirety.

Some of the questions that the exam review group will touch on in their four partial deliveries are: How can the exam as a form of final assessment support and contribute to realising the intentions of the curriculum renewal, overarching parts of the curricula, and the revised definition of competence? Is, for example, the current "exam lottery" compatible with these aspects? And how can technological developments provide us with new opportunities to implement, further develop and evaluate the exam?

The purpose of the first part of this partial delivery is to summarise the frameworks for the current exam system, the background for these (section 2) and the exam's purpose as documented in the applicable regulations, as well as provide a description of guidelines and procedures (section 3). The second part of the report features a summary of what we know about the quality of the exam system. We define key assessment concepts (section 4) and delve deeper into the validity of the exam (validity) (section 5) and reliability (reliability) (section 6), we explore the relationship between exams and classwork assessments, which together comprise the final assessment system (section 7) and investigate what it means to assess competency in subjects (section 8). In section 9, we consider the students' experiences of the exam. The third part of the report represents an investigation related to the curriculum renewal and summarises the advice that the research can give us when it comes to testing the expanded definition of competence in the curriculum renewal (section 10) and the possibilities and limitations involved in digital technology (section 11). In addition, we briefly

[1] English language programme subjects, foreign language programme subjects, mathematics programme subjects for natural sciences, and mathematics programme subjects for social studies.

summarise what we know of the contribution of teacher training to the teacher's assessment competency (section 12). In a separate, final section (section 13), we draw some conclusions and highlight key problems and questions that we shall investigate further in the coming months.

Part 1

Background and purpose of the exam system in primary and lower secondary education in Norway

2 The emergence of the current exam system

This section draws on a historical perspective of the emergence of the exam system in Norway, with a view to understanding the role and purpose of exams in current primary and lower secondary education. Similar to Denmark and Sweden, Norway has a long tradition of basing admission to higher education on exams, which are administered by teachers in close cooperation with national authorities. The strong position that teachers have in the current Norwegian exam system is the result of an eighteenth-century development in which the exam system underwent a number of changes.

The development of the overall achievement grades also merits consideration in this historical account, which identifies five key developments that have contributed to forming the current exam system:

(2.1) The evolution of the teaching profession in line with the secularisation of society in the 18th and 19th centuries, where the role of priests was gradually superseded by that of teachers, who were recognised as qualified to administer final exams to students

(2.2) The development of "comprehensive schools" in the 19th and 20th centuries, which created new needs for adapted education for new groups of students and eventually for selecting students for higher education. This led to a demand for new theories and technology to improve the validity and reliability of the teachers' assessments (see section 2 of this report for definitions of the most important measurement concepts)

(2.3) Post-war discussions concerning norm-referenced versus achievement-based marking and other assessment principles (see section 2 for definitions), including the distinct camps in the conflicts over marking in the 1960s and 1970s

(2.4) Various reforms and the aim of evaluating the students' "overall competence" in the 1990s, including the wider objectives for education in the general curriculum

(2.5) The Knowledge Promotion Reform (2006) and associated regulatory amendments clarifying that only academic achievements shall form the basis for determining marks awarded for classwork, and that the assessment shall be based on the curriculum's competence aims

2.1 The evolution of the teaching profession in the 18th and 19th centuries

The origins of written exam systems can be traced as far back as the Han Dynasty in China, which lasted from 206 BCE to 220 CE (Eckstein and Noah, 1993, p. 2). However, exam systems first appeared on the European continent between the end of the 18th century and the beginning of the 19th century – a period characterised by close connections between the school and the church. According to Jarning and Aaas (2008), the Examen Artium in Norway and Denmark (established in 1808) and the Studentexamen in Sweden and Finland (established in 1824) corresponded to the German Abitur (established in Prussia in 1788) and the French BaccalaurĀat (established in 1808). The Examen Artium was initially overseen by the University of Copenhagen and later by the University of Oslo (established in 1811), and acted as an entrance exam to higher education. It was through the examination system that the content and quality of gymnasium education began to be regulated and controlled (Lundahl and Tveit, 2014).

In 1848, the Norwegian Urban Schools Act (Byskoleloven) and the Norwegian Rural School Act (Landskoleloven) established the introduction of a public examination, to be overseen by a priest (Lysne, 1999, p. 68). A fundamental change occurred in 1884 when Johan Sverdrup became prime minister and started to work with comprehensive schools. That same year, the Examen Artium was transferred from the university to approved high schools (ibid., p. 81). After a long fight for increased recognition of their professional qualifications for identifying the knowledge students would need at university, the gymnasium teachers gained full control over the Examen Artium (Lysne, 1999, p. 47).

This recognition represented a milestone in the development of the teacher's role in the current exam system.

2.2 The development of "comprehensive schools" in the 19th and 20th centuries

Towards the end of the 19th century and the beginning of the 20th, the school system bore signs of a class divide, as only the children of the bourgeoisie would continue to middle school after five years. However, Norwegian primary schools, which were established in 1889 as *folkeskoler* or "schools for the people", had the intention of "achieving a school system that presupposed that all children had a school community up to the age of 10" (Dale, 2008, p. 54). The vision of a "comprehensive school" was that it would provide all children with an equal opportunity to receive the best education. This educational-political strategy led to a greater awareness of the outcome of having an education.

Progressive teachers became the driving force behind exams because knowledge of the results of teaching and of students' expectations of teaching was of fundamental importance to reforms and measures in the development of comprehensive schools. Dale (2008, p. 54) concluded that "the exam research that was on the offensive in the 1920s and 1930s laid the foundation for ideas on adapted education and differentiation" (ibid.). Headteacher Hans Eitrem's analysis of marking methodology in the Examen Artium between 1920 and 1924 contributed to an even greater awareness of the comparability of teachers' grading (Lysne, 1999, p. 106). This analysis, together with research from Bernhof Ribskog, head of the Norwegian Teachers Council (Læreskolerådet) from 1936 to 1957, led to the discovery of significantly large discrepancies between the teachers' marking and the examination marks. The tendency to overuse positive marks (Lysne, 1999, p. 112–131) contributed to the introduction into the 1939 curriculum of a normal distribution in marking, as well as a minimum requirement as part of the basis of marking classwork assessments and the mark requirements for exams (Dale, 2008).

2.3 Ambiguous assessment principles from the post-war period to the 1990s

After the Second World War, further development of Norwegian comprehensive schools was the key project in educational policy. In contrast with Sweden (Sejersted, 2005), Norway retained the continental European exam tradition (Lundahl and Tveit, 2014). The exam gained an even more important function as a tool for government control when it came to the teachers' policies for marking. "When Norwegian *realschule* (*framhaldsskole*) and Norwegian middle school were merged into a nine-year primary and lower secondary school in the 1960s, the demand for norm-referenced assessment increased" (Dale, 2008, p. 234). Dale highlights that "both school politicians and prominent teachers were focused on being able to compare the marks from primary and lower secondary school and from the gymnasium on a national basis" (ibid., p. 235).

For these reasons, it was suggested that a central exam board "for the production and administration of standardised tests" be introduced (ibid.). There was a desire to establish systems that could ensure "consistent marking in all schools in order to create equal conditions for competition" (ibid.). However, the Research Council for the Norwegian School System's (Forsøksrådet) (1969) trial of "standardised tests" did not gain approval as a standard assessment system. Shortly afterwards, the Norwegian Council for Higher Education (RVO) was established, and the exam was a principle focus of its work.

In the 1960s, a lot of frustration and dissatisfaction with the normal distribution in marks awarded for classwork from the 1939 curriculum had been building, for reasons including that many teachers were following this distribution much more than there was a basis for. This meant that it was easier to get good marks in a class with a low-achieving student group than it was in a class with a high-achieving group – an important reason why opposition to the marking system increased as it became increasingly important in the 1960s and 1970s.

The 1970s political discourse on decentralisation and education outside the school also acted as a

backdrop to the opposition. The discussions that followed were often of an ideological nature. Dale (2008, p. 236) highlights an attitude that "an important means of emphasising education outside the school was to dismantle the formal assessment system in the school". In a *NOU* from 1974, it was emphasised that a mark-free school would "increase the regions' opportunities to utilise schooling on its own terms" (ibid.).

However, the marking system was retained, and it therefore became necessary to find a form of relative marking other than norm-referenced assessment. Achievement-based grading means that the target/achievement descriptors are derived from more precise assessment criteria for indicating degrees of achievement for each mark level (Lysne, 1999, p. 39). From 1968, an achievement-based grading system was developed for assessments in upper secondary school. However, the achievement formulations were "not followed up with investigative work in order to develop common assessment criteria" (Dale, 2008, p. 236). Lysne (1999) describes the system for marking in the gymnasiums as "individually adapted goal-related assessment" (*tillempet målrelatering*). Emphasis was placed on the choice of teaching material, teaching aids and working methods with a view to aligning the individual student's learning process to a greater degree with their personal growth towards a range of targets for their education. According to Lysne (2004), this in practice entails an "estimated overall evaluation" (p. 120). In 1981, the principle of individually adapted goal-related assessment was also introduced into primary and lower secondary school (ibid., p. 197).

2.4 Reforms of the 1990s (Reform 94, L97)

In the education reforms of the 1990s, the principle of individually adapted goal-related assessment was formalised when both the general curricula goals and the subject curricula goals were included in what was referred to as "overall competence" (the Norwegian Ministry of Education, Research and Church Affairs, 1996, p. 13). The principles and guidelines of educational reform L97 laid out instructions for organisation and working methods in order to connect the general curriculum and the subject curricula's content. Similarly, it was attempted to integrate the general curriculum into the subjects of Reform 94 through a separate part of the curriculum called "common goals for the subject". The exam system remained unchanged in this period, as the discussions concerned assessment principles.

Lysne (1999) concludes that individually adapted goal-related assessment allowed "excessive room for individual estimation and can therefore provide little in terms of comparable marks" (ibid. p. 39). The principle contributed to large differences in overall achievement grades across schools and teachers. Multiple surveys have determined that the teachers lacked a common frame of reference due to unclear or ambiguous national criteria. Instead, the assessments still bore signs of normal distribution grading related to the student group (Hovdhaugen et al., 2014; Kommunerevisjonen in Oslo, 2013; Lie, Hopfenbeck and Turmo, 2005; Prøitz and Spord Borgen, 2010; Steffensen and Ziade, 2009), even though this group size is far too small for the principle of normal distribution to be applied.

2.5 The Knowledge Promotion Reform (2006) and the subsequent clarifications of regulations

In *White Paper No. 30 (2003–2004), Culture for learning* (*Kultur for læring*), it was recognised that "assessment of the students' 'overall competence' has been a contributing factor to decisions concerning individual assessment and curricula possibly being perceived as ambiguous, especially in upper secondary education and training (the Norwegian Ministry of Education and Research, 2004, p. 39). The Knowledge Promotion Reform clarified that only academic achievements shall form the basis for determining marks awarded for classwork, and that "the assessment shall not be norms-referenced" (ibid., p. 40). The scale of marks moved from norms-referenced connotations to achievement-based assessment, and guiding indicators of achievement were introduced (see section 2 of this report for more details). From a school history perspective, the achievement-oriented

structure in the LK06 curriculum represented a break from previous curricula's orientation towards activities and content descriptions.

In summary, this brief historical view of the development of the exam system demonstrates that the teaching profession's strong support of the exam system can be understood in light of the professionalism, authority and legitimacy that teachers have through being recognised as competent to assess the quality of students' performances and thus to control admission to higher education and professional life. The exam system in its entirety and the most important procedures have been relatively stable over the last few decades, but assessment principles and criteria have been debated significantly and changed over time. Marking has moved from a norm-referenced assessment principle to an achievement-based one, while the assessment criteria have been ambiguous throughout long periods of the 1900s, which has had negative consequences for the assessments' quality.

3 The exam's purpose and organisation

This section first summarises the exam and exam system's formal functions as they are defined in the Norwegian Education Act and associated regulations, and we briefly analyse which perspectives on the exam these formal definitions reflect.[2] We then describe how the exam is organised in order to fulfil its purpose. Exams are a part of the final assessment system and therefore have a specific framework, as well as some distinctive characteristics, such as the exam lottery and the private candidate scheme. Finally, we document changes to the current exam system that have taken place over the last few years. The changes are a result of input from users, officials and professionals and reflect two challenges in particular: that the exam system in a subject can limit the students' opportunity to clearly demonstrate their final competence, and that the increasing access to study aids requires new discussions on what is to be assessed and how.

3.1 The exam's formal purpose as a part of the final assessment system

The exam's role in the regulations associated with the Norwegian Education Act

Section 3-17. Final subject assessment

The final assessment shall provide information on the competency of the student, apprentice, candidate for experience-based trade certification and trainee upon completion of their education in the subject in the curriculum (cf. section 3-3).
Final assessments in primary and lower secondary school take the form of marks awarded for classwork and examination marks.
Final assessments in post-16 education take the form of marks awarded for classwork, examination marks, and marks for trade and journeyman's examinations, experience-based trade certifications, and tests of competency.
Final assessments are individual decisions and can be appealed in accordance with the rules laid out in section 5.
Students in primary and lower secondary education who have an individual curriculum shall be assessed according to the collective competency aims in the curriculum for the subject (cf. section 3-3).

Section 3-25. General provisions

The exam shall be in accordance with the curriculum.
The curriculum of individual subjects determines if and when the course shall have an exam. The curriculum also determines whether the student must sit an exam or whether they will be drawn for the exam in the exam lottery, what form the exam shall take, and whether the exam will be locally or centrally administered.
The department determines how many exams will take place in the Norwegian tenth grade (for students aged 15 to 16) and quarterly for the education programmes or programme areas in post-16 education.
Students at a junior level who finish a subject earlier than planned in the curriculum shall enter into the draw for that subject's exam for the year the subject was completed (cf. section 1-15). In accordance with the first sentence of section 1-15, if the student is drawn for the exam after the first point, they will sit this exam in addition to the other exams that the department has determined that they shall take in the tenth grade.
The exam shall be organised so that the student or the private candidate is able to demonstrate their competency in the subject.
The examination mark shall be determined on an individual basis and convey the competency of the student or private candidate as demonstrated in the exam.
The county council is obliged to inform students and private candidates in post-16 education of the rules that apply for new exams, postponed exams and extraordinary exams.
There are no new exams, postponed exams or extraordinary exams in primary and lower secondary school.
For trade and journeyman's examinations and tests of competency, the rules laid out from section 3-48 to section 3-68 apply.

The final assessment shall provide information on the competence of the student, apprentice, intern and trainee upon completion of their education in the subject (cf. section 3-17 of the regulations

[2] For a detailed analysis of the exam's more implicit and non-formalised roles in practice, see section 5.3.

associated with the Norwegian Education Act). As with the marks awarded for classwork, examination marks are a final assessment. They therefore both show the student's competency upon completion of the subject, but the regulations do not further clarify the interaction between these different marks.

The exam, including any preparatory periods, shall be in accordance with the curriculum and organised so that the student is able to demonstrate their competency in a subject. The examination marks shall be determined on an individual basis. Even though marks awarded for classwork and examination marks are equal forms of final assessment, the exam in some cases holds greater significance than the summative assessment. If a student has received a 1 (equivalent to an F) as a mark awarded for classwork, they will still pass the subject if they receive a 2 (equivalent to an E) or higher in the exam. This is in accordance with section 3-4 of the regulations associated with the Norwegian Education Act. The subject will never be passed if the exam is not passed.

Together with the marks awarded for classwork, the examination marks shall be entered onto a student's academic record and constitute the basis for admission to higher education and working life. This comes both from the rules on admission to upper secondary education and training in ch. 6 of the regulations associated with the Norwegian Education Act and from regulations concerning admission to higher education.

The two purposes of the exam – certification of competency and ranking of applicants – are both directly related to students. In addition, the regulations associated with the Norwegian Education Act charge school owners to "contribute to establishing an administrative system and to gathering information, both statistical and otherwise, that is needed to assess the condition and development of education" (section 2-2). The exam results are a part of this, even though this purpose is considerably less clear and can rather be seen as a means of quality assurance.

The quality assurance system shall contribute to quality development and transparency of and dialogue about the school's activities. It shall also lay the foundation for quality development of the individual school.

The *Education Reflection Report* (*Utdanningsspeilet*) contains figures and analyses of nurseries and primary and lower secondary education in Norway, and includes descriptions of the student's learning outcomes in terms of their marks awarded for classwork and examination marks.

The mark average and distribution of individual marks on a national level usually do not vary significantly from year to year. However, the mark average for counties may vary a fair amount between years, particularly in subjects with few students. The examination marks are an expression of the competency the student has demonstrated in the exam. If the papers differ from year to year, some variation in the mark average and distribution is to be expected. This means that the exam results are not directly comparable from year to year. Therefore, they cannot be used to form any judgements on changes in performance across classes.

The Norwegian Directorate for Education and Training (2018)

3.2 Regulatory frameworks for final assessments

Assessment of the students' final competency is regulated in ch.3 of the regulations associated with the Norwegian Education Act, and similarly in ch. 3 of the regulations associated with the Norwegian Charter Schools Act (Friskolelov). The curricula are also regulations. There is a clear connection between curricula and the regulations associated with the Norwegian Education Act. For example, according to section 3-3 in the regulations, the basis for assessment is the competence aims in a subject. The regulations set out guidelines for assessment practices, responsibilities and roles, and the exam forms are described in the curricula. The responsibility is distributed between national and local authorities and schools/teachers.

Udir is responsible for the development, implementation and management of the continuous test and assessment system. Within the system, Udir is responsible for centrally administered exams for students in the tenth grade and upper secondary education and training. The municipality and the county council are responsible for the locally administered exams in primary and lower secondary school and upper secondary education and training, respectively.

Exam forms in primary and lower secondary education

Centrally (national) administered exams
Written

Locally administered exams
Written
Oral
Practical
Combined oral and practical

Locally administered exams in upper secondary education and training may be oral, written, a combination of oral and practical, or practical. In addition, there is an interdisciplinary practical exam (which includes the common programme subjects) for the vocational programmes in upper secondary level 2. In primary and lower secondary school, only a local oral exam is administered. In addition, for example, the tenth-grade natural sciences curriculum provides guidelines for oral exams with practical elements.

The overview below shows the number of marks awarded in the school year 2015–2016 distributed across locally administered exams, centrally administered exams and classwork assessments. It should be noted that the exam lottery at each stage of upper secondary education and training means that the proportion of students who are drawn for centrally and locally administered exams can vary from year to year.

Marks awarded for examinations and classwork 2015–2016 (in number and %[2])

	Norwegian tenth grade (students aged 15–16)		Upper secondary university-preparatory subjects[3]		Upper secondary vocational subjects		Total upper secondary subjects	
	Number	%	Number		Number	%	Number	%
Locally administered exams	57,131	5%	62,994	7%	42,816	11%	105,826	8%
Centrally administered exams	74,435	7%	139,051	15%	2,678	1%	141,757	11%
Marks awarded for classwork	932,854	88%	736,711	78%	359,488	89%	1,084,480	81%
Total	1,064,420	100%	938,756	100%	404,982	100%	1,332,063	100%

This overview portrays the scope of marks that are awarded every year within the different categories.

[3] Including supplementary studies qualifying for higher education (upper secondary level 3), General Studies in Fishing and Forestry (upper secondary level 3), and General Studies in Media and Communication (upper secondary level 3). Private candidate exams and trade and journeyman's examinations are not included.

In both primary and lower secondary school and upper secondary education and training, each student completes a limited number of exams, and marks awarded for classwork therefore comprise a large majority of the marks on the student's academic record. The overview also highlights that the county council, the municipality and the school/teacher have significant roles and responsibility in the current final assessment system, including ensuring that the marks provide reliable and relevant information on the student's competency.

Udir has developed a framework for centrally administered written exams (Norwegian Directorate for Education and Training, 2018c) The aim is to establish a common foundation for quality assurance and quality control in Udir's work with exams.

The organisation of work with centrally administered written exams

The work with centrally administered written exams in Norway involves multiple various authorities who have various areas of responsibilities.

Udir is responsible for the development, implementation and management of the cohesive examination and assessment system. This includes centrally administered exams and associated information and guidance material. **Udir** can also annul an exam.

The county governor appoints examiners for the different subjects in primary and lower secondary school and upper secondary education and training according to recommendations from school leaders/headteachers. They also choose exams coordinators for primary and lower secondary school and for Norwegian in upper secondary level 3 for their region. The county governor is responsible for carrying out collective marking and managing the appeals procedure.

Municipalities and counties are locally responsible for centrally administered written exams in primary and lower secondary school and in upper secondary education and training. This applies to both centrally administered written exams and all locally administered exams (written, oral, practical and combined oral and practical). They are also responsible for selecting subjects and candidates from the exam lottery, according to the framework set by Udir. This also applies to charter schools. For the vocational subjects, the counties are additionally responsible for marking, appointing examiners, and appeals.

The examination boards prepare exam papers and are responsible for ensuring that they correspond with the curriculum and relevant conditions, on behalf of and in cooperation with Udir.

External consultants provide feedback to the examination boards and Udir on draft exam papers. External consultants shall advocate on the candidate's behalf and are an important part of assuring the quality of exam papers.

Examiners are responsible for marking exam responses in line with the competence aims in the curriculum and the features of achievement in the exam guidance on behalf of the county governor.

* In Norwegian upper secondary education and training, the term *formøteleder* is also used

(Norwegian Directorate for Education and Training, 2018c)

The regulations differentiate between centrally administered and locally administered exams when it comes to the range of knowledge they should test. Udir determines how the exam in individual subjects

Some frames for *locally administered exams* (cf. sections 3-29, 3-30):

- Written exams – up to 5 hours
- Oral exams – up to 30 minutes per candidate
- Combined oral and practical exams – up to 45 minutes per candidate
- Practical exams – up to 5 hours

Locally administered exams in lower secondary school are only oral.

Oral exams shall be accompanied by a preparatory period, wherein the candidate shall receive a topic or problem 24 hours before the exam itself. This period shall not be included in the basis for assessment.

The county decides whether private candidates shall receive this preparatory period.

The county decides whether the other locally administered exams shall have a preparatory period. This preparatory period may last for up to two days and shall not normally be included in the basis

shall be organised and what the exam papers shall be for centrally administered exams. This is stipulated in section 3-28 and is in line with the overarching regulations mentioned above. The time frame for centrally administered written exams is usually five hours, according to section 3-28a. Locally administered exams come in multiple forms with various time frames (see the text box).

The overarching regulations for centrally administered exams also apply to locally administered exams (mentioned above). In addition, the regulations contain the guidelines that locally administered exams shall provide the candidate with the opportunity to show their competence in as much of the subject as possible (cf. sections 3-29, 3-30). During the exam, the candidate shall be tested in more relevant aspects of the curriculum than can be garnered directly from any preparatory period. During an oral exam, the candidate shall present the topic or problem that they prepared in the preparatory period.

During the oral exam, the examination and assessment shall happen in "real time", where the dialogue between the candidate and the examiners is an important aspect of the exam. In written exams, both the formulation of the exam papers and the marking happen independently of the exam procedure itself and are verifiable. The amount of the curriculum that will be tested in the exam will therefore vary in terms of, for example, how the competence aims are formulated and the nature of the subject itself. The exam forms set guidelines for how far the candidate shall demonstrate their competency, either orally, practically or in writing. For individual subject curricula, written and oral skills are part of the competency to be tested (e.g. for language subjects), but for the majority of the curriculum this is not the case. In some cases, written or oral skills may affect the student's opportunity to demonstrate their competency in the subject.

3.3 The exam lottery (*trekkordningen*)

The exam lottery is discussed in *White Paper No. 20 (2012–2013)*: "The exam lottery (trekkordningen) means that students shall not be examined in every subject but shall instead prepare for exams in the subjects for which the exam is a possible final assessment in addition to the marks awarded for classwork" (the Norwegian Ministry of Education and Research, 2013, pp. 65–66). This means that the students may be examined in any subject, but that they don't know which subject until the lottery results are published. The exception from the exam lottery is that all the students who take Programmes for General Studies or the supplementary programme for general university and college admissions certification as part of a vocational education programmes shall sit the centrally

administered written exam in their main form of Norwegian (Bokmål or Nynorsk).

The exam lottery means that student groups are split up. Essentially, this shall be based on a random selection, in line with the principle of randomisation. However, it should be taken into account that as teachers have particular student groups in multiple subjects, they may have multiple groups taking exams on the same day or may be appointed as an external examiner at other schools for these days. One consequence of the exam lottery is the number of exams per student on their upper secondary education and training academic record, which may impact on the number of marks that form the basis of the average admission rates to higher education. For example, only 20 per cent of the students in upper secondary level 1 are drawn for an exam in the exam lottery (see the overview below).

The tenth-grade exam lottery

All students shall be drawn for a centrally administered written exam (mathematics, norwegian or english) and a locally administered oral exam.

The exam lottery in Programme for General Studies

- *Upper secondary level 1* Approximately 20 per cent of the students shall be drawn for an exam in one subject, which may take the form of a written, practical, oral or combined oral and practical exam.
- *Upper secondary level 2* All students shall be drawn for an exam in one subject, which may take the form of a written, practical, oral or combined oral and practical exam.
- *Upper secondary level 3* All students shall sit an obligatory written exam in their native language: their main form of Norwegian (Bokmål or Nynorsk) or Saami. For all students, an exam in the student's alternative form of Norwegian (Bokmål or Nynorsk) is part of the lottery. In addition to obligatory exams in their native tongue (either their main form of Norwegian or Saami), the students on the Education Programme for Specialisation in General Studies shall be drawn for two written programme subjects in addition to one oral, practical or combined oral and practical exam.

The exam lottery in vocational education programmes

- *Upper secondary levels 1 and 2* All students in upper secondary level 2 are required to take an obligatory multidisciplinary exam in a programme subject. In addition, approximately 20 per cent of students in upper secondary levels 1 and 2 shall be drawn for one subject. The 20 per cent shall be seen over a two-year period.
- *Upper secondary level 3* Supplementary programme for general university and college admissions certification: In addition to the obligatory exam in Norwegian, the students shall be drawn for one written and one oral, practical or combined oral and practical exam. For all students, an exam in the student's alternative form of Norwegian (Bokmål or Nynorsk) is part of the lottery.

Basing the exam lottery on the principle of randomisation has been debated for a long time among members of the School Student Union of Norway[4]. The students perceive the current exam lottery to be unfair because it does not provide everyone with the same opportunity to demonstrate their

[4] https://www.utdanningsnytt.no/nyheter/2018/november/elevorganisasjonen-vil-endre-eksamensordningen/ (in Norwegian)

competency, which is contrary to one of the two formal main aims described in the applicable regulations (see 3.1 and 3.2 above).

In addition, the day of the lottery itself is associated with a lot of pressure and stress among the students. In 2017, a multipartite workgroup appointed by the Norwegian Ministry of Education and Research delivered a report that assessed possible ways of organising the school year in light of challenges connected to teaching time, exams and exam preparations (the Norwegian Ministry of Education and Research, 2017). As a follow-up to this work, Udir has submitted four changes for further investigation on behalf of the Norwegian Ministry of Education and Research. The deadline for this is 21 March 2019.

Among other things, the workgroup concluded that a longer preparatory period between publishing the results of the lottery and the day of the exam is important, as the weeks before the exam can be planned more effectively. It also touched on the problems: What would happen if the students and the teachers were more aware of which subjects the students would have exams in beforehand? Are there arguments in favour of the draw for written exams being known longer in advance if it makes the school year easier to plan, so that the number of teaching hours for the subject in the year are met? On the one hand, students will have more time to prepare and immerse themselves in the subject if the lottery is drawn earlier in the school year. On the other hand, an earlier lottery may indirectly contribute to the lessons being dictated by the exam to a greater degree than they currently are. If the students know which subjects they shall be tested in early in the school year, this may impact the academic priorities of the students and the teachers and have a detrimental effect on the lessons in the subjects for which the students shall not sit an exam.

3.4 Arrangement for external candidates (privatistordningen)

Arrangement for external candidates is an alternative way to acquire vocational or academic qualifications. The original purpose of this scheme was for there to be an offer to document competency in a subject that the candidate either has not completed any education or final assessment in previously or wishes to improve their mark for. In Norway, the first group is referred to as "first-time external candidates" (*førstegangsprivatister*) and the latter group is referred to as "improving external candidates" (*forbedringsprivatister*). There are also students who have dropped individual subjects early or would like to retake subjects while they still have the formal status of student. The scheme arose prior to youth receiving the statutory right to upper secondary education and training through Reform 94.

In upper secondary education, the content of exams for private candidates is identical to that for current students, and the exams are centrally and locally administered and may be either written or oral. Registration as a external candidate in upper secondary education and training in Oslo from 2017 shows that the majority of external candidates take one exam, and almost 40 per cent have registered for two or more exams (the Norwegian Ministry of Education and Research, 2017).

Nesman and Kovač (2016) point out that the external candidate scheme may represent a good solution for individual groups, but that external candidates as a group have changed over time. Their studies demonstrate that external candidates are a disparate group, wherein the majority (52%) are currently taking exams to improve their mark in a subject they have already passed. A large proportion of external candidates are therefore still students in standard education. Students at a lower secondary level who have demonstrable competency in a subject to be able to study it at an upper secondary level may be able to skip subjects, according to section 1-15 in the regulations associated with the Norwegian Education Act. In addition, the arrangement for external candidate may have some unfortunate side effects. Many of those who register for exams as external candidates do not show up. Conversely, individual students may opt to not participate in lessons but still take exams as external candidates while they have student status.

These changes and challenges may be a reason to review how the external candidate arrangement works in practice. As an answer to letter of assignment 13-12 from the Norwegian Department of Education and Research, Udir recommended some measures that will contribute to reducing the extent of the scheme and preserving its original purpose. Examples of these measures include an increase in fees for external candidates and stricter requirements for first-time academic records.

Some changes have already been implemented in previous years. From 1 January 2018, the external candidate arrangement for programme subjects in vocational education programmes was adjusted, and it now enables all external candidates to take the exam in individual programme subjects. Previously, external candidates had to take the exam in more subjects than the ones they lacked a mark for. Figures from 2017 demonstrated that nearly 40 per cent of the external candidates would sit two or more exams in the spring of that year.

3.5 Development and changes in the exam

Administrative and content-related conditions of the exam are continuously under development due to input from users, officials and professionals. There are two challenges that have been particularly highlighted in the last few years: that the exam system of a subject itself can limit the students' opportunity to clearly demonstrate their final competency, and that the increasing access to study aids requires new discussions on what shall be assessed, and in which ways. Udir has therefore implemented some changes to the exam system in individual subjects, which are described in the next paragraph.

The role of study aids and resources in the associated regulations

According to the Knowledge Promotion Reform, the majority of subjects allow students to take study aids into the exam, but there are differences between what is allowed for written and oral exams. Udir determines which study aids are allowed to be used in each subject for centrally administered exams, and school owners determine this for locally administered exams (see section 3-31 in the regulations associated with the Norwegian Education Act and section 3-29 in the regulations associated with the Norwegian Charter Schools Act). For oral exams, the only study aids that the student or private candidate is allowed are their notes from the preparatory period. Simultaneously, the regulations specify that study aids must not undermine the basis for assessing the competence of the student or the private candidate (section 3-31). Udir has received feedback through examiner reports from and during examiner training for centrally administered written exams, wherein a number of examiners expressed that it may be challenging to assess the degree to which the students themselves have written the texts when they do not refer to sources they have (probably) used in their responses.

Digital resources for exams

The use of digital teaching materials in school has increased, and this has a significant effect on the study aids the students may use in an exam. In multiple subjects, part of the competency may involve being able to, for example, gather, assess and use sources in a relevant and verifiable way, including internet sources.

Since 2015, the students have had access to a selection of online resources in centrally administered exams in addition to other study aids. In 2017, there was clarification with a view to contributing to increased equal treatment of candidates within the same county. It is now obligatory for the counties and municipalities to offer a selection of online resources, and all candidates in the same county shall have access to the same online resources for an exam. As previously, it has been emphasised that students shall be made aware of online resources, and the selection must happen as part of a collaboration between schools and school owners.

Trial of unrestricted internet access during an exam

A trial of unrestricted internet access has also been conducted on the exam day itself in subjects where this is relevant. In the final report from the trial of unrestricted internet access during an exam (2012–2015), Rambøll writes (2015) that the majority of the students and teachers involved were satisfied with the exam form, and that it contributed to promoting relevant new pedagogical practices and competences in education. This exam form has led to some technical and/or practical challenges. The final report states that there were a few indications of an increase in scope of cheating and plagiarism, or that the exam form has a demonstrable impact on the students' exam results, both positively and negatively. The access to the internet was expanded to include all candidates from all schools sitting an exam in a particular subject from spring 2018.

The examiners were more critical of unrestricted internet access. Six of the ten examiners thought that accessing the internet in an exam is well suited to assessing the students' competency in a subject, but only three of the ten examiners thought the scheme should be continued, and even fewer (16%) wanted it to be the standard for all exams. Even though the examiners were divided in their responses, they saw little difference in answers with or without internet access. However, 73 per cent of the examiners and 51 per cent of the teachers thought that access to the internet makes it easier to cheat than in other exams taken on a computer. The teachers also clearly stated that high-achieving students are able to better utilise unrestricted internet access than low-achieving students (Rambøll, 2015).

A methodological limitation that it is important to highlight with these studies is that the schools themselves chose to participate in this trial. In other words, the entire evaluation of unrestricted

internet access during an exam has been conducted at schools that chose to test out this type of scheme. The schools can therefore probably be considered as a positive selection, which means that the findings cannot necessarily be generalised to apply to other schools.

Two-part exams – an example from mathematics

Two-part exams in mathematics, chemistry, physics and biology were introduced in accordance with the Knowledge Promotion Reform in 2008. Two-part exams in socioeconomics were introduced a little later (2013) due to an evaluation conducted by Rambøll and the Norwegian Institute for Teacher Education and School Research (ILS) in 2010. The introduction of two-part exams was justified by the fact that it was challenging both to test and to assess students' wider competency in the subjects where study aids were either allowed in or banned from the exam (with the exception of the internet and tools that allow for communication).

The change thus facilitated a more comprehensive testing of the student's competence and should be able to provide a better foundation for the marking. For example, in mathematics, the first part of the exam is without study aids and enables the students to be tested in mental arithmetic or estimation, while the second part of the exam allows students to use digital tools to solve more complex exercises.

An evaluation of the exam in maths for tenth-grade students in spring 2018 shows that the quality of the exam was consistently evaluated as good when it came to the correlation between exam papers and schooling, that the papers were understandable and appropriately challenging and that the scale corresponded to the time the students had available (Bjørnset et al., 2018). However, it also revealed that the students had unequal access to and training in digital tools both at school and in the home. This finding may mean that student groups with greater access to and more training in digital tools had better chances of success with the exam than students who did not have these opportunities. Bjørnset et al. (2018) highlight this as "a key mechanism for inequality creation", admittedly without being able to conclude anything from the data source on what role this mechanism had possibly played in the year's exam.

Level I foreign languages

A trial of a new exam system was conducted for selected level I exams in foreign languages in spring 2015, autumn 2015 and spring 2016 in Finnmark, Rogaland, Sør-Trøndelag and Troms. In the trial, the exam system was changed from a written five-hour exam to a combined written and oral exam. The background to the suggestion of this trial was both a large and systematic discrepancy between marks awarded for classwork and examination marks over time, and a desire to try out an exam form that could give students a better opportunity to demonstrate their competence in the subject and that tested more skills than just reading and writing. Both teachers and students provided feedback on whether they thought the model contributed to the students being able to demonstrate more of their cumulative knowledge in the foreign language. The marks that were reported showed no large increase in the mark average but were somewhat better than the marks in the standard five-hour written exam.

The model that was tested for level I foreign languages would have led to a number of administrative and financial challenges, particularly for the exam boards conducting exams for private candidates. Changes to the exam system would have meant an increase from one to two exams for all private candidates in foreign language. Therefore, the exam system was not changed following the conclusion of the trial, even though both teachers and students were positive about the new model.

Locally administered exams

Oral exams are administered locally. Oral exams comprise a majority of the total number of exams a student shall sit throughout their schooling. All students have at least two oral exams throughout their primary and lower secondary education. The rules for locally administered exams were changed by regulations on 26. September 2013. The purpose of these changes was to clarify the rules for oral exams and ensure a more unified national practice. The rules shall better facilitate the expectation that exams be predictable and fair for all students.

The changes caused a number of enquiries about what would be included in the basis of assessment. Feedback from the sector and the county meetings about the exam indicates that there are still schools that struggle with understanding how to interpret the decisions as they are described in the regulations ("the preparatory part shall not be included in the basis for assessment") and in circular Udir-2-2014 ("It is the competence the student demonstrates during the exam itself that the examiner shall assess. The notes that the student has produced in the preparatory period, for example the presentation, are not a part of the basis for assessment".). The latter has been clarified in the paragraph "Basis of assessment", where it states that "the academic competence the student demonstrates through the way the topic/problem is presented is also part of the basis of assessment of the student's overall competence". However, it has become apparent that the term "way" is subject to interpretation and not uniform practice across schools.

Trials of locally administered exams in practical subjects and the arts
As part of *White* Paper *28 (2015–2016), Subject – Specialisation – Understanding* (*Fag – Fordypning – Forståelse*), the Norwegian Ministry of Education and Research wishes to assess whether the practical subjects and the arts, including food and health, physical education, arts and crafts, and music shall be part of the exam lottery for locally administered exams in the tenth grade. Udir has therefore been commissioned to organise a trial of an exam-style test in these subjects.

The purpose of the trial is to gather experience in order to be able to make a decision on whether the subjects shall be entered into the exam lottery and on the form that this exam would take:

- oral exam with practical elements, lasting for 30 minutes per student with a 24-hour preparatory period
- a combined oral and practical exam, lasting for 45 minutes per student with a preparatory period of up to 48 hours

In the school year 2017–2018, 16 schools in 4 counties participated in this trial. In the school year 2018–2019, 19 schools in 5 counties participated. These schools were able to decide the form of exam for this trial themselves.

In total, 40 teachers tried out *an exam-style test* in practical subjects and the arts. Ten teachers participated in the trial in each of the subjects, and half of them tested each of the two forms of exam. Schools and teachers who participated in the 2017–2018 trial are positive about the trial. In a Questback survey in June 2018 the teachers answered that both they and the students were satisfied with the implementation. They had achieved a good division between the practical and oral parts, and the students were able to demonstrate their practical and oral competency in a good way. In addition, multiple teachers highlighted that the trial had led to a change in attitude to the subjects at the schools by the way that students and teachers were talking more about the importance of practical subjects and the arts. The trial has clearly led to a greater awareness of the curriculum's competence aims and of assessment. 14 of the 16 teachers who conducted the trials thought that an exam can contribute to strengthening the subjects' status.

Identified challenges:

- Papers that shall allow the students the opportunity to demonstrate their competence and assessment in the exam itself
- Regulations, times for setting classwork assessments and the exam lottery

Part 2

The quality of the current exam system

4 Key concepts: quality criteria and assessment types

In this section, we seek to clarify some key concepts that may be used in the development of the Norwegian exam system. In the following section, we will use these concepts to examine our current foundation of knowledge. As early as in *NOU 2015: 8*, there were requests for reviews of how the overall achievement grades and the exam system may together provide impartial and relevant information on a student's competency in a subject. The *NOU* indicated that teachers and examiners should be supported in their assessments by explicit goals, assessment criteria, guidance and quality assurance. The Stoltenberg Committee (Stoltenberg-utvalget) also recommended setting stricter quality requirements when designing and testing exam papers (*NOU 2019:* 3). Any implementation of a review of final assessment quality should be rooted in test-theoretical and assessment concepts, such as validity and reliability, which preserve different aspects of quality. In addition, it's important to take an integrated approach that ensures agreement between the quality criteria.

A significant challenge when it comes to the quality of an exam is that it is extremely difficult to know whether an exam paper has the desired qualities before it is used, as the papers must be kept secret prior to the exam. For example, in the Netherlands, the papers for the following year were test driven on a selection of students the year before. Therefore, it is possible to test an exam paper beforehand, but it requires a consensus across all involved parties. To assure the quality of the exam, it is therefore necessary to draft overarching questions in addition to examining individual criteria.

4.1 Validity (*legitimacy*)

Validity, or legitimacy, should be considered the key assessment-theoretical concept for exams. Research indicates that the individual interpretations of various parties are a significant factor in understanding the validity of:

- how far an interpretation, decision or action is rational;
- the type of evidence, reasoning or criteria used to judge how rational an interpretation is; and
- methods for improving the rationality of interpretations, decisions or actions (Moss, Girard and Haniford, 2006)

Validating a test or exam involves developing an argument about the type of evidence that shall be considered valid and about the results that are to be interpreted (Markus and Borsboom, 2013). It is almost impossible to judge the general quality or "impartiality" of an exam, as this must always be discussed in the context of its purpose. *Validation* is therefore a key process, wherein an exam's legitimacy is investigated and documented in this context (Kane, 2015). Therefore, any change to an exam's context or purpose triggers a need for a fresh validation of the test. If a test or exam shall have multiple applications, each of these requires individual validation. Pellegrino, Chudowsky, Glaser and the National Research Council (U.S.) (2001) specify that the more purposes an individual test or exam has, the greater the threat to each of these purposes.

Discussions of validity should also include whether there are any unintended negative consequences of tests and exams for certain population groups (e.g. minority-language students), undesirable systematic effects (e.g. stress, anxiety) or intentional or unintentional retroactive ("washback") effects on the teaching (consequential validity; Kane, 2015). Ensuring validity in exams entails reviewing the whole process: from the development of papers, via administering the exam and interpreting the results, to the way these interpretations are applied.

The development of exam papers is a well-established area in assessment research, which has resulted in explicit frameworks that describe important quality criteria, as well as the steps to be taken throughout the development process (see, for example, AEA Europe, 2017; Wilson, 2005). Exam development begins with precise definitions of its content and the assessment criteria, and includes tests to ensure that the paper has the desired qualities *before* it is implemented, which especially applies in the case of *high-stakes* exams. For possibilities when it comes to competency testing, see section 9.

In an attempt to also construct a theoretical model for the next steps in the quality assurance process, assessment researchers have developed the so-called validity chain (Crooks, Kane og Cohen, 1996). The eight stages have been described and adapted below:

1. *Administering* the papers that the students shall complete during the exam
2. *Marking* the students' performance on the exam papers
3. *Aggregating* results from individual exercises to calculate partial or total scores on the exam
4. *Generalising* from specific exam papers and results to the target area that shall be assessed (e.g. by discussing what one longer writing task or a collection of smaller writing tasks may say about the expectations of the student's writing competence, as expressed in the curriculum's competence aims)
5. *Extrapolating* the achievements that are assessed in the exam to a larger target area (e.g. general writing competence), which encompasses all tasks that may be relevant within this larger area
6. *Evaluating* the student's performance (in an exam context, this will normally involve coming to a decision on marks and forming a justification for this decision)
7. *Deciding* which actions or measures are relevant in light of the result (e.g. a student may decide to appeal a mark, or teachers and school leaders may decide to review the school's exam practice within a particular area)
8. *Noting* the effect on students and others who are impacted by the exam practice's process, interpretations and decisions (Crooks mfl., 1996)

Typical threats to validity in the various stages may include: some students receiving help from teachers to complete the papers in the exam situation, while others don't (administering); teachers emphasising what is easy to assess in the marking without examining the more complex aspects of the student's performance (marking); results from different types of papers being summarised in an inappropriate way (aggregating); the exam not containing enough questions to test enough of the student's competence (generalising); the exam containing no exercises from important parts of the target area (extrapolating); the student's performance being judged by the curriculum's competence aims but without any evidence that these have been met by the student (evaluating); the requirements that form the basis for follow-up measures after the exam being far too high or low (deciding); and the exam process having a negative impact on many students (noting). The party or group responsible for assuring the quality of a test or an exam should evaluate what the weakest stages are and attempt to strengthen them accordingly.

4.2 Reliability (*dependability*)

Reliability, or dependability, shows the extent to which the results from repeated assessments correspond with each other (Pellegrino et al., 2001). This may concern multiple assessments of the same construct within a trial exam or reviews of a single exam paper conducted by multiple examiners. Reliability is considered as a necessary, but not solely sufficient, condition for validity.

High reliability is a prerequisite for assessment quality in order to avoid coincidences. A student's exam result, in the sense of the quantifiable information on the student's competence, should be as

independent as possible from the examiner marking their answers, from the form of exam used, from the content selected for that particular exam, and from the time the exam took place. The requirement is that neither different examiners nor a repetition of the exam on the same day, with different papers or in a different form, will produce a different result, in the sense of other quantifiable information on the student's competence.

However, it is accepted that a certain degree of variation in results is unavoidable. This variation is known as an observational error. The greater the consequences of a result for a student – for example, in cases of final assessments, since a student's academic record forms the basis of admission to higher education and professional life – the more important it is to reduce observational errors and increase reliability as much as possible. Individual papers are often not very reliable, no matter whether they are standardised or non-standardised. Therefore, it is advisable to use as many different types of papers as possible and to let these be assessed by different examiners.

A challenge for assessments with a limited time frame is that aiming to increase their reliability may result in the exam papers being narrowed in their design and impact instead of being expanded in their number and type – in other words, the emphasis is placed on ensuring consistent information rather than collecting evidence for wide and important learning goals (Broadfoot, 2007). This problem highlights the necessity of having overarching frameworks for quality, systematic routines for monitoring quality, and accessible documentation of the processed results. Investigating reliability requires data that is as specific as possible. Usually, this will mean saving data from the examination results on a student level about each exam paper and assessment criterion.

4.3 Fairness (*impartiality*)

Fairness refers to the idea that all students must have the same chance to demonstrate their competency during an exam. In practice, this means that the exam is free from systematic inequalities for the group taking the test. It also means that the exam shall not be impacted by variables such as gender, language background, functionality, geographical location and similar.

Fairness is used when discussing a number of possible related problems: whether the exam papers give an advantage to individual student groups, whether all students are treated equally in the examination process and whether students have had the opportunity to learn what they are tested in (Pellegrino et al., 2001). There are also a number of additional factors that may impact the fairness of an exam. For example, the students' results may be impacted by language skills, motivation, fatigue, test anxiety, physical conditions when sitting the exam, or varying degrees of unethical exam preparation (Haladyna and Downing, 2005).

4.4 Assessment types (norm-referenced, achievement-based, standards-based and individual-based assessment)

As discussed in section 2, the theoretical basis for assessment has developed historically from being norm-referenced to being achievement-based, and in many educational contexts has recently developed further to being standards-based. The difference between the principles concerns what the assessment *is based on* or, rather, what it is compared with (William, 1996).

In a *norm-referenced assessment*, an exam response from one student is compared with responses from other students. A complete exam system will, as a rule, have a normal distribution of results that manifests as a symmetrical bell curve (also known as *the Gauss curve*). When the data set of results is large enough, it may be expected that the marks are distributed around a middle value. The majority

of students will receive a mark near this value, while the higher or lower marks are less frequent. It should be noted that this assumption does not apply to smaller units such as a class or school. However, many teachers have previously used this norm to assess their students' learning outcomes, which means that it is easier to receive high marks in a lower achieving class, and vice versa (see sections 2.3 and 2.4 for more information). In addition, norm referencing is problematic because, while the principle is suitable for ranking learning outcomes, it does not communicate expectations and requirements to the students. It therefore does not provide the teachers with a means of communicating with the students.

This criticism of the norm-referenced evaluation and assessment tradition formed the basis of the development of what is referred to in Scandinavia as *achievement-based assessment*. In American terminology, this was initially called *criterion-referenced assessment* (Popham and Husek, 1969), often referred to in Norwegian as *kriteriebasert vurdering*. *Achievement-based assessment* requires using explicit criteria as a basis for being able to assess achievement. According to Glaser and Klaus (1962), the difference between achievement-based and norm-referenced assessments is: "[c]riterion-referenced measures depend on an absolute standard of quality while norm-referenced measures depend on a relative standard" (ibid., p. 421). An advantage of having achievement and criteria as a basis of comparison is that it better facilitates the completion of assessment without needing a large number of students or a representative amount of the student group, as required by the norm-referenced principle.

Sadler (1987) further developed the understanding of achievement-based assessment to *standards-based assessment*, wherein a standard defines a particular level of quality that a group of students shall achieve and which is established by the authorities (Tveit, 2008; translation derived from Sadler, 1987, p. 194). In this type of standards-based approach, the assessment criteria are even clearer, specifically so that they describe higher and lower levels of achievement on the one hand, and on the other hand, one or more of these levels are defined as standards that all (in the case of *minimum* standards), as many students as possible (in the case of *norm* standards) or a specific proportion of students (in the case of *exceptional* standards) shall achieve. The concept of standards implicitly includes a responsibility perspective on the part of the teachers, as the school system is obliged to raise students up to pre-defined levels.

For a comprehensive description of key assessment concepts, it is important to include the principle of individual-based assessment. This principle is used a lot when providing feedback to students based on their previous achievements. *Individual-based assessment* therefore corresponds with *adapted education* as a basic value of education and the purpose of Norwegian primary and lower secondary education, and can be used in progress assessments. However, the principle is not compatible with an exam system that has a core value of fair competition for further educational and vocational opportunities.

5 The validity of the current exam system

It is key to the validation process that arguments are developed about the type of evidence that shall be considered as "valid", based on one or more purposes of an exam, and how the results shall be interpreted. This is used in this section to summarise the foundation of knowledge on the validity of the current exam system. We shall closely examine the two purposes of the exam that directly impact the students and that therefore have been identified as the primary focus of section 3: testing the students' individual competency in a subject as it is described in the curriculum and providing a basis for admission to higher education and professional life. Hovdhaugen, Prøitz and Seland (2018) note that being able to safeguard the exam's purposes requires dependency on the high legitimacy of the marking system.

5.1 The relationship between the exam and the curriculum

The exam system shall ensure validity through the cooperation of particularly knowledgeable professionals on papers that are based on national guidelines, with the opportunity for systematic feedback from the examining body. There is generally a lack of systematic research on the cohesion between the curriculum and the exam; however, there is experience-based knowledge and user insight in the field. We have decided to include this in the foundation of knowledge even though it is of varying quality and not systematically documented. In addition, only a small selection of subjects and exam forms have been investigated. Furthermore, this knowledge is largely based on surveys of various sample sizes and response rates. It is difficult to control, or at least know about, possible inequalities, which are often skewed positively. The scope of this section is therefore limited, and it must be acknowledged that there are uncertainties in larger areas that make it difficult to form precise conclusions. It is desirable to investigate the students' and teachers' experiences and viewpoints in a more systematic way, as well as conduct academic analyses of exam papers where these are viewed in the context of curricula and the exam purpose.

As part of the work with further developing the quality of centrally administered written exams, Fafo has evaluated the tenth-grade written mathematics exam for the 2017–2019 period. The marking procedure has also been investigated, and an assessment of the exam's content and formation has been provided. Research has also been conducted on how teachers and examiners assess the cohesion between the curriculum, teaching and the exam in mathematics, and on how the students experienced the exam.

The spring 2017 evaluation shows that the mathematics exam appears to be effective and fair (Andresen et al., 2017). This is a consistent perception among students, teachers and examiners. The majority of teachers and examiners thought the competence requirements and what was being tested in the exam corresponded well. The spring 2018 evaluation confirmed these main findings (Bjørnset et al., 2018). According to IRT analyses, the exam reliability is assessed as being high, which is interpreted as the exam measuring what it was intended to measure – the students' mathematical competence (Bjørnset et al., 2018). However, the teachers who were interviewed stated that the papers with a lot of text hindered students' ability to demonstrate their mathematical competency, particularly the minority language students and students with reading and writing difficulties.

An important validity question is whether the exam tests the same construct over the years, given that the exam papers differ. In connection with the evaluation of the mathematics exam, Udir is therefore conducting a quantitative analysis of the difficulty of exams in cooperation with the Research Unit for

Quantitative Educational Analyses (Enhet for kvantitative utdanningsanalyser, or EKVA) at ILS, through use of annual calibration tests. The calibration test is conducted in April every year and consists of the same papers, with the aim of comparing student performance over three years. Based on this, the researchers can gradually conclude whether it is the students' performance or the difficulty of the exam that may account for any variations in the exam results. The results from the surveys so far show that the student results are on the same skill scale for both 2017 and 2018.

Ensuring that the exam papers and competence aims correspond meaningfully and that the questions allow the students the opportunity to demonstrate their competency at different levels has generally been confirmed in the annual examiner survey conducted by Udir for centrally administered exams (Norwegian Directorate for Education and Training, examiner reports 2018). In addition, IRT analyses for the spring Biology 2 exam in 2017 and 2018 highlighted a good correspondence between the level of difficulty and the students' skills (IRT analysis of the 2017 and 2018 biology exam from the Norwegian Centre for Science Education), which can be interpreted as the exam being in line with the curriculum.

However, it is important to examine the limitations of these surveys. School leaders and school owners were asked about their opinions and perceptions of the exam in the survey *Questions for Norwegian Schools and School Leaders (Spørsmål til Skole-Norge)* in 2017, which included whether they thought that the exam allowed the students the opportunity to demonstrate their competence (Waagene et al., 2018). According to the report from the survey, school leaders and school owners were largely in agreement that oral and written exams allow the students the opportunity to demonstrate their competency (Waagene et al., 2018). However, there was disagreement on whether the exam is suitable for demonstrating competence in *all* subjects or just in some.

There was also disagreement on whether it is clear *which* competence the students are to demonstrate for the exam. Half of the school leaders thought that it was completely clear, while the other half answered that it was somewhat unclear. Among the school leaders in lower secondary education, a somewhat greater proportion answered that it was completely clear compared to leaders in other types of education. By comparison, the school leaders and school owners were in greater agreement that it was completely clear what competence the students are to demonstrate in *classwork assessments*. In this case, approximately three out of four thought that it was completely clear what competence the students are to demonstrate. The smallest schools seemed to be somewhat more positive about both exam forms than the larger schools and had a larger proportion of respondents who answered that oral and written exams allowed the students the opportunity to demonstrate their competence in all subjects.

5.2 A changing understanding of the curriculum

Even though there is a lack of research on the cohesion between the curriculum and the exam, there has been an increase in studies on curricula and assessment in recent years, which have provided insight into classroom practices. It would be natural to assume that there is a certain connection between classroom and exam practice.

Understanding the curriculum, including the definition of competence, is a prerequisite for developing and assessing an exam in accordance with the curricula in question. The FIVIS study highlighted that there may be a weakness in / lack of competence, cooperation, communities of interpretation and / or planning in the school sector when it comes to validity in the ongoing assessment of the classroom (Buland, Engvik, Fjørtoft, Langseth, Sandvik, and Mordal, 2014). Given that the definition of competence is even more complex in the curriculum renewal than it is in the Knowledge Promotion Reform, it may be concluded that the challenges will probably increase.

Sandvik et al. (2012) found that schools have different understandings of the competence thinking in the Knowledge Promotion Reform. Researchers highlight the challenge of using local learning goals that do not reflect or are not connected to the competence aims in the curriculum, and a danger that many narrow, local learning goals assessed through frequent testing may lead to fragmentation and surface learning (Sandvik et al., 2012; Hodgson et al., 2011; 2012). A literature review of research reports shows that there is inconsistency between the competence aims in LK06 and local subject curricula (Andreassen, 2016). Feedback from officials indicates that it is a challenge that a number of teachers and school leaders do not see the various aspects of the curriculum in conjunction with the overall achievement grades (Udir, 2015). Individual officials highlight that teacher have difficulty describing the students' competency in a subject in appeal cases (Udir, 2015).

Simultaneously, as a result of local development processes and national measures (e.g. the Assessment for Learning initiative, known as Vurdering for læring in Norwegian), there has been a lot of attention directed at the assessment field in recent years, and primary and lower secondary education is characterised by a steadily increasing degree of assessment cultures that promote learning (the Norwegian Ministry of Education and Research, 2016). In one of NIFU's bi-annual surveys for Norwegian schools and school owners in spring 2017, 95 per cent of school leaders answered that the work with the Assessment for Learning has increased awareness of the connection between assessment and local work with curricula and has contributed to a more active use of curricula and to the school developing a more learning-oriented assessment culture (Federici et al., 2017). The school leaders also predominantly had the impression that a majority of teachers regarded the competence aims in context, and that marks for classwork were awarded based on a wide selection of sources. A large number of school leaders also thought that the teachers' overall achievement grades are well supported by the curriculum.

The degree to which this development of practice contributes to ensuring the validity of exams according to the curriculum renewal, requires further investigation. NOU 2015: 8 highlights that the challenges connected to understanding the definition of competence will probably increase given the complexity of the definition in the curriculum renewal. The investigation demands various measures for quality assuring the final assessment, not least clarifying the requirements and criteria that make up the basis of the final assessment. The key elements include competence aims in the curricula categorised in stages, preferably with different levels of achievement, and guidance and support material (such as student responses). In accordance with requirements from school owners, schools and teachers, the administration sees an additional need to strengthen the regulations for the overall achievement grades, as the current regulations only specify quality requirements or assessment processes to a small extent, which may lead to differences in assessment results (NOU 2015: 8; see also NOU 2019: 3).

5.3 The exam's various roles in practice

As mentioned above (see section 4.1), validation is a process in which a test's validity is researched in the context of its function. However, empirical studies show that the exam and exam systems in practice may have more functions than the ones formally defined in the applicable regulations (Newton, 2007; Herman and Baker, 2009; Stobart, 2008). The non-defined implicit functions are referred to as "roles" in research. These are not always desirable, but they exist and must be monitored. It is a known problem that when an exam has various purposes and roles, tensions and contradictions may arise between them.

It is therefore important to define the exam's main purposes and to clarify any additional roles the exam has in practice. This should be done to avoid these roles obstructing the main purposes of, or providing a basis for different interpretations of, exam results, and because they can represent a threat to the validity of exams. In Norway, there is a severe lack of research on this question. In this section, we present an analytical framework that differentiates between different purposes and roles of the exam and describes them in detail. The framework may provide a starting point for investigating the degree to which the exam in Norway has implicit roles beyond those explicitly defined.

The exam's purpose to certify learning and rank students

In accordance with international research, Tveit and Olsen (2018) differentiate between various purposes and roles that the exam may have. Firstly, an exam may be used summatively to certify the students' competency and select students for higher education and professional life through subsequent marking and ranking. Both the examination marks and marks awarded for classwork shall provide quantifiable information on the students' competency in a subject at the end of their education in that subject. The marks from the final assessment are extremely significant for the certification of competence and for admission of students to higher education or professional life, or, at tenth grade, for admission to upper secondary education and training. These two purposes are clearly described in the applicable regulations, and we have identified these as the main purposes of the exam (see section 3.1).

The exam's role in quality assuring students' results

The exam may play a role in quality assuring the students' results, as the students receive an external assessment of their own competency in a subject (*White* Paper *No. 28 (2015–2016).* In a final assessment system that is largely based on the subject teacher's assessment, the exam may be considered as an important external quality element. For example, subjects that are assessed by a centrally administered exam have identical sets of questions for everyone, which may contribute to the students receiving a more similar academic record, as the examination marks are awarded according to the same assessment basis (the Norwegian Ministry of Education and Research, 2016, p. 62).[5] The quality assurance role is also featured more implicitly in *White Paper No. 30 (2003–2004)*, where the exam is described as "especially quality-assured tests" (Utdannings- og forskningsdepartementet – now the Norwegian Ministry of Education and Research, 2004, p. 37) because they were developed in line with explicit quality criteria.

The role of the exam in further developing assessment practices

The exam results can contribute to teachers and the school further developing both their own practice and work with assessment. Both *White Paper No. 20 (2013)* and *White Paper No. 28 (2016)* emphasise that an exam's functions include enhancing the competence of examiners. Teachers who act as examiners may participate in many measures that improve their assessment practice, for example, examiner training and meetings with other examiners in order to develop a common understanding (referred to as "communities of interpretation"). This knowledge/experience is brought back to their schools, where it may be passed on to other teachers.

In addition, marks awarded for exams are feedback to the school on how external examiners assess the students' exam performances. This applies to both centrally and locally administered exams. Tveit and Olsen (2018) highlight that upper secondary education and training has few other sources of statistical information, and it is therefore natural that the exam is a source of knowledge for assessing learning outcomes.

However, various studies demonstrate that the relationship between examination marks and marks

[5] In the Swedish system, it is sometimes expected that teachers place significant emphasis on the results from the national tests when marking (Gustafsson and Erickson, 2018).

awarded for classwork is unclear for many teachers and school leaders (Hovdhaugen et al., 2014, 2018; Prøitz and Sport Borgen, 2010). There are various ideas on whether or the degree to which the two marks should harmonise, and whether the exam represents a narrower testing of competency than classwork assessments. Differing understandings of the exam's role may, for example, provide a basis for different interpretations and use of mark statistics in local quality assessment systems, which has consequences for the local development work.

In *White Paper No. 28 (2015–2016)*, however, it is emphasised that a comparison of marks is most suitable when it is used to investigate whether there are systematic deviations from the national average difference between classwork assessments and the exam over time. The paper goes on to say that this should be *one of many sources* of knowledge on practice in school. Hovdhaugen et al. (2018) also highlight that the idea that the exam may act as a calibration tool of marks awarded for classwork has multiple flaws. For example, the two assessment forms may:

- be extremely different and clearly stand out from each other in terms of practice
- have clear differences in their legal definition
- have completely different premises for marking

The exam's role in guiding the learning

The exam may also play a role in guiding the understanding and enactment of curricula. Research shows that exams may have a retroactive ("washback") effect on teaching, since the exam system recognises what is considered to be important in the curriculum, intentionally or unintentionally[6] (see Nordenbo et al., 2009). Simultaneously, it may be argued that this need not be a problem as long as the exam reflects the curriculum.

The 2008 *Education Reflection Report* (Norwegian Directorate for Education and Training, 2009) indicates that the assessment guidelines for marking exams "shall have a learning promoting effect, since the teachers can convey the indicators to the students before the exam" (p. 105). In line with this, previously released exam papers and assessments of these may be examples that schools, school leaders and individual teachers can use as a foundation for interpreting and analysing the definition of competence, curriculum and competence aims in individual subjects. Section 2 on the emergence of the current exam system demonstrated that the exam was originally used in this way to guide the education system, before the current main purpose of certification and selection gradually increased in importance.

The role of the exam in supporting learning and teaching

The exam may have a formative role as teachers use previous exam papers to exemplify/clarify the expected competence at the end of the subject and as a starting point for discussing competency in the subject, progression and indicators of achievement with the students. In this way, the exam may be used to support learning processes and adapt education. Admittedly, this role applies primarily to progress assessments and processes in the classroom and not to the exam as part of the final assessment. However, the exam may be formatively used, as the teacher may use the results to analyse strengths and weaknesses in student responses and look at the analysis in the context of their teaching. They may then use it as a starting point for adjusting their teaching of the subject in the next school year.

[6]The exam lottery should possibly also be discussed in this context because it is to ensure that the students "prepare for exams in the subjects for which the exam is a possible final assessment in addition to the marks awarded for classwork" (the Norwegian Ministry of Education and Research, 2013, pp. 65–66). However, there are no empirical studies that say that the exam lottery actually has this type of guiding role.

Sometimes, it is also highlighted that the exam may have a formative role by being used as external motivation so that the students maintain their focus towards the end of the school year. In this context, the exam lottery should again be discussed, as it may possibly contribute intentionally or unintentionally to securing this effect in multiple subjects. In addition, it may be assumed that the preparatory part of an exam has its own learning effect beyond the teaching time.

In summary, this analytical framework demonstrates that the current exam system may have multiple implicit roles other than those defined in the applicable regulations (see section 3.1 on these) in certifying, selecting, quality assuring, further developing assessment practices, guiding the teaching and, sometimes, supporting teaching in Norwegian lower secondary education. It is probable that the exam has more implicit roles than the ones that are described as its main purpose in the applicable regulations. Multiple and various purposes and roles may lead to various interpretations of exam results and various side effects of any subsequent changes. It is therefore very important to clarify the implicit roles the exam has in practice. However, there is a lack of research investigating this field, and it is therefore difficult to draw definite conclusions.

6 Reliability in the current exam system

It is an important indicator of quality that an exam paper receives consistent assessments from multiple examiners to ensure that the marking is not characterised by coincidence. This requires unambiguous exam papers with clear instructions, explicit assessment criteria (i.e. indicators of achievement) and comprehensive examiner training to ensure communities of interpretation. At the same time, the examiners will always differ in their assessments to a certain degree. However, sufficient reliability is a prerequisite for quality when working with assessments. This section describes the current frameworks for ensuring high reliability and summarises the foundation of knowledge we have in this area.

One challenge with the current data source, in regard to being able to research the reliability of exams, is that examiner information for each subject is gathered only at a student level and not at a question level within a student's exam. This makes it difficult to retrospectively investigate causes of possible problems with examiner consensus. As indicated earlier (section 5), testing an exam before its official use as a means of ensuring high reliability presents an additional challenge, since the exam papers must be kept secret.

6.1 Frameworks for marking exams

Together with the county governors, Udir is responsible for the marking of centrally administered written exams, and the municipality/county is responsible for the marking of locally administered exams (sections 3-28, 3-29, 3-30). Exams are marked by two external examiners. For locally administered exams, one examiner may be the student's teacher for that subject. In cases of disagreement, the final mark will be decided by an exams coordinator for centrally administered written exams and by the external examiner for locally administered exams.

The current regulations contain general guidelines and frameworks for final assessments (cf. 3.2). There is no equivalent framework for *the requirements for* the exam marking process. For example, it is not stated who sets requirements for the quality of the marking. The marking process depends on the type of exam in question, and may take several forms depending on, for example, whether it is a centrally administered written exam or an oral exam. While universal exam papers, assessment criteria and examiner training are developed for centrally administered written exams, oral exams have different papers, assessment criteria and examiner training. Irrespective of the exam form, the marking is based

on a system wherein the examiners "discuss their way to a mark". This demonstrates the importance of developing a community of interpretation among the examiners in order to increase the reliability of exams in the current system.

An effective quality assurance system should include systematic approaches that guarantee examiner consensus and reliability in an effective way, regardless of the exam form. Oral exams are assessed in real time and are not currently testable in the same way as written exams. However, oral exams allow an important opportunity for the students to demonstrate their competence in a different way than in written exams. Therefore, various approaches may be needed to guarantee high exam quality and reliable examination results across the various exam forms.

6.2 Indicators of achievement

There are no guidelines for developing indicators of achievement or explicit assessment criteria related to exams. The students have the right to know what will be emphasised in the assessment of their competence (cf. section 3-1 of the associated regulations). There is a difference between centrally administered and locally administered exams, as well as across schools and school owners, in the extent to which such indicators or assessment criteria exist, and how they are used.

Udir is developing exam guidelines with indicators of achievement for all centrally administered and marked exams. These indicators shall be used in marking and are the starting point for discussion of examiner training and the Examiners' meetings. Udir has also developed reference indicators of achievement in selected tenth-grade subjects to support classwork and progress assessments. Their use is voluntary. Offering a common starting point for assessing competency in a subject may contribute to promoting a more equal and fair assessment throughout Norway. The indicators are based on the curricula and are descriptions of the quality of competence in subjects across the main areas. Competence is described at different levels; currently, the indicators are formulated for the mark groups 2 (E), 3–4 (D–C) and 5–6 (B–A). Teachers at a school are expected to discuss the indicators and work to develop a common understanding of them. These indicators may also be used as a starting point for developing indicators for locally administered exams.

Various surveys show that indicators of achievement are used a lot as a form of exam guidance and are considered useful in the schools' assessment work (Hovedhaugen et al., 2014; Gjerustad et al., 2015). 65 per cent of school leaders also specify using assessed exam responses (Waagene et al., 2018), and half of school leaders use exam reports. In contrast, only a minority of school leaders state that they use predicted grade reports to develop a common assessment basis at their school. The *Questions for Norwegian Schools and School Leaders (Spørsmål til Skole-Norge)* survey from autumn 2014 shows that virtually all school leaders and owners specify having prepared and used local indicators for assessment work (Gjerustad, Waagene and Salvanes, 2015). There is no systematic information on the content and quality of these indicators. It must be noted that either the sample size or response percentage was limited for all studies. If there is a particular danger of the responses having inequalities – which is often the case in surveys – these are often skewed positively.

Udir's indicators of achievement leave some room for interpretation, which must be developed and discussed in collaboration with other teachers. Hovdhaugen et al. (2014) found that the teachers think it is easiest to assess exam responses at either end of the scale, but that it takes more work to justify why a 3 (D) over a 4 (C) was awarded than, for example, a 5 (B) or even a 6 (A). Because the marks 3 and 4 comprise a particularly large proportion of the marks, the teachers have requested clearer assessment

criteria for various types of papers, as well as examples of responses, which will make it easier to differentiate between a 3 and a 4 (Krogh, 2016).

In summary, it may be stated that there is a lack of systematic research, with the exception of surveys on how school leaders and teachers work with indicators of achievement. The foundation of knowledge gives reason to assume that they have different experiences of indicators and assessment criteria, which may lead to differences in the assessment process.

6.3 The significance of communities of interpretation

There are many counties and municipalities that have developed guidelines for oral exams, but these involve various scopes and say differing things about subject-specific conditions. There is a lack of systematic knowledge about how municipalities and counties work qualitatively with examination results from locally administered exams. The counties have an established collaboration for work on locally administered exams and have various collaborative arenas and areas, such as developing common exam papers for locally administered written exams for individual subjects.

In NIFU's spring 2017 survey for Norwegian schools and school owners, a vast majority of municipalities (68%) and counties (87%) stated that they facilitate arenas for learning and sharing, wherein teachers may further develop their assessment practices (e.g. networks / scheduled gatherings / meeting places) (Federici et al., 2017). Fewer school owners, 47 per cent of municipalities and 53 per cent of counties respectively, stated that they facilitate discussions on the curricula's content.

Hovdhaugen et al. (2014) also found that there are various forms of subject collaboration between teachers, and that the school leadership had in some places implemented specific measures for developing and forming subject collaboration between the teachers, and in other places had left this to the departments. A large majority of school leaders in NIFU's bi-annual surveys express that the way the teachers' assessment practices can help the students learn and achieve their goals is discussed to a large degree at their school (Federici et al., 2017). Furthermore, the school leaders largely have the impression that all or the vast majority of teachers in the same subject/discipline work together to achieve a common understanding of the competency level in their subject. There is strong support for the perception of both school owners and school leaders as being driving forces for the development of assessment practices, but school leaders in particular are viewed in this way.

The various surveys show that there is a wide range of forms of collaboration related to local assessment, but do not mention anything about the quality of these collaborative areas and the degree to which or how this work is connected to locally administered exams. However, we may assume that collaboration on assessments will also indirectly impact locally administered exams. Simultaneously, geographical location in the municipality or county is a factor that may impact the possibility of collaboration between schools and participation in any examiner training connected to locally administered exams.

Together with the county governors, Udir arranges collective marking for all centrally administered written exams (with the exception of centrally administered exams with local marking) and has various measures that shall collectively contribute to communities of interpretation (see the text box below). Examiner training is a part of the collective marking and is currently not obligatory. However, all examiners are encouraged to participate, and there is generally a large turnout at these meetings. The

schools are also not obliged to have teachers participate in centrally administered exam marking, which may mean that there are schools (through the years) without teachers who have marked centrally administered exams. These schools will therefore not have teachers who can bring back their experience from the training and the Examiners' meetings.

Teachers who have participated in examiner training have found it extremely useful (Hovdhaugen et al., 2014), which corresponds with Udir's experience. Teachers with examiner experience express trust in the communities of interpretation that emerge in the work with collective marking (Hovdhaugen et al., 2014). According to researchers, the Examiners' meetings may be a type of neutral ground for the teachers, where the subject itself is the focus, the responses are anonymous, and only the indicators/assessment criteria are used. The Examiners' meetings strengthen the teachers' perceptions of safety and verifiability when marking exams.

Udir has multiple measures that collectively contribute to a community of interpretation when marking a centrally administered exam (the only exception being centrally administered exams with local marking):

- **Predicted grade reports** in primary and lower secondary schools and for Norwegian classes in upper secondary level 3 for all exams coordinators: These provide guidance for examiner training led by exams coordinators.
- **Examiner training and collective marking**: In examiner training, a selection of genuine exam responses is used to discuss the competency that is demonstrated within them and their respective marks. The community of interpretation that results from the examiner training sets guidelines for the marking of all exam responses in the subject.
- **Exam guidance including indicators of achievement**: This provides information on an exam and how it shall be assessed. Indicators of achievement also contribute to ensuring an overall assessment of competency. The examiners shall use this guidance as a common reference framework in their work. The guidance shall be provided in good time before the exam.
- **Exam responses with explanations for various marks**: These are published on Udir.no for various subjects in primary and lower secondary schools and upper secondary schools. For each response, an explanation of the mark is provided. These are to be used as a reference when marking and may be a starting point for the development of a local community of interpretation.
- **Exam reports** in a selected subject: These have the purpose of providing teachers and candidates with better insight into how the exam papers are rooted in the curriculum and experience of the exam implementation and collective marking. The reports also include mark statistics.

Moreover, the researchers indicate multiple positive effects of examiner training/meetings (regardless of who arranges them):

- Examiner training is an important assessment competency enhancement and for some is potentially their only "training" in marking.
- Many teachers, school leaders and school owners highlight that collaboration on examination results has led to a comprehensive way of thinking about assessments and given them tools for conducting assessments in a community of interpretation.
- The teachers also find this to be valuable in their own assessment practice and that it can help strengthen assessment communities at their school.

- Many teachers, school leaders and school owners support obligatory training, as it may benefit the entire subject community at school and contribute to a more uniform assessment within the whole school.
- Marking can create meeting places and subject communities, which many teachers say strengthens professionalism in the teaching profession generally and particularly for assessments.

Generally, the school leaders find that the examiners' experiences contribute to improving assessment competence throughout the school for oral and written exams. Eight out of ten school leaders think they contribute to written exams to some or a large degree, and nine out of ten think they contribute to oral exams (Waagene et al., 2018).

6.4 Examiner consensus

High consistency between the examiners when assessing the exam is crucial for quality. While it is not realistic to expect that the examiners always consider exam answers equally, the aim should be to avoid greater variation in the assessment as a whole. It is important to note that standardised exam papers do not necessarily have higher examiner consensus than non-standardised papers. Examiner consensus is often related to whether it was possible to develop clear exam papers, instructions and assessment criteria beforehand. Reliability is also related to the extensiveness of the examiner training. Furthermore, even though there is no empirical research on this topic, it is not impossible that the number of exam papers an examiner marks impacts both the quality of the marking and the examiner consensus, as this may affect how much time the examiners actually have to discuss and assess written responses.

Examiners' meetings and examiner training have been added as part of the quality assurance for centrally administered exams. The examiners conduct a preliminary assessment of the papers before the Examiners' meetings and make a final assessment based on the community of interpretation.

Fafo's evaluation of the mathematics exam for tenth grade students in 2017 showed that there was good consensus between the examiners in their mark suggestions before the collective Examiners' meeting, even though some examiners called for better guidance in marking individual question responses. This particularly applied to receiving clearer guidelines for marking papers that require the use of digital aids (Andresen et al., 2017). The researchers conclude that there was high examiner consensus for spring 2018 – in other words, the examiners were fairly similar in their assessments (Bjørnset et al., 2018).

Professionalisation of the assessment:

External marking entails the teachers having to discuss the curriculum, assessment and marking with other teachers both before and after the exam. For centrally administered exams, the examiners are recruited from throughout Norway, and examiner training is conducted in order to professionalise the assessment of the responses and contribute to communities of interpretation and impartial marking. Oral, combined oral and practical, and practical exams are exam forms that entail the wide involvement of teachers locally. Therefore, teachers are able to get an outside perspective of their own practice, through being examiners both for their own students and at other schools. This may be a starting point for discussing, adjusting and further developing their own training and assessment practice.

Even though the project happened a little while ago, the KAL project (Quality assurance of learning outcomes in written Norwegian) must be mentioned when discussing exam evaluation in Norway(Berge mfl., 2005). The project comprises a study of 3,330 exam texts from 1998 to 2001 and remains the most comprehensive study of students' writing in the Norwegian subject exam and its subsequent marking to date. Among its findings, the KAL project showed that the students were fairly capable writers, and even the low-achieving students were able to produce simple narrative texts. Other discoveries included gender differences in favour of the girls, and that the students prefer to write subjective narrative texts over factual ones. This tendency was directly challenged through the Knowledge Promotion Reform, where non-fiction and fiction texts were given equal weight, and through exam papers that have made it obligatory for students to demonstrate that they can write factual texts.

A particularly important finding in this context is related to the examiners' assessment of the exam papers. The KAL researchers concluded that the consensus between examiners in primary and lower secondary schools was not as high as was desirable, but better than many had anticipated. They highlighted that primary and lower secondary schools develop a conversation culture about students' performances and the quality of their texts. Furthermore, the KAL report provided clear guidance that teachers' conversation culture is a strategic master key to further developing the quality of literacy programmes in primary and lower secondary schools (Berge et al., 2005).

Based on a survey conducted by NIFU, it seems that the need for clearer assessment criteria and a larger community of interpretation is particularly great in upper secondary schools (Seland, Lødding and Prøitz, 2015). A methodical survey conducted by EKVA on behalf of Udir indicates that examiner consensus for exams is a particular challenge in Norwegian language subjects. The surveys showed that agreement on the assessment of answers between examiner 1 and examiner 2 before the establishment of a community of interpretation during the collective marking was not particularly high. The fact that examiner consensus in subjects such as Norwegian was lower than in subjects such as mathematics may be rooted in the wide variety of question formats that the candidates are examined in for these subjects, and the degree to which their responses provide room for and need professional interpretation. A master's thesis on examiner reliability in Norwegian language subjects for the school year 2015 supports this interpretation and additionally highlights ambiguity relating to the weighting of assessment criteria and the weighting of short- and long-answer questions (Krogh, 2016).

In his doctoral thesis, Bøhn (2017) particularly focused on how the assessment of oral exams in the core subject English in upper secondary schools works. His conclusion was that there is an acceptable level of examiner correspondence in the use of overarching criteria. The reliability of the examiners in the study, which included 80 informants, was largely good. However, Bøhn also indicated that there were challenges in assessing the English exam, which are related to the assessment of the expression, content and level determination of individual criteria. In this respect, it may be useful to develop clearer indicators of achievement. An additional finding was that a common understanding of assessment criteria does not automatically mean that the teachers assess performance equally. It is also important that teachers are in agreement on how the level of a performance shall be determined on the scale of marks. This report indicated that the English teachers for students in vocational education had a tendency to assess the students "more kindly" than teachers for the Education Programme for Specialisation in General Studies.

Carlsen (2003) has also investigated examiner-based assessment of oral language skills in the case of Norwegian as a second language. Her findings confirmed Bøhn's. She concluded that the examiners

should agree with each other on the marking and emphasise the same features in their assessments. Otherwise, the assessment was in danger of being characterised by coincidences and was therefore not to be trusted.

Simultaneously, research shows us how it is possible to achieve a greater community of interpretation and higher reliability – even when it comes to constructs that have had as poorly defined a basis as that for writing. A lot of the research that exists in Norway on assessing writing has been carried out in the context of the national writing projects and the Developing National Standards for the Assessment of Writing – a Tool for Teaching and Learning Project (Normprosjektet). Experiences from this context and the main findings may be applicable to exams.

Kvistad and Smemo (2015) discovered that the students' texts and subsequent assessment benefited most from explicit expectations, particularly related to the exam's purpose (Otnes, 2015), as well as detailed requirements concerning content and structure (Smemo and Solem, 2015). Vaguely formulated papers were not only decisive for the students' performance, but also difficult to assess (Solheim and Matre, 2014). The authors identified that use of example texts was well-suited for developing communities of interpretation among the examiners, as they demonstrate the various assessment standards. When it comes to the number of examiners, Borgström and Ledin (2014) conducted a study in Sweden and concluded that a textual assessment requires three examiners to guarantee satisfactory reliability.

In order to achieve common expectations and assessment criteria ("standards"), the Normprosjekt utilised a *bottom-up* process, which involved a larger number of experienced teachers (Solheim and Matre, 2014; Evensen et al., 2016). Through investigating how these teachers assessed student texts, a multi-dimensional matrix was able to be developed that clearly specified the assessment standards. Staging an intervention, where other teachers received information on writing and assessment, further contributed to significantly developing the teachers' assessment competency.

7 The relationship between the exam and classwork assessments

Exam results that are published in Udir's school portal (Skoleporten) and statistical portal (Statistikkportalen) provide Udir, county governors, school owners and the schools a certain foundation of knowledge on the mark distribution and average for the exam. Marks and mark suggestions from the results of centrally administered written exams that are registered in PAS (Udir's service for conducting centrally administered written exams) are a source of information that can be used in the work on further developing exams.

Examination marks and marks awarded for classwork for a subject should be expressions of the same competency, but there are often questions on the difference between the two. Differences in marks are not necessarily problematic in themselves. However, there is possibly a question of how significant the differences can be before the marks are no longer an expression of the same competency in the curriculum. Simultaneously, there is a consensus that exams and classwork assessments cover different perspectives.

Therefore, differences do not necessarily indicate a student's over- or underachievement and are not enough in themselves to cast doubt on or legitimise either examinations marks or marks awarded for classwork. However, differences in these marks should not be systematically related to the class year, subject, class, school, geographic location or any other type of group. If the differences can be systematically connected to external conditions other than the students' competency, they then represent inequalities that are not compatible with fair assessments. [7] This is investigated below.

Research clearly documents that there are differences between the exam and classwork assessments (see, for example, Hovdhaugen, Prøitz and Seland, 2018), and that these have existed for a long time (Hægeland et al., 2005). Here are some of the findings from research on the relationship between examination marks and marks awarded for classwork:

- Nationally, the average marks for centrally administered exams are normally somewhat below the average marks awarded for classwork. However, we have little knowledge of the causes of these differences. For example, one study shows that the difference between classwork assessments and exams depends on the exam form and subject, and that there are greater differences in Norwegian language assessments than in mathematics.[1] The biggest difference in mathematics between marks awarded for classwork and examination marks is in practical mathematics, where there is an entire mark level's difference between the average mark awarded for classwork and the average examination mark. 78 per cent of the students receive lower marks for the exam than for classwork in practical mathematics for upper secondary level 1. (The Norwegian Directorate for Education and Training, 2017)

- The marks awarded for classwork in subjects with written exams are relatively constant over time, while the examination marks are more varied (Hovdhaugen et al., 2014). Therefore, the examination marks are the least stable of the two. Based on this finding, researchers question the exam's function as an objective measuring point. There are questions on whether changes to examination marks represent changes at a competency level or changes to an exam's difficulty.

[7] It is important to remember that the rankings may only be "roughly" the same, as there is always random variation, especially in an exam that is taken only once and preferably by a small selection of students at a school. Random circumstances such as exam anxiety, illness, bad moods, misfortune with the specific questions being asked on that particular day, etc. will lead to an imperfect single measurement. However, this will not lead to systematic differences.

- Students at small schools (fewer than 50 students over 7 school years) and schools with low examination mark averages receive better marks awarded for classwork than those at larger schools (around 40 students or more on average per year) and schools with high examination mark averages. Between 40 and 50 per cent of the schools stand out as having particularly high or low marks awarded for classwork compared to examination marks. The marking practice at each school is largely stable across subjects. In other words, if a school awards high marks for classwork in one subject, they will also award high marks for classwork in other subjects. The marking practice also remains stable over the years (Galloway, Kirkebøen and Rønning, 2011). It seems that teachers implicitly adhere to an internal social norm at their school when awarding marks for classwork assessments. This means that they orientate towards the general level at their school. Marks awarded for classwork by a teacher at a school with high-achieving students will probably therefore evaluate the same final competence as being at a slightly lower level than a teacher at a school with a high proportion of low-achieving students. As centrally administered exams have the same papers and marking practice throughout Norway, this often results in marks that are a little higher for exams than for classwork assessments at high-achieving schools than at low-achieving schools.

- There are also systematic differences in the discrepancy between exam and marks awarded for classwork when comparing high- and low-achieving students. About half of the students receive a different examination mark for an exam than for their classwork assessment in a subject. While 75 per cent of the students who receive a 5 (B) or 6 (A) for their classwork assessment go down a mark for their exam, fewer than 50 per cent who receive a 2 (E), 3 (D) or 4 (C) receive a lower examination mark (the Norwegian Directorate for Education and Training, 2013). This may primarily be an effect driven by natural fluctuations, as students who receive a 5 (B) or 6 (A) may go down in mark but rarely go up, while those who have marks in the middle of the scale may go either up and down in mark.

- Girls perform – relatively – better in classwork assessments than exams (which can be seen in the Education Reflection Report over many years), and this particularly applies to the Norwegian and foreign language subjects (wherein the girls' grade point average is 0.4 to 0.5 higher than the boys' for the exam) (Wollscheid et al., 2018; Borgonovi, Ferrara and Maghnouj, 2018). The Stoltenberg Committee has investigated gender differences in school performance and is of the opinion that the assessment system seems to be disadvantageous for boys if it features more marks awarded for classwork than examination marks (NOU 2019: 3).

- Another potentially systematic difference is found between state and private schools, as the difference between exam and marks awarded for classwork is greater in private than in state upper secondary schools (Hovdhaugen, Seland, Lødding, Prøitz and Vibe, 2014). This is probably as students at private schools receive higher marks awarded for classwork (in the cases where that the same academic skill is demonstrated in the exam) (the Norwegian Directorate for Education and Training, 2013).

- The *School Results 2008* (*Skoleresultater 2008*) report, drawn up by Statistics Norway (Statistisk sentralbyrå) on behalf of Udir, presents a survey of results in primary and lower secondary schools and upper secondary schools. It shows a strong connection between the students' marks in primary and lower secondary schools and upper secondary schools (Steffensen and Ziade, 2009). Subject marks from primary and lower secondary schools

are generally a good indicator of marks in the same subjects in upper secondary education and training, even when controlling for differences in family background. The report also contains an analysis of the failure rate across selected subjects and student groups in upper secondary education and training. There is a lower failure rate in the Norwegian and English language subjects than in mathematics, and the proportion who fail mathematics is clearly lower in theoretical than practical mathematics. Furthermore, the proportion who fail is lower in university-preparatory subjects than in vocational education programmes.

The differences between exams and classwork assessments documented here (between boys and girls, state and public schools, large and small schools, and high-achieving and low-achieving schools or pupils) show that there are *systematic* deviations between examination marks and marks awarded for classwork. This again demonstrates that the differences cannot be reasonably connected to the students' final competency in that subject. Multiple studies have documented that the teachers potentially emphasise other factors than the curriculum achievements when awarding marks for classwork assessment in subjects, such as the student's efforts or organisation and behaviour (Dale and Wærness, 2006; Prøitz and Spord-Borgen, 2010; Sjyvollen, 2007; Tveit, 2007b).

In addition, it seems that the teachers potentially adhere to norm-referenced assessments wherein they compare students in a class or the school with each other instead of exclusively marking on the basis of individual achievement (Galloway, Kirkebøen and Rønning, 2011; see section 4.4 for definitions of the concept). These types of differences create a danger of different pupils not having the same opportunities in assessments. That being said, it is certainly important to highlight that an achievement-based assessment will presuppose clear goals and operationalisation, which probably do not currently exist sufficiently.

In addition to the group-related differences documented above, there are differences between exams and classwork assessments related to school year. It means that variation at the marking level over the years may turn out to be a source of unfair competition for the same places on a programme of study.

There are also variations in marking an exam compared to a classwork assessment across subjects. Students who specialise in science subjects have a higher average mark in the core subjects than students who, for example, take social sciences. However, the students studying science subjects receive lower marks on their academic record in general studies subjects than the other student groups (Angell, Lie and Rohtgi, 2011). This means that the mark requirements seem to vary across the general studies subjects, and that, for example, a 5 (B) does not have the same meaning in science subjects as it does in social sciences. A similar trend can be observed when comparing foreign languages with social sciences. Hovdhaugen (2014) indicates that this type of subject-specific difference is potentially due to different assessment approaches. In addition, the differences may be explained by aspects of the subjects' epistemology, particularly the subject-specific structures (for more details, see section 8).

Another phenomenon that may have consequences for admission to higher education and professional life **is** inequalities due to the subjects' value on the student's academic record. Subjects with a low number of teaching hours have the same weight for admission as subjects with a high number of teaching hours. The number of marks per subject does not necessarily match the number of class hours in upper secondary education and training. For example, there may be up to six marks on an academic record for Norwegian in upper secondary level 3. As there are significant gender differences in the language subjects, the number of marks for languages will benefit the girls, according to the Stoltenberg Committee (*NOU 2019: 3*). Therefore, the administration recommends investigating the weighting of marks according to teaching hours or other models.

In summary, the foundation of knowledge shows that the relationship between examination marks and marks awarded for classwork is characterised by systematic differences that are related to external circumstances and potentially not to the students' competency – which undermines the fairness of the assessments. An important question is the ways in which it is possible to counteract or compensate for these inequalities. So far, our review shows that there is little research on this, either in Norway or internationally.

8 Subject assessment – subject differences

Section 7 documents that assessment and assessment results vary between school subjects. One explanation for this is the variation between the subjects' various and particular characteristics, which are often referred to in a Norwegian context as the subjects' unique nature. Another explanation is that teachers and examiners in assessment situations draw on specific and different epistemological and ideological concepts about assessment in different subjects. Testing competence in line with the curriculum renewal and based on the new curricula will involve having to acknowledge the subjects' individual content and structure. Therefore, this section summarises the perceptions that researchers and teachers have of the different subjects, and how the implicit or explicit ideologies impact attitudes towards assessment.

8.1 Subject perceptions

Muller (2009) differentiates between subjects with different conceptual and contextual contexts. Subjects that have a strong conceptual context have more explicit disciplinary roots in higher education (the research discipline that the subject refers to) and have a more rigid, hierarchical and sequential structure that gives teachers a clearer framework for assessment. On the other hand, there are subjects with stronger conceptual contexts that have a weaker connection to the subject's reference discipline, are less hierarchical and more segmented, and that require constant development of common frameworks for the subject's areas of competence and what is deemed important knowledge in the subject, which should thus be considered. Therefore, school subjects comprise the basis of teachers' and examiners' constructions of frameworks for assessing performance and associated practices for marking (Wiliam, 1996).

Based on comparisons of teachers' statements on assessing in English, natural sciences and mathematics, Black argues that teachers of mathematics and science consider their subjects to have unique and objectively defined goals, while teachers of English (in an English-speaking context) consider there to be a number of goals that may be relevant for students to achieve at a certain point (Black et al., 2003, p. 68). We also find this in Norwegian studies, as evidenced in repeated rounds of interviews with over 100 Norwegian teachers in lower secondary schools and upper secondary education and training on assessment in their subjects conducted from 2009 to today (Prøitz and Borgen, 2010; Prøitz, 2013; Hovdhaugen et al., 2014; Seeland et al., 2018; Prøitz, 2018).

8.2 Subject assessment perceptions

In national and international studies of teacher-reported considerations, we also find that understandings of a subject's unique nature impact the teachers' assessment practices. For example, assessment in subjects such as English is often characterised as comprehensive, intuitive, non-numerical and based on observation and dialogue, while assessment in subjects such as mathematics is categorised as rational-analytical with fixed standards and criteria, and with value-free and stable indicators (Wyatt-Smith and Klenowski, 2013). We can see descriptions of assessment in subjects with a narrower or wider basis, where the narrow approach is dominated by use of a single assessment form, often written, or a very short test situation. The wide approach includes a wider selection of assessments and particularly a combination of written, oral and/or practical testing for exams that

facilitates a wider test of competence. In a Norwegian context, there is a lot of evidence that indicates that the exam form for the individual subject contributes to defining these narrow or wide frameworks for assessment (Prøitz, 2018).

We know from research that teachers and examiners are largely loyal to their framework and guidelines on assessment, but research also shows that following new systems and rules within the subject's framework may lead to problems if the policy behind new assessment systems does not harmonise equally well with the subject's pre-existing framework (Prøitz, 2014). For example, we know that some school subjects seem to fit better with the current competence-based thinking than other subjects (Muller, 2009; Prøitz, 2014).

Previous studies in Norway have shown that there may be a weaker connection between subjects, content and national frameworks for assessment, particularly in more contextually anchored school subjects such as Norwegian and arts and crafts (Prøitz and Borgen, 2009; Prøitz, 2013). This may reflect a weakness connected to assessment in Norway in connection with the development or revision of subject content in the national curriculum, where assessment often enters into the discussions too late or is "tacked on the end" and is therefore not an integrated part of the work with the curriculum documents (Lysne 2006; Gjone, 1983). This often leads to extensive postproduction and adaptations in order to ensure high quality and effective assessments.

The subjects' unique nature has to a very limited extent been a focus of assessment research. National and international assessment research has largely aimed to contribute to increased knowledge of, and to define good practice for, assessments on a more general and universal basis, despite the research most often occurring within the frameworks of school subjects. Therefore, it does not focus on whether more studies on how universal principles for good assessment can be developed or supported (Brookhart, 2013; Wyatt-Smith and Klenowski, 2013), but whether the school subjects' content and structure have been sufficiently recognised as central factors within assessment research.

The curriculum renewal tries to meet this challenge by defining key elements that shall cover the most important aspects of the subjects and provide an explicit prioritisation of what the students shall learn (https://www.udir.no/laring-og-trivsel/lareplanverket/fagfornyelsen/kjerneelementer/ in Norwegian). Areas of competence, methods, concepts, ways of thinking and styles of expression that have been identified as key elements shall characterise the curricula's content and progression and contribute to the students' developing an understanding of the content and contexts of the subject over time. In this way, the key elements can contribute to the subjects' content and structure being recognised, but whether that happens is an empirical question and should be thoroughly investigated (including any unintended side effects).

9 Students' experience of exams

To form a comprehensive picture of how exams work, it is crucial that we listen to what the students say about their experiences of the exam system. Through the School Student Union of Norway, students have been highlighting weaknesses of the exam system since 1963. These weaknesses include the students not getting to demonstrate their full competence, the current form impacting the students' performance to a significant degree, and the exams drawn from the lottery being largely random.

Exams are a part of a complex and psychological reality. They can be exciting and demanding while simultaneously something that many students dread. Exams can lead to anxiety, but they can also equip students for further professional life and studies. Harris and Brown (2016) indicate that social and psychological factors impact a number of aspects in a school system: decisions on how teaching shall be judged and taught, students' participation in assessment practices, and how assessment results are interpreted, understood and utilised. For example, students may worry about receiving poor marks, and teachers may be impacted by time constraints, mood swings, prejudices and similar when they are marking exam responses. The social, historical and cultural frameworks in and surrounding the education system impact students' views of exams, motivation, self-image and self-esteem, as well as opportunities for collaboration. Similarly, the political and legal frameworks for exams may match or conflict with teachers' assumptions, values, attitudes, etc. Harris and Brown (2016) indicate that the human conditions for assessment should therefore form the basis of how we understand the design, implementation and marking of exams and other assessment situations.

Section 3-32 of the regulations associated with the Norwegian Education Act allow for local facilitation for students who need it, so that they are able to demonstrate their competency in the subject. The facilitation should not lead to these students gaining an advantage over the others.

There is little research available on the student perspective in the final assessment system generally, and this also applies to the students' own experience of the exam. This section summarises some findings that have been collected from international research, as well as some feedback that Udir has received through surveys.

9.1 The student voice, motivation, exam anxiety, stress and performance

One way to gain greater insight into how the students think about, for example, types of exams, time constraints and time-related stress is to invite the student voice in to work with exams and marking. As a trial, Udir has, in consultation with the School Student Union of Norway, included the student voice by conducting surveys among students who completed exams in English after the tenth grade and in English as a core subject for upper secondary levels 1 and 2 in 2016 and 2017. In spring 2017, the students also participated in a predicted grade report in English for the first time and explained what they thought about the exercises.

The student/user perspective is dealt with in the exam coordinators' reports and in the exam reports from the tenth-grade English exam in spring 2017. Among other questions, they were asked about their reasoning for choice of exercises, the purpose of the preparatory period and the exam length (the Norwegian Directorate for Education and Training, 2018d). Provided below are only 2 of the 500 student responses who gave their opinions on the preparatory period for the tenth-grade English exam in spring 2017. The vast majority were positive about "engaging with the subject and topic" before they embarked on the exam itself:

"Honestly, you don't need the preparation booklet, there's almost no point..."

"Great, as then you get into working mode and get to think about the subject for a day before you have such a big assessment..."

The connection between motivation and performance is important to consider in the development of the exam field (Eccles, 1983). Both extremely high and extremely low levels of motivation variables may be less desirable than anything in between. For example, if students find that the significance of the exam paper is low, they may choose not to spend energy on making an effort to master it (Natriello and Dornbusch, 1984). However, if students find that the performance demands of an exam are extremely high, anxiety may impede performance (Tobias, 1985). In the same way, if the students have an extremely low level of expectation of mastery for an exam, it is very unlikely that they will approach it with a lot of enthusiasm or endurance. If they have high expectations of mastery, they may risk not giving the task sufficient attention to achieve a good result (Schunk, 1984).

When it comes to students' exam anxiety, studies (Hill, 1984) report that students' nervousness is desirable when they perceive an exam to be significant, when it is expected to be difficult and when the circumstances surrounding the exam situation are intrusive (e.g. rigid time frames, associated time constraints, special test instructions and unknown exam form). Even though students' mistakes on previous papers impact the development of anxiety, nervousness is not caused solely by a lack of knowledge or skills required to answer the questions. Studies have shown that students with high exam anxiety do better and perform at levels closer to their less anxious peers on the same cognitive exercises when the exams are administered under less stressful conditions (Hill, 1984; Hill and Wigfield, 1994).

9.2 Students' experience of different exam forms

There is not that much research at all on how students in Norway experience the different exam forms. We have therefore chosen to look for international research on assessment in *high stakes* contexts and exams in order to get an indication of possible challenges that Norwegian pupils could struggle with. But it must be noted that we cannot know with certainty whether the international research status also applies to Norway. What certainly can be said is that the exam is a test with great consequences, so it may be assumed to be possible that these types of challenges exist.

Collectively, international research shows that students prefer assessment formats that reduce stress and anxiety (Nassar, Qaraeen and Naba'h, 2011; van de Watering, Gijbels, Dochy and van der Rijt, 2008). A study by Birenbaum and Feldman (1998) found that students with little/no anxiety about exams prefer open exams. However, students with high levels of anxiety largely prefer multiple choice papers because they associate them with more security in assessment situations. This finding is consistent with findings from a study by Nassar et al. (2011), wherein the students thought that multiple choice papers in the exam are less difficult, more explicit and fairer than the long-answer questions. However, the students thought that both types of exams are valuable. Birenbaum and Feldman therefore assume that if the students get their preferred assessment form, they will be more motivated to do their best.

An earlier study by Ben-Chaim and Zoller (1997) found that students in scientific subjects in upper secondary schools prefer exams that are written, with an unlimited time frame and wherein they can use supplementary material. Time limits are found to be stressful and a cause of worry and pressure. Assessment forms that reduce stress will, according to Ben-Chaim and Zoller (1997), increase the chance of success, and students prefer exams that emphasise understanding instead of surface learning. Baeten et al. (2008) found that the preferences for various forms of exam seem to be related to various learning strategies and approaches; students with an in-depth approach seem to prefer long-answer questions, while students with a surface approach prefer multiple choice questions for exams (Baeten, Struyven and Dochy, 2008; Birenbaum and Feldman, 1998).

A meta study by Beller and Gafni (2000) found that there are gender differences in the preferences of test forms and exams. In these cases, girls prefer long-answer questions and boys show a slight preference for multiple choice questions (e.g. Gellman and Berkowitz, 1993). Furthermore, Beller and Gnafni (2000) found that boys score better on multiple choice questions than girls, and that girls score better than boys on open questions than on multiple choice questions (e.g. Ben-Shakhar and Sinai, 1991). However, there are also individual studies that indicate the opposite in terms of gender differences in exam forms. The evidence is therefore a little unclear.

Nassar et al.'s (2011) study found a divide between low- and high-achieving students when it comes to preference for long-answer questions as a test format in the exam. They also found that high-achieving students prefer long-answer questions for an exam more than moderate- or low-achieving students.

McDowell (1995) suggests that students think new assessment forms in school are interesting and motivating. The students are still aware of the need to achieve high marks, but the degree to which they are in favour of this varies. Alternative assessment forms (exams) may contribute to transforming an exam culture controlled by a traditional exam form to an assessment culture that emphasises the cohesion between teaching and assessment (Birenbaum and Dochy, 1996; Dochy and McDowell, 1997). Research shows that alternative assessment methods (e.g. portfolio assessments, group projects, use of cases) are less threatening for the majority of students than traditional testing. These alternatives are also perceived as fair test formats (Sambell, McDowell and Brown, 1997).

Dochy and McDowell (1997, p. 292) indicate that a change to assessment forms is an effective way to encourage the students to change their learning methods. Furthermore, it is highlighted that assessments are one of the most effective tools for innovation in both teaching and learning. "When assessments stay the same, students often will not accept the need to change their approaches to learning; for example, students often prepare for exams by rote learning even if this is not appropriate" (Dochy and McDowell, 1997, p. 292). However, the researchers caution against a belief that new assessment formats are automatically better, as they think that there are no ideal individual assessment forms. A single assessment form cannot serve multiple purposes, and each assessment form has its own method variation that interacts with people.

In summary, this foundation of knowledge shows large research holes in terms of the students' experience of exams and exam forms in Norway. The few surveys that exist indicate that there is a lot to gain by listening to the students, as the results can contribute to increased validity. The students have over time named weaknesses in the exam system as a whole and in individual exam forms, such as ambiguous questions or instructions. Based on international research, it seems to be important to vary exam forms as much as possible so that different student groups have the opportunity to perform in the best possible way.

Part 3

Predicting the curriculum renewal

10 The curriculum renewal's expanded definition of competence and the exam

The competency-based curricula were introduced with the Knowledge Promotion Reform. Evaluations of the Knowledge Promotion Reform indicate local differences when it came to curriculum understanding in the schools (cf. point 6.3). However, later reports (Norwegian Directorate for Education and Training, 2018e) show that the teachers have still gained an increased awareness and understanding of the definition of competence and the curricula. As stated in the introduction, the curriculum renewal's revised definition of competence, which forms the basis of the curricula's design, emphasises the students' use of knowledge and skills in both familiar and unfamiliar situations, and the ability to understand, reflect and think critically is an important part of their competency.

The definition of competence in the curriculum renewal LK20:
Competency is the ability to acquire and apply knowledge and skills to master challenges and solve tasks in *familiar and unfamiliar contexts* and situations. Competence includes *understanding and the ability to reflect and think critically*.

The guidelines for designing subject curricula in the curriculum renewal (LK20 and LK20S; Norwegian Directorate for Education and Training, edited 11 October 2018) contain directions for ensuring that they describe the relevant competence, clear priorities and clear progression, and a seamless cohesion within and between subjects. These curricula are to be effective tools for supporting and guiding teachers, school leaders and school owners. Furthermore, the guidelines state that the curricula shall facilitate various forms of teaching and assessment that promote in-depth learning. In-depth learning in the curriculum renewal is defined as "gradually developing knowledge and lasting understanding of concepts, methods and contexts in subjects and across disciplines. This involves the students reflecting on their own learning and applying this learning in different ways, in both familiar and unfamiliar situations, alone or together with others" (Norwegian Directorate for Education and Training, 2018f).

In-depth learning and the definition of competence contain elements that overlap with and correspond to each other. Both concepts highlight understanding and the use of knowledge and skills in familiar and unfamiliar situations. In addition, they both emphasise teaching a student to learn and reflect on their own learning. Therefore, in-depth learning can be considered a prerequisite for developing complex competency as expressed in the curriculum renewal.

Creating open and overarching goals that enable the students to transfer what they have learnt to both familiar and unfamiliar situations on the one hand, and clearly expressing what the students are to learn and the type of competence that will be subject to a final assessment on the other, is a difficult balancing act (see section 10.1). Moreover, the guidelines state that "the competence aims may in some cases also be somewhat narrower and express a limited competence". Section 10.2 presents what international research says on matters such as the factors that can contribute to developing exams in the direction of complex competence, and the conditions it is necessary to take into consideration when testing and assessing this collective competence in an exam. In what way can the student perspective be accounted for in an exam, and what consequences does involve students before and/or during an exam have for the curriculum renewal (see section 10.3)?

There is a lack of research on competence-oriented exams, especially in relation to their advantages and disadvantages, psychometric qualities, and retroactive effect on teaching, including unintended side effects. Medical education is an exception in this regard, particularly in the clinical form *Objective Structured Clinical Examination* (OSCE). Therefore, section 10.4 also provides knowledge and experience of testing and assessing complex competence taken from the higher education sector.

10.1 Possibilities and challenges when measuring competence in an exam

Competence-oriented exams are designed to measure more complex abilities and knowledge. This type of exam can lead to in-depth learning processes even in the preparatory phase, in regard to both competence in the subject and the ability to utilise this competence in different contexts. An example of this is problem-solving, which involves analysis and evaluation. In addition, competence-oriented exams make it easier for students to see the relevance of their knowledge and skills, which may stimulate in-depth learning and perseverance. Competence-oriented exams that reflect a connection between the assessment, teaching practice and desired learning outcomes in line with Biggs' model of *constructive alignment* (2003) can help guide learning towards complex competence early on, instead of knowledge detached from competence.

Developing competence-oriented exams is often more demanding than traditional exams (Schaper, Hilkenmeier and Bender, 2013). It is often more difficult to assess competence because more complex abilities and knowledge are usually less precisely defined, and because it is not always possible to develop explicit criteria that define the extent to which an answer is correct or not. These types of exams require additional criteria that can preserve qualitative differences in the responses as well as the degree to which the criteria are met. This can lead to a reduction in objectivity and/or reliability. In all cases, competence-oriented exams require more professional opinion and extensive training of exam developers/exam boards and examiners, as well as facilitation of experience-sharing and reflection (ibid).

It's not just Norway that has challenges with redefining the assessment system to ensure a revised and expanded definition of competence. Schleicher's book *World Class: How to build a 21st-century School System* (2018) describes the challenges in this way: "The dilemma for educators is that routine cognitive skills, the skills that are easiest to teach and easiest to test, are exactly the skills that are also easiest to digitise, automate and outsource." The way we relate to these problems will be significant for whether we succeed in meeting the requirements specified by the curriculum renewal.

10.2 Developing exams that measure competence

Before the exam form is chosen – be it long-answer questions, multiple-choice questions, oral tests or portfolio assessments – it's useful to imagine the situations wherein the students will use the competence later in life (Schaper, Hilkenmeier and Bender, 2013). Therefore, it will be helpful to consider the type of questions that can assess this competence before making a final decision on the exam form. In competence-oriented exams, the exam exercises will typically involve solving and evaluating problems taken from real life in varying complexity (case- or scenario-based exams). Pure reproduction of knowledge will be less appropriate. A well-known disadvantage of these types of exams is an increased uncertainty amongst the students about whether they have found the "correct" solution (ibid.).

The complexity of the curriculum renewal's definition of competence is almost impossible to test with a single exam or exam form. It requires thinking comprehensively about the final assessment as a system. In this case, each question in a single exam paper (or group of questions) concentrates on a single aspect (or group of aspects) of the competency. However, all the questions should collectively cover the competence in its entirety, and – if possible – be integrated into a wider case study or more far-reaching scenario (ibid.). If measurement flaws are inherent in all exam questions, it is better to have multiple small exercises than one large one.

Since competence-oriented exams leave more room for interpretation than traditional exam forms, it is necessary to predefine what can be considered high and low achievement, as well as set clear thresholds for these levels and the development of possible solutions to questions in order to guarantee

accurate assessment (ibid.). Exam developers must establish progression descriptions for the competency and its content dimensions.

It is difficult to imagine that all the elements of the revised definition of competence can be tested through the traditionally established exam forms or solely through an exam. For example, there are limitations to the breadth of the definition of competence and when the student is to learn and reflect on their own learning, as well as work with an area for the long term. The portfolio assessment has been discussed as a new exam form in this context because it could compensate for the fact that the current exam form has signs of being a snapshot or a sample of a student's competence. Furthermore, it will strengthen the diversity of exam forms, which can benefit various student groups. In addition, it has been identified as an assessment form that may be able to include the student perspective by offering options (see section 10.3 for a closer investigation of other options to involve the students).

However, there is a tension between its flexibility and the opportunity for comparison, as identified by Koretz (1998, p. 332):

> "Portfolio assessments have attributes that make them particularly appealing to those who wish to use assessments to encourage richer instruction – for example, the 'authentic' nature of some tasks, the reliance on large tasks, the lack of standardisation, and the close integration of assessment with instruction. But some of these attributes may undermine the ability of the assessments to provide performance data of comparable meaning across large numbers of schools."

Black, Harrison, Hodgen, Marshall and Serret (2011) investigated the components necessary to guarantee a student portfolio's validity in respect of the competence requirements. They concluded that a collection of multiple exercises was required. The validity of summative assessments was dependent on the range of and balance between the content of each student's portfolio, in that the content should reflect the scope and goals of the subject as well as vary in style (form). A potential concern in this context was knowing who actually completed the exercises included in the portfolio, as students work on them at home.

10.3 Student involvement in the exam

The student's active role in the learning process is at the heart of the curriculum renewal's expanded definition of competence and emphasis on in-depth learning. According to *White Paper No. 28 (2015–2016)*, assessment forms and the quality assurance system must support schooling that places a greater emphasis on in-depth learning and systematic progression (p. 123). As a consequence of the new elements of the curriculum renewal, it will be natural to examine the student's role more closely before and/or during the exam.

Involving the students in their own learning, including their own assessments of their own academic performance, is a part of a continuous assessment and has been an important focus area in recent years, among others through the Norwegian Assessment for Learning initiative (2010–2018). However, the questions related to self-assessment and student participation in Udir's annual student survey (Elevundersøkelsen) show that there is still work to be done before this becomes an established practice. The survey also shows that practices vary between schools. Simultaneously, there has been a relatively positive development of these issues compared with other assessment issues in the period 2013–2017.

The current exam systems allow student involvement in exam to a certain degree. For example, the preparatory period of an exam gives the students the opportunity to prepare alone or together with others, and less restrictive question types allow students to choose from different approaches.

10.4 Reliability and validity of assessments of complex competence

From exam research in medical education, it appears that an appropriate selection of different types of exam exercises, context and multiple examiners can guarantee high reliability (Wass, Van der Vleuten, Shatzer and Jones, 2001). Research findings show that sufficient reliability can be achieved in all exam forms – even if there are no standardised tests – given that an acceptable selection of exercises are included of different types, in various contexts and marked by various examiners (Norcini, J. et al., 2018).

The most significant recommendation is to have multiple exercises for each exam which are each marked by different examiners. Research also shows that an adequate selection of exercises has a greater impact on reliability than standardisation; a wisely designed exam can generate reliable results in a reasonable time (ibid.).

Reliability is additionally related to the time aspect of shorter exams being less reliable than those that last longer. Irrespective of the exam form, students' performance on one exercise will not necessarily predict how the students perform on other exercises (Wass, Van der Vleuten, Shatzer and Jones, 2001). Moreover, some exam *forms* may be less reliable than others, such as long-answer questions and oral exams. A consequence of this is that the exam must be of a certain length and cover sufficient aspects of the competence in order to guarantee results that may rightfully be used for the exam's purpose. Using a wider range of exam formats involves including forms that alone are potentially less reliable. However, aggregating various methods and contexts solves this concern (Van der Vleuten and Schuwirth, 2005).

Example from medical education

Objective Structured Clinical Examinations (OSCE):

- A multi-case format consisting of a series of tasks and encounters (stations).
- Introduced to assess higher cognitive abilities and to increase the validity of exams
- Authenticity is achieved by offering candidates simulated real world challenges, either on paper, in computerized forms or in a laboratory setting.
- The items are contextual, vignette-based, or problem-oriented and require reasoning skills rather than straightforward recall of facts.
- Technological development is regarded to have potential in this context because computer simulations can replace written or verbal scenarios and raise the standard of clinical testing.

In summary, this research can indicate that there are examples of how complex competence can be assessed without threatening quality requirements such as validity, reliability and fairness. However, it is a challenge to have an exam that is standardised with an expanded definition of competence as it is difficult to specify competence in such a way that can be reliably measured. This has been done successfully in medical education – however, that system was developed over a long time and with the help of significant resources. Whether this approach is appropriate for such a sizeable system as exams in tenth grade and upper secondary school and training remains a question. To answer this, a careful investigation and thorough discussion are needed. Simultaneously, the final assessment may be considered an integrated system, wherein classwork assumes an important role in competence assessment in order to maintain and strengthen the robustness of the current exam system, which satisfies requirements for reliability, validity and overarching fairness.

11 The significance of technology for exams

Technological development is significant for exams in different ways. This may range from distributing and delivering exams in a digital system to developing exam papers on a digital platform and exploiting the opportunities that this entails, which can also have a significant impact on the assessment aspect and exam content. Assessing responses can also be supported by technology. A recurrent discussion topic is the degree to which the exam should reflect technological development and the large degree of digitalisation happening in the majority of social arenas and the ways that this can happen. This technological development involves new opportunities for assessment, some challenges and, not least, prerequisites with respect to competency and access to digital equipment.

This section summarises our preliminary foundation of knowledge on the significance of technology for exams, with an emphasis on Norwegian research. This knowledge base is characterised by the fact that we currently have little experience in digitally assessing competence in the exam field. In addition, it again appears that the foundation of knowledge is largely based only on surveys. It may be questioned if this approach is suitable for investigating the effects of technology usage or if another type of research is needed, such as intervention studies.

This section is divided into areas that are impacted by digitalisation, prerequisites for change, and preliminary experiences with digital exams.

11.1 Areas that are impacted by digitalisation

In this subsection, we have chosen to highlight the following areas that digital technology can impact or change in an exam:

- Administration and exam implementation
- Technological support and resources for exams
- Exam content
- Exam marking

Administration and exam implementation

One aspect of digitalisation concerns delivering the submission itself or implementing the exam through a digital process. The main purpose of this type of digitalisation is to increase efficiency, information security and privacy. Technology also opens doors for new formats of assessment. Sound files, videos, multimodal texts, and software are just a few examples of digital products that may be relevant for final assessments. The current exam system is digital in the sense that the students can download the exam questions and submit their responses digitally. National development over recent years has primarily focused on renewing administration solutions for exams.

Udir's exam service

PAS, Udir's service for administering exams, and PGS, Udir's service for implementing exams, were developed to prepare, implement, and administer both tests and centrally administered exams.

The gradual introduction of PAS and PGS started in 2008 and has contributed significantly to raising the implementation quality of centrally administered written exams through increased efficiency and improved security.

The two systems currently comprise a digital service for exams that has been used to develop exam questions, gather materials, and register students for exams. Moreover, they are used during the exam implementation itself and for marking and appeals procedures. In autumn 2015, the systems also began to be used for locally administered written exams.

Udir has renewed the administration solution and started the process of acquiring a new implementation solution by 2021. The new solution for developing and implementing exams and tests will be able to offer new question formats and functionality that facilitate assessing competency in new ways and support the marking process.

Technological support and resources for exams

Technological support concerns the use of digital tools in education to support and enrich teaching, learning and assessment processes. In terms of exams, this primarily concerns various digital resources in an exam situation. These types of digital resources can include unrestricted access to the internet or software that supports reading or writing or is subject specific. As mentioned in section 3, internet access trials in exams for selected subjects in upper secondary education and training have been conducted annually from 2012 up to and including 2015. These trials were evaluated on behalf of Udir (Rambøll, 2012; Rambøll, 2013; Rambøll, 2014; Rambøll, 2015). The evaluation reports examined aspects such as the preparation and implementation of unrestricted internet access in exams and revealed associated benefits and satisfaction, as well as results and implications. The subsection on the experience of digital exams summarises the main findings from the final report published in January 2019.

Exam content

The potential of technology to provide new exam forms and assessment tools has led to discourse on the degree to which digital assessment has the potential to measure competencies that previously have been difficult to discern, such as competency related to metacognition (Erstad, 2008; Redecker and Johannessen, 2013). However, it is more difficult to find data on how this can specifically be done, and the positions in the discourse are only vaguely founded on evidence. However, this increasing focus on which competencies can be measured is considered as a sign of a paradigm shift in digital assessment from focusing on technological use to streamlining assessment processes and increasing scoring reliability (Redecker and Johannessen, 2013).

How technology can be used to further develop the exam system is particularly relevant in light of the revised definition of competence in the curriculum renewal (Norwegian Ministry of Education and Research, 2016), and there is a need for more evidence-based knowledge in this area. Technological development is also a driving force for changing the school's content and by extension the competencies it is relevant to measure (*NOU 2015: 8*). This can be evidenced by the new areas and topics that are being introduced into school curricula, by the changing weighting between content

areas, and by the new multidisciplinary topics or overarching subject competences that are appearing in the curricula. Examples of this include the introduction of programming into mathematics subjects, digital text forms and textual expressions in Norwegian subjects, source criticism competence, and digital skills as a foundational skill (Hultin and Berge, 2014).

Exam marking

Technology provides opportunities for scoring papers automatically and may therefore support examiners when assessing exam responses. The quality of this automatic scoring will vary by paper type, but, when appropriate, it will be able to significantly reduce marking time, as well as reduce the risk of scoring mistakes.

Another possibility of digital marking is that it enables examiner reliability to be investigated by saving data. If data from all examiners is saved at student and question level, then corresponding studies will have access to more information that they do currently, which will lead to a more meaningful and enlightening investigation of examiner reliability.

11.2 Digital competence and prerequisites

The Technology and Programming for Everyone report (*Teknologi og programmering for alle*) describes how digital technology can be used to generate new opportunities to improve the quality and efficiency of learning and teaching processes, but it emphasises that these opportunities have some prerequisites and require some changes, particularly in terms of students' and teachers' digital competency (Sanne et al., 2016).

The curriculum presupposes that teachers utilise digital tools in their lessons as well as help develop the students' digital skills in the subject. This has been a premise for all subjects since the introduction of the basic skills through the Knowledge Promotion Reform. In order to instruct students in digital competency, teachers need to possess a professional digital competency themselves (Norwegian Directorate for Education and Training, 2018a). Having a professional digital competency partly involves having knowledge of digital assessment forms and the skills to utilise them in teaching and learning processes. If the students are familiar with various forms of digital assessment, they have a better basis for handling a digital exam situation, but this presupposes that the teachers have the competency to include these assessment forms in their teaching. The Norwegian researchers who worked with the international comparative study International Computer and Information Literacy Study (ICILS) described the professional-pedagogic competency among teachers when it comes to using digital resources in a qualified manner as lacking (Hatlevik and Throndsen, 2015).

The ICILS study also found that around a fourth of ninth grade Norwegian students (students aged 14–15) had such weak digital skills that they would struggle with being able to fully participate in education, work or society (Hatlevik and Throndsen, 2015). A third of the Norwegian students were able to search for information, critically analyse sources and create digital presentations according to more specific criteria. Approximately half of the students demonstrated that they could use the computer as a tool and were able to use digital resources to solve simple problems. They had some awareness of privacy but simultaneously lacked critical assessment abilities for how personal information online can be used. A fourth of the students had knowledge solely of elementary file management and text editing, and had only a surface-level understanding of data security and online behaviour. The SMIL (*Sammenhengen mellom IKT-bruk og læringsutbytte*) study on the connection between IT use and learning output in upper secondary education and training at an organisational level shows that the competency of students in upper secondary education and training in terms of

professional use of ICT and digital learning strategies is generally too low (Krumsvik et al., 2013).

We have observed significant differences between schools in terms of access to digital resources and the extent to which training in the use of these resources is prioritised in lessons, which also appear in the national Monitor surveys that map the school's digital status (Egeberg, Hultin and Berge, 2016; Harlevik, Egeberg, Gudmundsdottir, Loftsgarden and Loi, 2013). Access to various forms of digital equipment is generally high in Norwegian schools. However, findings from the annual Monitor surveys and ICILS show that the quality of the equipment and associated infrastructure varies somewhat, and there are significant differences in levels of access between schools (Egeberg et al., 2016; Hatlevik et al., 2013; Hatlevik and Throndsen, 2015). The *Monitor 2016 survey* (*Monitor skole, 2016*) investigated the digital maturity of primary and lower secondary school at an organisational level and found that, among the factors examined, equipment was reported as having the greatest variation of perceived quality among the participating schools (Egeberg et al., 2016).

However, the SMIL study found that differences between IT usage and learning outcomes in upper secondary education and training were primarily caused by student groups' usage patterns and are no longer based on unequal access to technology (Krumsvik et al., 2013). This finding is in line with a general development that is often described as a transition from first generation digital skills to second generation (Hatlevik and Throndsen, 2015). The *Digital Division Lines* (*Digital skillelinjer*) sub-report in the evaluation of the tenth-grade mathematics exam investigated the type of teaching the students had received in using the digital resources relevant for the exam, and how they have been prepared to use these in the exam itself (Bjørnset, Fossum, Rogstad, Smestad and Talberg, 2018). The report describes certain student groups who are digitally privileged, in the sense that they have better prerequisites for succeeding in the exam than other students. This advantage may be connected to technical conditions, such as access to equipment and infrastructure, or teaching conditions, such as the scope of quality of education in digital skills.

11.3 Digital exam experiences

Up to and including 2012, the standard system for written exams in primary and lower secondary school and upper secondary school has been to deliver the exam responses electronically through a digital exam implementation system (Norwegian Directorate for Education and Training, 2016). Evaluations of these implementations largely focus on unrestricted internet access and use of digital resources. The bi-annual survey *Questions for Norwegian Schools and School Leaders* (*Spørsmål til Skole-Norge*) in spring 2017 included questions on the use of net-based resources for centrally administered exams (Federici, Gjerustad, Vaagland, Larsen, Rønsen and Hovdhaugen, 2017). The survey showed that around two out of three school owners and school leaders answered that they offer net-based resources. Net-based resources are most used in upper secondary education and training, where 88 per cent answered that they offer this. Among primary and lower secondary schools, 62 per cent answered that they offer this. However, the SMIL study shows that central digital resources in school subjects and the student's multi-media and multimodal learning in upper secondary education and training are drawn on by the exam forms only to a small degree (Krumsvik, Egelandsdal, Sarastuen, Jones and Eikeland, 2013).

Digital assessment forms highlight the significance of various digital competencies, such as production competence, tool knowledge and genre understanding. As evidenced in the evaluation of trials of unrestricted internet access in an exam in upper secondary school, this exam form can have a "washback" effect on the teaching (Rambøll, 2014). The teachers at the schools that participated in the trial use the internet in teaching to a larger degree than teachers at the reference schools. These teachers implement exams and all-day exams that allow the students access to the internet and focus their teaching on using and critically analysing sources. This effect can also extend beyond the defined

competence aims in the curriculum. The qualitative Monitor survey from 2010 focuses on teachers in lower secondary education who prioritise teaching text formatting in order to successfully meet the exam's structural digital requirements, and this competence requires going beyond the curriculum outline to be measured (Hatlevik, Tømte, Skaug and Ottestad, 2010).

In January 2019, a final report from the evaluation of unrestricted internet access in the exams for seven subjects from the Education Programme for Specialisation in General Studies in upper secondary school (Rambøll, 2019). The report mainly described findings from surveys completed by students, teachers, people responsible for administering exams and IT support technicians at schools and throughout the county. The survey was conducted from May to June 2018. Key themes in the report include technical maturity and the implementation of exams with unrestricted internet access, marking and frameworks with respect to detecting cheating in the year's exams, authenticity and relevance connected to the exam system's accordance with teaching practice for the subject in question, and formulating questions and assessments, as well as support in terms of the school's facilitation for students with specific needs.

To summarise the main findings from the evaluation of unrestricted internet access in exams (Rambøll, 2019):

- There are few technical or practical challenges related to the implementation of exams.
- The majority of schools have implemented preventative measures such as monitoring internet use during an
  exam, training invigilators and recruiting more and more digitally competent invigilators.
- 90 per cent of people responsible for administering exams have informed the students about cheating and plagiarism before the exam. However, the proportion of students who stated that they have received this information is lower.
- Qualitative interviews indicate that the students have sufficient understanding of cheating and plagiarism, but that there are grey areas that require clarification.
- Only one case of cheating in the exam for the subjects in question was reported.
- 93 per cent of the teachers in the target group stated that the use of the internet is included as an important part of the students' learning in their teaching.
- 96 per cent of the teachers in the target group stated that their students have received training in source analysis.
- The teachers in the target group stated that they implement other exams with access to the internet to a larger degree than other teachers in the control group do.
- The students found it useful to have access to the internet in the exam, but both examiners and teachers were more unsure of the benefits of the access.
- The exam papers in the spring were found to be well suited for this form of exam . Simultaneously, 62 per cent of the examiners and 36 per cent of the teachers stated that access to the internet requires new exam exercises.
- One out of five students found that the exam form was more stressful than the exams without the internet.
- This especially applied to girls and students in the subjects Politics and Human Rights and Social English.
- Exam guidelines and assessment criteria were perceived as clear among the examiners.
- Some examiners reported that they were stricter in their assessment of responses from students who had access to the internet than responses from students who did not have access to the internet, although this cannot be determined by the marks.
- The scope of support for implementing the exam is about equal for students in the target group and the control group.

12 Teacher education and assessment competency

Investigating teacher education or suggesting changes to it are not explicit parts of the exam review group's mandate. However, since teacher education undoubtedly presents the best opportunity to develop teachers' assessment competency for both formative and summative assessments, so that they can guarantee validity, reliability and fairness to the highest possible degree, we have decided to include it here. We know that teachers are deeply involved in final assessments in tenth grade and in upper secondary education and training through their work developing exam papers, acting as examiners and, in particular, implementing the overall achievement grades and marking. These tasks require an expanded assessment competency.

Assessment is discussed in section 2 of the Norwegian teacher education curriculum regulations for years 1–7 and 5–10 as well as for years 8–13, which set requirements for the learning outcomes of the various study programmes in line with the Norwegian qualification framework. In the teacher education curriculum regulations for years 1–7 and 5–10, which were established in 2013, assessment is discussed in two points in the "Knowledge" and "Skills" parts respectively. The points emphasise that, following the conclusion of their education, the student teachers should have thorough knowledge of matters such as assessment and mapping tools. The student teachers should also be able

Norwegian Curriculum Regulations for Differentiated Primary and Lower Secondary Teacher Education Programmes for Years 1–7 and Years 5–10

Section 2, Learning outcomes
Knowledge

- Student teachers have thorough knowledge of teaching at a beginner's level, basic skills, assessment and mapping tools, class management and assessment of the students' learning, and how to promote learning in a subject

Skills

- Student teachers can analyse, assess, and document students' learning, provide feedback that promotes learning, adapt their teaching to the students' needs and expectations, utilise a variety of teaching methods and contribute to the students' ability to reflect on their own learning and development

Norwegian Curriculum Regulations for Practical Pedagogical Education Programmes for General Subjects (PPU-A) and for University-Led Teacher Education Programmes for Years 8–13

Section 2, Learning outcomes

- Knowledge

Skills

- Student teachers can describe indicators of competency, assess and document students' learning, provide feedback that promotes learning and contribute to the students' ability to reflect on their own learning and development

Norwegian Curriculum Regulations for Practical Pedagogical Education Programmes for Vocational Subjects (PPU-Y) and for University-Led Teacher Education Programmes for Years 8–13

Section 2, Learning outcomes

- Knowledge

Skills

- Student teachers can describe and document students' learning, provide feedback that promotes learning and contribute to the students'/apprentices' ability to reflect on their own learning

to assess students' learning and provide feedback that promotes learning. The curriculum regulations for university-led teacher education were established in 2018 and refer to the revised definition of competence and indicators of achievement in the "Skills" part, but do not mention any areas of competence that the student teachers are to be instructed in.

Teacher education has changed significantly over recent years. This is particularly in relation to primary and lower secondary teacher education since, as of and including autumn 2017, student teachers are required to undertake a five-year master's degree. In this section, we refer to the most recent curriculum frameworks and guidelines and limit ourselves to the five biggest study programmes. Assessment is discussed in section 2 of all the above-mentioned curriculum regulations, which set requirements for the learning outcomes of the various study programmes in line with the Norwegian qualification framework.

In the teacher education curriculum regulations for years 1–7 and 5–10, which were established in 2016, assessment is discussed in two points in the "Knowledge" and "Skills" parts respectively. The points emphasise that, following the conclusion of the education, the student teachers should have thorough knowledge of matters such as assessment and mapping tools. The student teachers should also be able to assess students' learning and provide feedback that promotes learning. The latter is also included in the regulations for PPU[8]-Y, which was established in 2013. The curriculum regulations for PPU-A and university-led teacher education were established in 2015 and 2013, respectively. Both refer to the revised definition of competence as well as indicators of achievement in the "Skills" part, but do not mention any areas of competence that the student teachers are to be instructed in.

Marking is not explicitly mentioned in any of the regulations as an area for which teachers require specific competency. There are a number of other factors in these points and other learning outcomes in the curriculum regulations that presuppose assessment competence, such as adapting lessons and understanding what promotes learning and guarantees progression, but these are less directly identifiable as "assessment".

The new national curriculum regulations for teacher education, which were established in the teacher education unit of Universities Norway (UHR-LU[9]) in 2017, are not very specific regarding the programmes' core features and only require that "the teacher education shall qualify the student teachers to conduct ethically sound assessments". However, the guidelines for the university-led teacher education more clearly specify assessment knowledge and skills (NRLU, 2017). Emphasis is placed on assessment for learning and progress assessments, but the final assessment is explicitly mentioned in the teaching methodology part as an area of competence, wherein the student teachers "learn to conduct assessment for learning and final assessments, use assessment criteria and provide thorough reasoning for their assessment in the subject". The teaching placement is also mentioned, the purpose of which is to ensure that the student teacher "has experience-based knowledge of students' learning processes and assessment for and of teaching". Assessments are referred to by the curriculum regulations for PPU-A as a reoccurring matter that must be safeguarded by the institutions and that adds learning outcomes such as "extensive knowledge of the forms of teaching, work and assessment, both generally and for specific subjects". In line with university-led teacher education, the PPU student teachers should be able to "administer progress and final assessments", provide an

[8] Praktisk-pedagogisk utdanning (PPU): One-year undergraduate teacher training programme *(requirement for employment in primary and secondary school for candidates with a vocational or general academic educational background.)* PPU-Y is for candidates with a vocational educational background. PPU-A is for candidates with general academic educational background.

[9] Universitets- og Høgskolerådet- Lærerutdanningen

explanation of the mark and have the opportunity to "test forms of formative and summative assessments that they themselves can use as teachers" (NRLU, 2017b). PPU-Y also emphasises the latter aspect and additionally mentions practical exams (NRLU, 2018).

There is extremely limited research and up-to-date systematic information and knowledge about how to ensure these assessment-related qualification requirements in teacher education. Furthermore, we know little of the learning outcomes or the effect of teacher education when it comes to assessment competence. This applies to a particularly high degree to summative assessments and marking.

According to an old survey, assessment competence is one of the areas of teacher education with the poorest quality of results (Finne et al., 2011). School leaders evaluate this area of the education significantly less positively than, for example, social competency and professional identity. However, this difference in the evaluation of results quality also applies to student teachers and teacher educators. The teacher education unit of Universities Norway (UHR) (2011) refers to these results in a separate report from the same year and demands that more explicit national guidelines and means of ensuring that a student has adhered to these guidelines be included in assessment education. A survey from 2013 showed that the majority of teacher educators report a solid understanding of assessment *for* learning[10], but it is unclear what conclusions we can draw from this about their knowledge of summative assessments.

A slightly more recent report on changes to teacher education does not discuss these changes in detail; however, it is not known whether the new teacher education models will perform better regarding assessment competence (Munthe et al., 2014). However, the programme at the Centre for Professional Learning in Teacher Education (ProTed) may possibly be understood as an indication that the main focus of teacher education is directed at areas other than summative assessments and marking. ProTed is Norway's first centre for professional learning and the result of a long-term collaboration project between the University of Oslo and the University of Tromsø. It is financed by the Norwegian Agency for Quality Assurance in Education (NOKUT). The centre's objective is to promote quality in higher education, and it has subsequently developed impressive measures in teacher education. Its multiple projects also include assessment education, but based on annual reports and other documents, these would appear to revolve around formative assessments, otherwise known as assessments for learning (see for example ProTed, 2016; 2017).

There is a lack of knowledge generally about the extent to which student teachers participate in assessment work when they are on their placement. For example, placement periods do not often coincide with the end of the autumn and spring terms, when the students take their final assessments, as the student teachers are generally in the middle of intense exam and assessment periods themselves. There is also a question of how accessible the school's assessment work is to the student teachers when so much of the work related to, for example, classwork assessments and the lead-up to the exam is conducted by individual teachers and/or during their office or meeting hours – which are not as available to student teachers on their placement. However, this is a relationship that we know little about and for which we need to gather more systematic knowledge.

In UHR-LU[11], work is occasionally done on assessments in teacher education, but these cases

[10] https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/2013/ntnu_a_bidra_til_skolebasert_kompetanseutvikling.pdf (in Norwegian)
[11] https://www.uhr.no/strategiske-enheter/fagstrategiske-enheter/uhr-larerutdanning/om-uhr-larerutdanning/ (in Norwegian)

emphasise the assessment of the student teachers. The basic concept is that teacher education can contribute to the provision of useful examples of work in teaching and assessment. As a result, it is important that we be extremely conscious of how assessments are conducted, even within teacher education. Examples of this concept include projects from ProTed, a report from the survey of marks in mathematics in primary and lower secondary school teacher education in 2014 (workgroup appointed by the Norwegian Association for Teacher Education (NRLU), 2015) and the 2019 Teacher Education Conference on Futuristic Assessments in Teacher Education (Lærerutdanningskonferansen om framtidsrettet vurdering i lærerutdanning). ProTed in particular has developed new assessment methods by using tablets to assess student teachers on their placement and provide automatic feedback in the exam (NOKUT, 2015).

Generally, it is important to highlight that teacher education is a basic education, and that teachers also learn through informal supplementary training and formal continuing education, such as the Qualifications for Quality (Kompetanse for kvalitet) strategy. More recent political documents emphasise the need for collaboration between basic and higher education and in the professional community, such as the Norwegian Teacher Education Strategy 2025 (Lærerutdanning 2025)[12]. Work in developing the competence of teachers in schools does exist regionally, such as the SKUV project[13] at NTNU in Trøndelag. The measure is an example of a partnership between school owners and the university, initiated from within the teaching placement.

General pedagogical or didactic textbooks provide teachers and teacher education with little support in their work on summative assessments. There are introductory books in English, but these are poorly adapted to the Norwegian context and the Norwegian overall achievement grade system in particular. There is a lack of systemised knowledge about the development of student teachers' assessment competence during their placement, except for comprehensive development programmes that have focused on formative assessment, such as the national assessment for learning program (2010-2018)[14] and the national initiative on the lower secondary level (2012-2017) [15][16] (Ungdomstrinn i utvikling) (which featured assessment for learning as a reoccurring theme). Multiple county governors arrange examiner training and collections of overall achievement grades, both on their own initiative and in collaboration with Udir. However, it is notable that competence development programmes in relation to summative assessments and marking have not been as comprehensive as those in relation to formative assessments in recent years.

[12] Lærerutdanningene 2025. Nasjonal strategi for kvalitet og samarbeid i lærerutdanningene - regjeringen.no (in Norwegian)
[13] https://www.ntnu.no/ilu/skuv (in Norwegian)
[14] the-norwegian-assessment-for-learning-programme_final-report-2018.pdf (udir.no)
[15] Ungdomstrinn i utvikling (UiU) 2012-2017 – sluttrapport (udir.no) (in Norwegian)
[16] Sluttrapport: Ungdomstrinn i utvikling (udir.no) (in Norwegian)

13 Status of the foundation of knowledge and problems with the exam system in Norway

This final section of the report presents the exam review group's assessment of the foundation of knowledge on the exam. The first part presents the main conclusions and resulting actions. The second part briefly summarises the most important findings from each section. Based on these findings, the third part identifies problems and questions relating to the exam system that still need to be addressed, and that the exam review group intends to investigate in its future work. Two further partial deliveries shall investigate and complement these initial discussions, which will enable us to advise on the curricula in partial delivery 2 during March 2019 and to provide recommendations for changes to the exam system as a result of the curriculum renewal and technological developments. These will be submitted for a final decision on 15 March 2019.

13.1 Status of the foundation of knowledge and main conclusions

In this report, we have assembled a foundation of knowledge on the exam system in Norway. This collection is based on the current exam system, how it has emerged and its officially defined purpose. We have adopted a broad perspective on the quality of the current exam system, including criteria such as validity, reliability and fairness, as well as the student perspective. We have also examined the relationship between examination marks and marks awarded for classwork as two forms of final assessment. The curriculum and the regulations associated with the Norwegian Education Act provide frameworks for the content, organisation, and assessment of exams, as well as guidelines to ensure the quality of the exam system, and have therefore been included as an important perspective.

With the stipulation that this is preliminary documentation of the foundation of knowledge, one main conclusion is that while there are some user insight and experience-based knowledge on the exam, there are large gaps in this research. We must emphasise that a lack of research does not necessarily mean that no good qualitative work has been conducted on exams, only that there is a lack of systematic research evidence in this area. The studies that do exist focus almost exclusively on surveys that reflect what the participants think or remember, but otherwise have some shortcomings as a systematic investigation of processes, effects and long-term consequences. Robust information on processes, effects and consequences requires experiments and trials in schools and should be illustrated both quantitatively and qualitatively.

Considering the significance of the exam for the individual student and the exam's status in society, there has been relatively little research on exam quality. This is particularly significant when compared with the research devoted to the large international surveys and national tests and the public attention exams receive every year. An inadequate foundation of knowledge makes it difficult for the exam review group to give detailed answers to the most immediate challenges. It can be challenging to generalise results from other countries to Norway – or to try to apply results from the university sector to primary and lower secondary education – as the context and framework are so different. However, this research provides indications of what may be relevant for further work in this mandate.

The review of the foundation of knowledge provides a strong basis for investigating a more integrated approach to the final assessment. This applies both to the relationship between the exam and classwork assessments and to quality assurance:

- When the final assessment is planned as an integrated and coordinated system, it may become more apparent which competence is to be tested in an exam and which competence is to be tested through classwork assessments. Therefore, we need a comprehensive framework for

final assessments that relates the various final assessment systems to the revised definition of competence, so that the entire definition of competence is tested systematically.

- Quality assurance of the exam should also be examined and planned in an integrated manner. There are some routines for monitoring quality, but not all the data is developed into documentation that is then made available. If there is no framework in place, it is difficult to see whether quality assurance is happening in an integrated manner. It may be useful to use one of the established frameworks to fulfil this purpose (for example, AEA Europe, 2017; Stobart, 2009).

  It will be a challenge to prioritise between the different considerations. For example, ensuring high levels of validity, reliability and fairness (marking without systematic discrepancies for certain groups) is difficult as well as time-consuming and resource intensive, as the measures that can strengthen validity may weaken reliability (and vice versa). Moreover, some quality criteria are clearly related to the exam itself (validity, reliability and fairness), while others are related to wider contexts that are more difficult to control (interpretation of results and consequences of exam implementation). In addition, it is probable that the priorities look different in exam research from a measurement perspective – which often emphasises quality assurance of the exams themselves (e.g. construct validity and examiner reliability), preferably before they are implemented – compared to a school perspective – which often examines how exams regulate actions and knowledge in practice (consequential validity). One approach in these cases may be combining the two perspectives in order to achieve a better balance between them.

  Moreover, thorough quality assurance requires a new approach to data storage. For example, data for marking exists only at an overall level (the marks for an exam from the examiners), while effective analyses require data at the lowest level possible: a registration of points per exam question or assessment criterion, per examiner. This level of data would enable a closer investigation of potential causes of agreement/disagreement among examiners. For example, it is possible that an examiner's assessment differs from one exam question to another (unstable marking, low intra-examiner reliability), or that the extent of disagreement varies with question *type* due to an examiner valuing aspects of assessment differently. It is also possible that the examiners' assessments correspond across questions or assessment criteria but still result in different marks due to differences in the overall assessment. If data exists at the marking level, the details that make up the marks are not visible and it becomes difficult to address causes of disagreement, for example, in examiner training.

The review of the foundation of knowledge also demonstrates that there is a great need for knowledge related to three key themes in the report:

- Given that *validity* (legitimacy) is the most important criterion of quality in an exam context, this perspective has been highly prioritised in this foundation of knowledge. Even though there is a lot of knowledge available on this theoretical approach, and even though the foundation of knowledge is developing a clear understanding of the concept of validity, it has been difficult to find studies that investigate the legitimacy of the current Norwegian exam system. The exam boards are currently an important part of the system for securing validity (exam content), but we have little systematised research on, for example, the validity of exam

content per subject or on the exams that are administered to different students. A significant aspect of fairness is that an exam should measure equivalent competence throughout the years that students are competing for admission to higher education.

The *reliability* (dependability) of an exam has been better investigated, at least when it comes to mathematics, writing in Norwegian and marking. However, there are some studies on oral or other exam forms on subjects other than Norwegian and mathematics, and some studies that have investigated the marking process of each question and not just marking as the final stage.

The relationship between examination marks and marks awarded for classwork has been thoroughly researched in terms of mark discrepancies across selected criteria, but there is little research on how this is possible or whether systematic mark discrepancies should be counteracted or compensated for, for example, across gender, regions, or subjects. If the differences are related to something other than the students' competence, this may undermine the fairness of an exam. Overall, this gap in the research reflects a general ambiguity around the relationship between exams and classwork assessments.

- The student perspective has not been studied much either, as there is almost no research that directly investigates the student's perception of exam forms or exam questions in Norway. Reports tend to focus only on the teachers' impressions of their students' subjective evaluations of exam forms and questions. The few studies we have indicate that there are systematic differences in perceptions and mastery of various exam forms, and that questions as well as instructions are sometimes found to be ambiguous. Evidently, more systematic studies are needed on how the students interpret and master the questions. In addition, it is difficult to differentiate between the possible reasons for exam anxiety or pressure because individual expectations for tackling stress and circumstances outside the school are rarely considered in the existing studies.

- Even though the education sector has acquired experience and competence in developing and assessing exams from the competence-based curricula in the Knowledge Promotion Reform, the revised definition of competence will set new requirements for test development and marking. Research from other countries and from the university sector can provide us with an initial indicator of how these demands may be met, but whether the recommendations are appropriate for the Norwegian exam system remains an open question to be investigated. The significant changes in the aftermath of the curriculum renewal are to be evaluated, and the same should apply to possible changes in the exam system and the effect of this.

Therefore, the main conclusions of this report are that the final assessment and its quality assurance require an integrated approach, and that there is a significant need for research that investigates the curriculum renewal and exam system's prerequisites, processes and results. Both measures can contribute to new types of discussion about exams. Ensuring validity, reliability and a fair exam system requires pilot trials and time to consider whether the conclusions drawn from the assessment can be considered legitimate. This is particularly important when it applies to *high-stakes* situations, such as exams.

13.2 Summary of the foundation of knowledge

The emergence of the current exam system

Norway has a long tradition of basing admission to higher education on exams, which are administered by teachers in close cooperation with national authorities. The exam system has historically represented governing authorities' most important tool for controlling the teachers' marking policies. The historical review shows that teachers have been recognised as competent to assess the quality of students' performances and in this way have been largely responsible for controlling admission to higher education and professional life. The exam system in its entirety and the most important procedures have been relatively stable over the last few decades, but assessment criteria have been debated significantly and changed from a norm-referenced assessment principle to an achievement-based principle.

The exam's purpose and organisation

The exam's purpose is indicated by the regulations associated with the Norwegian Education Act, where examination marks, along with marks awarded for classwork, are to be an expression of students' competence upon the end of their education in a subject. Examination marks and marks awarded for classwork together provide a basis for admission to both upper secondary and higher education. This gives exams a formal function beyond the expression of a student's final competence in a subject. It may be argued that the legitimacy of an exam as part of a ranking system rests entirely on the examination mark being an expression of a student's competence and being equally expressive of the student's competence regardless of the subject. Simultaneously, it may be argued that the legitimacy of the requirement that students demonstrate their competence and that there be an additional quantification of this competence rests entirely on the examination mark being used for something meaningful (such as admission to upper secondary and higher education). The role of the admission system will therefore be important in the discussion about what an exam shall or shall not be.

Examination marks and marks awarded for classwork are both currently an expression of the student's competence at the end of their education in a subject, but they must be considered as distinct from each other. The regulations associated with the Norwegian Education Act may appear to be ambiguous about whether the exam is to test the entire breadth of the curriculum. This ambiguity can result in the various actors in the system interpreting the relationship between exams and classwork assessments differently.

The exam lottery means that student groups are split up, in most cases based on random selection. The foundation of knowledge shows that this lottery does not perhaps allow all students the opportunity to demonstrate their competence in a valid way. This is not compatible with the concept of final assessment as an integrated system. The students themselves may perceive the allocation of different exams as unfair because they do not get the same opportunity to show their competency. Additionally, the subjects and the number of exams on a student's academic record can vary, which may impact the average admission rates to upper secondary and higher education.

The private candidate scheme is an offer to document competency in a subject that the candidate has not previously completed any education or final assessment in, or wishes to improve their mark for. The number of exams taken to improve an existing mark has grown over time and currently comprises a significant proportion of private candidates, which brings the purpose of this scheme into question. In addition, the scheme is administratively challenging to implement, which in many cases has consequences for the possibility of further developing the exam.

Even though the exam system has been relatively stable, some trials and changes have taken place in

recent years as a result of input from users, officials and professionals. Measures such as new exam forms or access to study aids reflect the fact that the exam system in a subject can limit the students' opportunity to clearly demonstrate their final competency, and that the increasing access to study aids requires new discussions on what is to be assessed and how.

The quality of the current exam system

The final assessment is to provide impartial and relevant information on a students' competency in subjects studied. To achieve this, teachers and examiners should be supported in their assessments by explicit goals, assessment criteria, guidance and quality assurance. This report has investigated various quality criteria (particularly validity, reliability and fairness), taken an integrated approach that safeguards agreement between these criteria and examined the whole process – from the development of exam papers, via the administration of the exam and finalising marks, to the way the results are interpreted and used in practice. A challenge when it comes to the quality of an exam is that it is difficult to know whether an exam paper has the desired qualities before it is used, such as through pilot trials, as the papers must be kept secret prior to the exam.

Exam validity

The mathematics exam is almost the only exam form that has been investigated to some degree in terms of validity. In this case, the majority of teachers agree that there is a meaningful correspondence between the competence aims and what the students are tested in. Findings from the KAL project provide some information on validity in the exam in Norwegian writing, even though this survey is relatively old (Berge et al., 2005). In general, there is a lack of research on the cohesion between the curriculum and the exam and on the extent to which there are differences between subjects and the various exam forms. There is disagreement among school leaders and school owners on whether exams are suitable for demonstrating competency in all subjects. There is also disagreement on the extent to which it is clear what competence the students are to demonstrate in exams.

The foundation of knowledge demonstrates that the current exam system may have multiple implicit roles apart from its legally defined purpose of certifying, quality assuring, further developing assessment practices, guiding the teaching and, sometimes, supporting teaching in Norwegian primary and lower secondary education. Multiple and various purposes and roles may lead to various interpretations of exam results and various side effects of any subsequent changes. It is therefore very important to clarify the implicit roles the exam has in practice. However, there is little research in this area, and some of the roles the exam has in the education system are therefore possibly understated. For example, this applies to the exam's role in raising the examiners' competency and the exam's contribution to the professionalisation of the assessment. Research documents that school leaders and teachers perceive participation in marking as important for the professional development of both the school as an organisation and the individual teacher.

Exam reliability

An important indicator of quality is that an exam paper receives consistent assessments from multiple examiners to ensure that the marking is not characterised by coincidence. This requires unambiguous exam papers with clear instructions, explicit assessment criteria (i.e. indicators of achievement) and comprehensive examiner training to ensure communities of interpretation.

Indicators of achievement from exams with central marking are used a lot and perceived as useful in the schools' assessment work. In addition, school leaders and school owners have prepared local indicators. The indicators are formulated for the mark groups 2 (E), 3–4 (C–D) and 5–6 (A–B). Because 3 and 4 comprise a particularly large proportion of the marks, the teachers have requested clearer assessment criteria, which will make it easier to differentiate between a 3 and a 4.

Teachers who have participated in examiner training find it to be extremely useful, and school leaders perceive that the examiners' experiences contribute to improving assessment competence throughout the school. However, a lack of examiner consensus in the assessment of exams seems to be a problem in the majority of subjects. The students' responses and the examiners' assessments benefit from explicit expectations, clear aims and detailed requirements for content, structure and criteria weighting. To achieve high levels of reliability, it is imperative that exams consist of a greater number of questions. This is more important than standardising questions. In addition, the questions within an exam should be assessed by different teachers and examiners.

There is a lack of systematic research on how municipalities and counties work qualitatively with examination results from locally administered exams. User insight and various surveys show that there is a range of collaboration forms related to assessment but do not mention anything about the quality of these collaborative areas and the degree to which this work is connected to locally administered exams.

One challenge with the current data source from research on exam reliability is that examiner information for each subject is gathered only at a student level and not at a question level within a student's exam. This makes it difficult to retrospectively investigate the causes of possible problems in examiner consensus.

Systematic differences between exams and classwork assessments

Differences between exams and classwork assessments need not be a cause for concern in themselves unless they are due to systematic differences between groups. Gender-related differences and differences between private and state schools, large and small schools, and high-achieving and low-achieving schools, as well as differences across regions, demonstrate that there are systematic differences in what classwork assessments and exams measure that cannot be reasonably connected to the students' final competency in that subject. This leads to a situation where different students are given different opportunities, which is not compatible with the concept of a fair assessment.

There is little research on the causes of these systematic differences. Examples indicate possible procedural weaknesses or culturally conditioned subject norms in the final assessment in Norwegian schools. Exams and external marking can be thought to lessen these undesirable side-effects if the marks awarded for classwork reflect factors beyond the curriculum goals, such as student effort or organisation and behaviour.

In addition to systematic differences across student groups, there are differences between exams and classwork assessments connected to year and subject. This type of variation at the marking level may turn out to be a source of unfair competition for the same places on a programme of study. Another phenomenon that may have consequences for admission to higher education and professional life is inequalities due to the subjects' value on the student's academic record based on the teaching hours.

Final subject assessment

The unique nature of individual subjects has been a focus of assessment research to a very limited extent. When testing competence in line with the new curricula in the curriculum renewal, it will be important to recognise curriculum content and structure, which are often minimised factors. There are subjects that are clearly anchored in research disciplines in higher education and have a more rigid, hierarchical and sequential structure. On the other hand, there are subjects that have a weaker connection to their academic research disciplines and are less hierarchical and more segmented. The curriculum renewal's idea to define the key elements may contribute to recognising the subject's content and structure.

There are school subjects that comprise the basis for the teachers' and examiners' assessment. Subject assessments are based on a narrower or wider basis, where the narrow approach is dominated by use of a single assessment form, such as a written exam. The wide approach is dominated by a wider selection of assessments, such as written and oral or oral and practical assessments in an exam. In the Norwegian context, there is a lot of evidence that indicates that the exam form for the individual subject contributes to defining these narrow or wide frameworks for assessment.

Students' experience of exams

Through the School Student Union of Norway, students have been highlighting weaknesses of the exam system since 1963. These weaknesses include the students not getting to demonstrate their full competence, the current form impacting the students' performance to a significant degree and the exams from the lottery being drawn largely randomly. Research indicates that listening to the students, for example on whether they find questions or instructions to be ambiguous or unambiguous, may contribute to increasing the validity in the development of exam questions.

The foundation of knowledge shows that students' exam anxiety increases when they perceive an exam to be significant, when it is expected to be difficult and when the circumstances surrounding the exam situation are stressful. Students prefer assessment formats that reduce stress and nervousness, but there is no ideal assessment form. The students' preferences vary with factors such as the degree of question transparency, gender, performance requirements and learning strategies. Based on international research, it seems to be important to vary exam forms as much as possible so that different student groups have the opportunity to perform in the best possible way.

Testing the curriculum renewal's expanded definition of competence

Competence-oriented exams are designed to measure more complex abilities and knowledge. The curriculum renewal's definition of competence highlights understanding, the use of knowledge and skills in familiar and unfamiliar situations and learning to learn and reflect on this learning. In-depth learning can be considered a prerequisite for developing complex competency as expressed in the curriculum renewal. However, the development of competence-oriented exams is often more demanding because more complex abilities and knowledge are usually less precisely defined and because it is not always possible to develop explicit criteria that define the extent to which an answer is correct or not.

Experience from medical education shows that before the exam form is chosen, it is useful to imagine the situations wherein students will use the competence later in life and what types of questions are suitable for assessing this competence. The complexity of the curriculum renewal's definition of competence is almost impossible to test with a single exam or exam form and requires thinking about final assessments as an integrated system. Portfolio assessments have been highlighted as a potentially useful element in testing complex competence because they can compensate for the fact that the current exam form has signs of being a snapshot or a sample of a student's competence. Furthermore, they will strengthen the diversity of exam forms, which can benefit various student groups. But this form of assessment also has some challenges related to assessment work. An example of this is that there is a lack of "control over" whether the students have completed the work that is included in the portfolio.

Creating open and overarching goals that enable the students to transfer what they have learnt to new situations on the one hand, and clearly expressing what the students are to learn and the type of competence that will be subject to a final assessment on the other, is a difficult balancing act. There is a significant lack of research on competence-oriented exams, especially in relation to their advantages

and disadvantages, psychometric qualities and retroactive effect on teaching, including unintended side effects. From exam research in medical education, it appears that a large selection of different types of exam exercises, context and multiple examiners can guarantee high reliability.

The significance of technology for exams

Digital technology can impact or change different areas of the exam, such as exam administration, use of study aids, exam content and marking process. Access to various forms of digital equipment is generally high in Norwegian schools, but there are some prerequisites to utilising opportunities to improve the quality and efficiency of an exam, and there is a need for change, particularly in terms of students' and teachers' digital competency. The professional-pedagogic competency among teachers when it comes to using digital resources varies greatly. Student groups who have greater access to digital tools (access to equipment and infrastructure) and receive more training in using these tools (scope of quality of education in digital skills) have a greater chance of succeeding in the exam than other students.

The current exam system is digital in the sense that students can download the exam questions and submit their responses digitally. Its main purpose is to increase efficiency, information security and privacy, but it also opens doors for new formats of assessment, such as sound files, videos or multimodal texts. The new solution for developing and implementing exams, which will be acquired by 2021, will be capable of offering new question formats and giving additional support to the marking process. Moreover, technological developments can provide opportunities for the exam questions to reflect the width of the definition of competence and subsequently become more valid. By gaining access to automatic paper scoring, technology will be able to significantly reduce marking time.

Technological development is also a driving force for changing school content and, by extension, the competencies it is relevant to measure. Examples of this include the introduction of programming into mathematics subjects, digital text forms and textual expressions in Norwegian subjects, source criticism competence and digital skills as a foundational skill. In addition, technological support can extend to using digital resources in an exam situation, such as unrestricted access to the internet or software that supports reading or writing or is subject-specific.

Simultaneously, technological developments create new challenges: the new opportunities must be balanced with academic traditions and the requirement that students be able to demonstrate their mastery of basic knowledge and skills in individual subjects. Furthermore, it takes time for students to learn to use study aids (paper and digital) appropriately, and examiners need a community of interpretation in order to ensure a common understanding of what characterises an effective use of sources. A survey conducted after the trial of open internet access in exams in upper secondary school indicates that the majority of students found it useful to have access to the internet in the exam, but both examiners and teachers were less sure of the benefits of the access. Simultaneously, girls in particular found that the exams with unrestricted internet access were more stressful than the exams without the internet.

Teacher education and assessment competency

The curriculum regulations for teacher education are different for years 1–7 and 5–10, and for PPU and the university-led programme. The teacher education curriculum regulations for years 1–7 and 5–10, which were established in 2013, emphasise that, following the conclusion of the education, the student teachers should have thorough knowledge of matters such as assessment and mapping tools. The student teachers should also be able to assess students' learning and provide feedback that promotes learning. The curriculum regulations for PPU and university-led teacher education refer to the revised definition of competence and indicators of achievement in the "Skills" part, but they do

not mention any areas of competence that the student teachers are to be instructed in. However, these guidelines do explicitly mention the final assessment and assessment of learning. Marking is not explicitly mentioned in any of these curriculum regulations or guidelines.

There is extremely limited research and up-to-date systematic information or knowledge about how to ensure these assessment-related qualification requirements in teacher education. Furthermore, we know little of the learning outcomes or the effect of teacher education when it comes to assessment competence. There is some regional work in competence development for schoolteachers, but there is a noticeable lack of similarly comprehensive competence development programmes for summative assessments and marking when compared to the development programmes for formative assessments and assessment for learning, which can possibly be interpreted as a common thread running from teacher education to supplementary training and continuing education.

13.3 Problems and questions for further work

The foundation of knowledge that we have compiled here, even though it is limited to some degree, leads to some formative questions that the exam review group shall investigate in their further work. The areas highlighted as particularly important include:

- clearly defining the purpose of exams
- discussing opportunities to test the expanded definition of competence in a subject in an exam
- examining the relationship between classwork assessments and exams
- assessing whether the exam lottery is appropriate
- further developing the quality assurance of exams on the basis of validity, reliability and fairness
- assessing the significance of new technology for exams

Some discussions as a potential starting point for further work

The review of the foundation of knowledge demonstrates that the exam system has multiple roles beyond its formally defined purpose, which are not always equally apparent. Simultaneously, the exam is to accommodate validity requirements, which makes it important to clarify the purpose of the exam and pay attention to the implicit roles it may have so that these can form the basis of validation processes connected to the exam's design and implementation.

If the main purpose of an examination mark is to be an external assessment in addition to the mark awarded for classwork, we can question ourselves as to why the majority of marks awarded for classwork are not followed up with an exam. Simultaneously, the extent to which it is appropriate to combine the exam's quality assurance function with the overall achievement grades is not clear.

Currently, marks awarded for classwork comprise approximately 80 per cent of a student's academic record, while examination marks comprise 20 per cent. Therefore, the following questions are relevant: Is it reasonable that examination marks and marks awarded for classwork are valued equally on an academic record, or that written and oral exams are valued equally when the quality of written exams can be assured in a different way to oral ones? Is it reasonable that some subjects are weighted more on a student's academic record than others, considering the number of teaching hours for that subject in upper secondary school?

The exam lottery must be approached from a wider perspective that includes systematic thinking about the exam's fairness, predictability and how exams are organised. It may be challenging to

imagine how the exam lottery can enter into an integrated system for exams and marks awarded for classwork that efficiently guarantees the perspectives mentioned here.

Whether exams can or should test the extent of the students' competency, or whether exams should test only certain parts of the competency, is a relevant issue. If the exam and marks awarded for classwork complement each other as parts of an integrated system for assessment and if they are planned, these issues can be avoided. Simultaneously, this type of integrated approach creates an opportunity to include a wide spectrum of exam forms that collectively test complex competence.

A question that can be posed in this context is whether all exam forms are equally appropriate. Some competence aims may be less suitable for testing in a written or oral exam. If a larger proportion of the competence is not suitable for the exam form that is traditionally used, then new forms should be investigated. A possible approach is to base the choice of exam forms to a greater degree on the competency/work form the students will require in everyday or professional life, as well as in continuing and higher education.

It may be challenging to test students' teamwork abilities and/or solutions and products they have developed together in an individual exam. The current framework allows for different solutions as long as the assessment is still individual. Simultaneously, it is possible to view exams and classwork assessments as an integrated system, where classwork assessments better guarantee some aspects of competence than exams.

Exam forms and assessment processes are anchored in the subjects' content and structure, and this relationship should not be taken for granted. Considering the curriculum renewal's multidisciplinary topics and work with in-depth learning in multiple parallel subjects, this is a topical problem that should be addressed in further discussions on the exam and final assessments. Research on validity will be able to illustrate how an exam is designed and used in various contexts and for different purposes.

The foundation of knowledge shows differences in the marking between schools, subjects, genders, class years, etc. Research should investigate how the differences can be understood and the degree to which they can be justified or changed. For example, how should the levels of different subjects be reconciled with each other? To what degree does the exam for one subject measure the same competence as the exam for the same subject the previous year? Fairness is compromised if the requirements are systematically higher in certain subjects or years compared to others.

Thorough quality assurance requires measures to be systematically planned based on a framework (see the attachment for an example), a new approach to data storage and documentation of results being made available. In some respects, it is relatively easy to take steps to improve an exam's quality. This particularly applies to improving reliability, which can be achieved by developing indicators of achievement for all marks, using a larger number of questions in one exam paper that are assessed by different examiners and developing explicit assessment criteria beforehand that clarify expectations as well as detailed requirements for content, structure and criteria weighting.

From the student perspective, it is natural to question whether they are given sufficient opportunity to influence the exam content and how they are tested. Another part of the discussion in this context could focus on whether the exam can be followed up with a more detailed form of feedback than the mark. This type of feedback would be resource-intensive if not given automatically. Therefore, its usefulness must be investigated first.

There is currently no collective overview of the total costs of both locally administered and centrally administered exams. Even though Udir has an overview of its own costs for centrally administered written exams, including question development and production (approx. 34 million), system support

and IT management (approx. 27 million), and marking and appeals procedures (approx. 128 million), there is no statement showing local costs. In 2019, Udir will conduct an investigation to provide a better basis for assessing the guidance on complying with the applicable budget in the exam review group's mandate.

Dialogue should be opened with teacher education in order to strengthen the assessment competency in newly qualified teachers. Teachers have a significant responsibility in assessing and examining students. Teacher education presents the best opportunity to build the solid foundation for assessment competency that is necessary for conducting assessments in a valid, reliable and fair manner. Increased production of Norwegian specialised literature/textbooks on summative assessments and textbooks adapted to the various types of teacher education would be particularly desirable.

14 Bibliography

Andresen, S., Fossum, A., Rogstad, J. & Smestad, B. (2017). *På prøve. Evaluering av matematikkeksamen på 10. trinn våren 2017*. Fafo.

Workgroup appointed by the Norwegian Association for Teacher Education (NRLU): Christiansen, A., Enge, O. & Lode, B. (2015). *Rapport fra karakterundersøkelsen i matematikk i GLU-utdanningene i 2014.* Retrieved from https://docplayer.me/6669704-Rapport-fra-karakterundersokelsen-i-matematikk-i-glu-utdanningene-i-2014.html (last accessed: 11 Feb 2019)

Association of Educational Assessment – Europe (AEA Europe). (2017). *European Framework of Standards for Educational Assessment 1.0.* Retrieved from https://www.aea-europe.net/wp-content/uploads/2017/07/SW_Framework_of_European_Standards.pdf (last accessed: 06 February 2019)

Backmann, K. & Sivesind, K. (2012). Kunnskapsløftet som reformprogram: fra betingelser til forventninger. In T. Englund, E. Forsberg & D. Sundberg (Eds.), *Vad räknas som kunskap? Läroplansteoretiska utsikter ock inblickar i lärarutbildningen ock skola* (pp. 240–260). Stockholm: Liber.

Baeten, M., Struyven, K. & Dochy, F. (2008). *Students' assessment preferences and approaches to learning in new learning environments: A replica study*. New York: AERA (paper presented at AERA March 2008).

Baird, J.-A., & Hopfenbeck, T.N. (2016). Curriculum in the Twenty-First Century and the Future of Examinations. In D. Wyse, L. Hayward, & J. Pandya (Eds.), *The SAGE handbook of curriculum, pedagogy and assessment* (pp. 821–837). Los Angeles; London; New Delhi; Singapore; Washington, DC: SAGE.

Beller, M. & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles: A Journal of Research, 42* (pp. 1–21).

Ben-Chaim, D. & Zoller, U. (1997). Examination-type preferences of secondary school students and their teachers in the science disciplines. *Instructional Science, 25(5)* (pp. 347–367).

Ben-Shakar, G. & Sinai, Y. (1991). Gender differences in multiple choice tests: The role of differential guessing. *Journal of Educational Measurement, 28* (pp. 23–35).

Biggs, J.B. (2003). *Teaching for quality learning at university (2nd edition)*. Buckingham: Open University Press/Society for Research into Higher Education.

Birenbaum, M. & Dochy, F. (1996). Introduction. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. xiii–xv). Boston: Kluwer.

Birenbaum, M. & Feldman, R.A. (1998). Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research, 40(1)* (pp. 90–97).

Bjørnset, M., Fossum, A., Rogstad, J., Smestad, B. & Talberg, N. (2018). *Digitale skillelinjer: Evaluering av matematikkeksamen på 10. trinn våren 2018*. Fafo-rapport 2018:36.

Black, P., Harrison, C., Lee, C.S., Marshall, B. & William, D. (2003). *Assessment for learning, putting it into practice*. Open University Press.

Black, P., Harrison, C., Hodgen, J., Marshall, B. & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy and Practice, 18(4)* (pp. 451–469).

Borgonovi, F., Ferraram, A. & Maghnouj, S. (2018). The gender gap in educational outcomes in Norway. *OECD Education Working Papers, 183.* OECD Publishing: Paris. http://dx.doi.org/10.1787/f8ef1489-en

Broadfoot, P. (2007). *An introduction to assessment*. London; New York: Continuum.

Brookhart, S.M. (2013). Grading. In J.H. McMillan (Ed.), *SAGE Handbook of Research on Classroom Assessment* (pp. 257–272). USA: SAGE.

Buland, T., Engvik, G., Fjørtoft, H., Langseth, I., Sandvik, L.V. & Mordal, S. (2014). *Vurdering i skolen. Utvikling av kompetanse og fellesskap. Sluttrapport fra prosjektet Forskning på individuell vurdering i skolen* (FIVIS). NTNU.

Bøhn, H. (2017). *What is to be assessed? Teachers' understanding of constructs in an oral English examination in Norway* (doctoral thesis). Universitetet i Oslo.

Carlsen, C. (2013). *Guarding the Guardians. Rating scale and rater training effects on reliability and validity of scores of an oral test of Norwegian as a second language* (doctoral thesis). Universitetet i Bergen.

Crooks, T.J., Kane, M.T. & Cohen, A.S. (1996). Threats to the Valid Use of Assessments. *Assessment in Education: Principles, Policy and Practice, 3(3)* (pp. 265–286).

Dale, E.L. (2008). *Fellesskolen – reproduksjon av sosial ulikhet*. Oslo: Cappelen akademisk forlag.

Dale, E.L. & Wærness, J.I. (2006). *Vurdering og læring i en elevaktiv skole*. Oslo: Universitetsforlaget.

Dochy, F. & McDowell, L. (1997). Introduction assessment as a tool for learning. *Studies in Educational Evaluation, 23(4)* (pp. 279–298).

Duncan, C.R. & Noonan, B. (2005). Factors Affecting Teachers' Grading and Assessment Practices. *The Alberta Journal of Educational Research, 53(1)* (pp. 1–21).

Eccles, J. (1983). Expectancies, values and academic behavior. In J.T. Spence (Ed.), *Academic and achievement motives*. San Francisco: Freeman.

Eckstein, M.A. & Noah, H.J. (1993). *Secondary School Examinations. International Perspectives on Policies and Practice*. New Haven: Open University Press.

Egeberg, G., Hultin, H. & Berge, O. (2016). *Monitor skole 2016: Skolens digitale tilstand*. Oslo: Senter for IKT i utdanningen.

Eggen, A.E. (2004). *Alfa and Omega in Student Assessment; Exploring Identities of Secondary School Science Teachers* (doctoral thesis). Department of Teacher Education and School Research, University

of Oslo.

Erstad, O. (2008). Changing Assessment Practices and the Role of IT. In J. Voogt & G. Knezek (Eds.), *International Handbook of Information Technology in Primary and Secondary Education, 20* (pp. 181–194). Springer US.

Evensen, L.S., Berge, K.L., Thygesen, R., Matre, S. & Solheim, R. (2016). Standards as a tool for teaching and assessing cross-curricular writing. *The Curriculum Journal, 27* (pp. 229–245).

Federici, R.A., Gjerustad, C., Vaagland, K., Larsen, E.H., Rønsen, E. & Hovdhaugen, E. (2017). Spørsmål til Skole-Norge våren 2017. *NIFU-rapport 2017, 12.* Oslo.

Finne, H., Jensberg, H., Aaslid, B.E., Haugsbakken, H., Holth Mathiesen, I. & Mordal, S. (2011). *Oppfatninger av studiekvalitet i lærerutdanningen blant studenter, lærerutdannere, øvingslærere og rektorer* (=SINTEF rapport; A18011). Trondheim: SINTEF.

Forsøksrådet for skoleverket. (1969). S*tandardiserte prøver i skolen. Forsøk og reform i skolen – nr 16.* Oslo: Universitetsforlaget.

Galloway, T.A., Kirkebøen, L.J. & Rønning, M. (2011). *Karakterpraksis i grunnskoler: sammenheng mellom standpunkt- og eksamenskarakter*. SSB.

Gellman, E. & Berkowitz, M. (1993). Test-item type: What students prefer and why. *College Student Journal, 27(1)* (pp. 17–26).

Gjerustad, C., Waagene, E. & Salvanes, K.V. (2015). Spørsmål til Skole-Norge våren 2014. NIFU.

Gjone, G. (1993). Types of problems and how students in Norway solve them. In N. Mogens (Ed.), *Cases of assessment in Mathematics Education: An ICMI Study* (pp. 107–118). Amsterdam: Kluwer Academic Press.

Gustafsson, J.E. & Erickson, G. (2018). Nationella prov i Sverige – tradition, utmaning och förändring. *Acta Didactica Norge, 12(4)*. http://dx.doi.org/10.5617/adno.6434

Haladyna, T.M. & Downing, S.M. (2005). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice, 23(1)* (pp. 17–27).

Harlen, W. (2005). Teachers' Summative Practices and Assessment for Learning – Tensions and Synergies. *The Curriculum Journal, 16(2)* (pp. 207–223).

Harris, L.R. & Brown, G.T.L. (2016). The Human and Social Experience of Assessment: Valuing the Person and the Context. In G.T.L. Brown & L.R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 1–17). New York: Routledge, Taylor & Francis Group.

Hatlevik, O.E., Tømte, K., Skaug, J.H. & Ottestad, G. (2010). *Monitor skole 2010: Samtaler om IKT i skolen*. Oslo: Senter for IKT i utdanningen.

Hatlevik, O.E., Egeberg, G., Gudmundsdottir, G.B., Loftsgarden, M. & Loi, M. (2013). *Monitor skole.* Oslo: Senter for IKT i utdanningen.

Hatlevik, O.E. & Throndsen, I. (2015). *Læring av IKT : Elevenes digitale ferdigheter og bruk av IKT i*

ICILS 2013. Oslo: Universitetsforlaget.

Herman, J.L. & Baker, E.L. (2009). Assessment policy: Making sense of the babel. In G. Sykes, B. Schneider & D. Plank (Eds.), *Handbook of Education Policy Research.* London: SAGE.

Hill, K.T. (1984). Debilitating motivation and testing: A major educational problem – Possible solutions and policy applications. In R.E. Ames & C. Ames (Eds.), *Research on motivation in education, 1 : Student motivation.* New York: Academic Press.

Hill, K.T. & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. *Elementary School Journal, 85* (pp. 105–126).

Hovde, P. & Olsen, S.O. (2015). *Utredning – Digital eksamen NTNU 2015–2019*. NTNU.

Hovdhaugen, E., Seland, I., Lødding, B., Prøitz, T. & Vibe, N. (2014). Karakter i offentlige og private videregående skoler. En analyse av eksamens- og standpunktkarakter i norsk og matematikk og rutiner for standpunktvurdering i offentlig og private videregående skoler. NIFU. *Rapport 24/2014.*

Hovdhaugen, E., Prøitz, T. & Seland, I. (2018). Eksamens- og standpunktkarakterer – to sider av samme sak? *Acta Didactica Norge, 12(4)*.

Hultin, H. & Berge, O. (2014). *Notat til utvalgsarbeid om digital kompetanse*. Oslo: Senter for IKT i utdanningen.

Hægeland, T., Kirkebøen, L.J., Raaum, O. & Salvanes, K.G. (2005). *Skolebidragsindikatorer: Beregnet for avgangskarakterer fra grunnskolen for skoleårene 2002–2003 og 2003–2004 (Rapporter SSB 2005/33)*. Oslo: Statistisk sentralbyrå.

Jarning, H. & Aas, G.H. (2008). Between Common Schooling and the Academe: The International Examinations Inquiry in Norway, 1935–1961. In M. Lawn (Ed.), *An Atlantic Crossing? The Work of the International Examination Inquiry, its Researchers, Methods and Influence* (pp. 181–204). Oxford: Symposium Books.

Kane, M.T. (2015). Explicating validity. Assessment in Education: *Principles, Policy and Practice, 23(2)* (pp. 1–14).

Kirke-, utdannings- og forskningsdepartementet. (1996). *Om elevvurdering, skolebasert vurdering og nasjonalt vurderingssystem* (Meld. St. 47 (1995–1996)). Oslo: Departementet.

Kommunerevisjonen. (2013). *Standpunktkarakterer i videregående skole – likebehandles elevene?* Oslo: Oslo kommune kommunerevisjonen.

Koretz, D. (1998). Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education: Principles, Policy and Practice, 5(3)* (pp. 309–334).

Krogh, L.C. (2016). *Kreativitet og ambivalens: En undersøkelse av variasjoner i vurdering og kjennetegn ved sprikvurderte tekster fra eksamen i norsk hovedmål 2015* (master's thesis). Høgskolen i Sørøst-Norge.

Krumsvik, R.J., Egelandsdal, K., Sarastuen, N.K., Jones, L.Ø. & Eikeland, O.J. (2013). *Sammenhengen mellom IKT-bruk og læringsutbytte (SMIL) i videregående opplæring*. Bergen.

Kunnskapsdepartementet. (2013). *På rett vei.* (Meld. Paper 20 (2012–2013)).

Kunnskapsdepartementet. (2016). *Fag – Fordypning – Forståelse — En fornyelse av Kunnskapsløftet.* (Meld. Paper No. 28 (2015–2016).

Kunnskapsdepartementet. (2017). *Organisering av skoleåret i videregående opplæring*, rapport fra arbeidsgruppa oppnevnt av KD. Retrieved from https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/01.11.2018/ntnu_a_bidra_til_skolebasert_kompetanseutvikling.pdf (last accessed: 01 Nov 2018)

Lawn, M. (2008). *M. Lawn (Ed.), The Work of the International Examination Inquiry, its Researchers, Methods and Influence*. Oxford: Symposium Books.

Lekholm, A.K. & Cliffordsson, C. (2008). Discrepancies between School Grades and Test Scores at Individual and School Level: Effects of Gender and Family Background. *Educational Research and Evaluation, 14(2)* (pp. 181–199).

Lekholm, A.K. & Cliffordsson, C. (2009). Effects of Student Characteristics on Grades in Compulsory School. *Educational Research and Evaluation, 15(1)* (pp. 1–23).

Lysne, A. (1999). *Karakterer og kompetanse. Stridstema i norsk skolehistorie*. AVA forlag.

Lysne, A. (2004). *Karakterer og kompetanse. Kampen om skolen*. AVA forlag.

Lysne, A. (2006). Assessment Theory and Practice of Students' Outcomes in the Nordic Countries. *Scandinavian Journal of Educational Research, 50(3)* (pp. 327–359).

Lundahl, Ch. & Tveit, S. (2014). Att legitimera nationella prov i Sverige och i Norge – en fråga om profession och tradition. *Pedagogisk Forskning i Sverige, 19(4–5)* (pp. 297–323).

Markus, K.A. & Borsboom, D. (2013). *Frontiers of test validity theory: measurement, causation and meaning*. New York: Routledge, Taylor & Francis Group.

McDowell, L. (1995). The impact of innovative assessment on student learning. *Innovations in Education and Training International, 32(4)* (pp. 302–313).

McMillan, J.H. (2003). Understanding and improving teachers' classroom assessment decision-making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 2(4)* (pp. 34–43).

Moss, P.A. (2007). Reconstructing Validity. *Educational Researcher, 36(8)* (pp. 470–476).

Moss, Pamela A., Girard, B.J. & Haniford, L.C. (2006). Validity in Educational Assessment. *Review of Research in Education, 30* (Special Issue on Rethinking Learning: What Counts as Learning and What Learning Counts) (pp. 109–162).

Muller, J. (2009). Forms of Knowledge and Curriculum Coherence. *Journal of Education and Work, 22(4)* (pp. 205–226).

Munthe, E., Solbakken, J.I., Hjetland, H. & Hustad, B.C. (2014). *Lærerutdanninger i endring: Indre utvikling – ytre kontekstuelle og strukturelle hinder (Følgegruppen for lærerutdanningsreformen; Rapport Nr. 4).*

Nassar, Y.H.B., Qaraeen, K. & Naba'h, A.A. (2011). Secondary School Students' Perceptions of Essay and Multiple-Choice Type Exams. *Dirasat, Educational Sciences, 38(1)* (pp. 345–358).

Natriello, G. & Dornbusch, S.M. (1984). *Teacher evaluative standards and student effort*. New York: Longman.

Nesman, M. & Kovač, V.B. (2016). Privatister – hvem er de, og hva motiverer dem til å lykke på eksamen? Kartlegging av bakgrunnsvariabler og deres intensjon i lys av en utvidet versjon av teorien om planlagt adferd. *Nordisk tidsskrift for pedagogikk og kritikk*.

Newton, P.E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education, 14(2)* (pp. 149–170).

NOKUT. (2015). *Centre for Professional Learning in Teacher Education (ProTed): Mid-term evaluation – Centre of Excellence in Higher Education.* Retrieved from https://www.uv.uio.no/proted/om/arsrapporter/proted_mid-term_evaluation_report_2015.pdf (last accessed: 12 Feb 2019)

Nordenbo, S.E., Allerup, P., Andersen, H.L., Dolin, J., Korp, H., Larsen, M.S. & Østergaard, S. (2009). *Pædagogisk brug af test – Et systematisk review*. Copenhagen: Danmarks Pædagogiske Universitets Forlag.

Norcini, J., Brownell Anderson, M., Bollela, V., Burch, V., Costa, M.J., Duvivier, R., Hays, R., Palacios Mackay, M.F., Roberts, T. & Swanson, D. (2018). 2018 Consensus framework for good assessment. *Medical Teacher, 9* (pp. 1–8). https://doi.org/10.1080/0142159X.2018.1500016

NOU (Norges offentlige utredninger) 2015: 8. (2015). *Fremtidens skole: Fornyelse av fag og kompetanser*. Oslo: Kunnskapsdepartementet.

NOU (Norges offentlige utredninger) 2018: 15. (2018). *Kvalifisert, forberedt og motivert — Et kunnskapsgrunnlag om struktur og innhold i videregående opplæring*. Oslo: Kunnskapsdepartementet.

NOU (Norges offentlige utredninger) 2019: 3. (2019). *Nye sjanser – bedre læring: Kjønnsforskjeller i skoleprestasjoner og utdanningsløp*. Oslo: Kunnskapsdepartementet.

Norgesuniversitetet. (2015). *Sluttrapport fra Ekspertgruppa for digital vurdering og eksamen per februar 2015.* Retrieved from https://norgesuniversitetet.no/files/images/content/sluttrapport_ekspertgruppa_for_digital_vurdering_og_eksamen_februar_2015.pdf (last accessed: 27 Nov 2018)

Nasjonalt råd for lærerutdanning (NRLU). (2017). *Nasjonale retningslinjer for lektorutdanning for trinn 8–13.* Retrieved from https://www.uhr.no/_f/p1/i4d4335f1-1715-4f6e-ab44-0dca372d7488/lektorutdanning_8_13_vedtatt_13_11_2017.pdf (last accessed: 12 Feb 2019)

Nasjonalt råd for lærerutdanning (NRLU). (2017b). *Nasjonale retningslinjer for praktisk pedagogisk utdanning – allmennfag.* Retrieved from https://www.uhr.no/_f/p1/i13d351d8-d4a8-4c93-ac64-f0d2fbbdc6c6/godkjente-retningslinjer-ppu.pdf (last accessed: 18 Feb 2019)

Nasjonalt råd for lærerutdanning (NRLU). (2018). *Nasjonale retningslinjer for praktisk-pedagogisk*

utdanning for yrkesfag trinn 8-13. Retrieved from https://www.uhr.no/_f/p1/i6c34f03d-e46c-4ce8-8c90-9c47b488bbc1/nasjonale-retningslinjer-for-praktisk-pedagogisk-utdanning-for-yrkesfag-trinn-8-13_ferdig.pdf (last accessed: 19 Feb 2019)

Nygård Arntzen, H. (2015). *Matematikkeksamen gjennom tre reformer: En analyse av avgangseksamen på høyeste nivå i den videregående skolen* (master's thesis). Universitetet i Tromsø.

Pellegrino, J.W., Chudowsky, N., Glaser, R. & National Research Council (U.S.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Popham, W.J. & Husek, T.R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6(1)* (pp. 1–9).

ProTed. (2016). *Centre for Professional Learning in Teacher Education: Annual report for 2016.* Retrieved from https://www.uv.uio.no/proted/om/arsrapporter/annual_report_2016_proted.pdf (last accessed: 12 Feb 2019)

ProTed. (2017). Centre for Professional Learning in Teacher Education: *Annual report for 2017.* Retrieved from https://www.uv.uio.no/proted/om/arsrapporter/annual-report-for-proted_2017.pdf (last accessed 12 Feb 2019)

Prøitz, T.S. & Borgen, J.S. (2010). *Rettferdig standpunktvurdering – det (u)muliges kunst?* NIFU STEP report 16/2010.

Prøitz, T.S. & Borgen, J.S. Variations in grading practice—subjects matter. *Education Inquiry, 4(3)* (pp. 1–22) (in press, forthcoming September 2013).

Prøitz, T.S. (2018). *Ten years later – variations in grading practices revisited*. Paper presented at LOaPP project meeting 30 Nov 2018 USN.

Rambøll. (2012). *Forsøk med internett til eksamen: Sluttrapport*. Retrieved from https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Eksamen-med-tilgang-til-Internett/ (last accessed: 27 Nov 2018)

Rambøll. (2013). *Evaluering av eksamen med tilgang til internett: Sluttrapport.* Retrieved from https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Eksamen-med-tilgan-til- internett/ (last accessed: 27 Nov 2018)

Rambøll. (2014). *Forsøk med tilgang til internett på eksamen*. Retrieved from https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Forsokt-med-Internett-pa-eksamen/ (last accessed: 27 Nov 2018)

Rambøll. (2015). *Evaluering av forsøk med tilgang til internett på eksamen 2014–2015*. Retrieved from https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/forsok-med-tilgang-til-internett-pa-eksamen/ (last accessed: 27 Nov 2018)

Rambøll. (2019). *Evaluering av åpent internett til eksamen: Sluttrapport*.

Redecker, C. & Johannessen, Ø. (2013). Changing Assessment: Towards a New Assessment Paradigm Using ICT. *European Journal of Education, 48(1)* (pp. 79–96). Blackwell.

Resh, N. (2009). Justice in Grades Allocation: Teachers' Perspective. *Social Psychology of Education, 12(3)* (pp. 315–325).

Sambell, K., McDowell, L. & 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)* (pp. 349–371).

Buland, T., Engvik, G., Fjørtoft, H., Langseth, I., Sandvik, L.V. & Mordal, S. (2012). *Vurdering i skolen. Intensjoner og forståelse. Delrapport 1 fra prosjektet Forskning på individuell vurdering i skolen (FIVIS)*. Trondheim: NTNU.

Sandvik, L.V. & Buland, T. (2013). *Vurdering i skolen. Operasjonaliseringer og praksiser Delrapport 2 fra prosjektet Forskning på individuell vurdering i skolen (FIVIS)*. Trondheim: NTNU/SINTEF.

Sanne, A., Berge, O., Bungum, B., Jørgensen, E.C., Kluge, A., Kristensen, T.E., Mørken, K.M., Svorkmo, A. & Voll, L.O. (2016). *Teknologi og programmering for alle: En faggjennomgang med forslag til endringer i grunnopplæringen*. Oslo. Retrieved from https://www.udir.no/globalassets/filer/tall-og-forskning/forskningsrapporter/teknologi-og-programmering-for-alle.pdf (last accessed: 19 Nov 2018)

Schaper, N., Hilkenmeier, F. & Bender, E. (2013). *Umsetzungshilfen für kompetenzorientiertes Prüfen: HRK-Zusatzgutachten*. Germany. Retrieved from https://www.hrk-nexus.de/fileadmin/redaktion/hrk-nexus/07-Downloads/07-03-Material/zusatzgutachten.pdf (last accessed: 06 Feb 2019)

Schunk, D. (1984). Self-efficacy perspective on achievement behavior. *Educational Psychologist, 19* (pp. 48–58).

Sejersted, F. (2005). *Sosialdemokratiets tidsalder. Norge og Sverige i det 20. århundre. Andre del av historieverket Norge og Sverige gjennom 200 år.* Oslo: Pax forlag.

Seland, I., Lødding, B. & Prøitz, T.S. (2015). *Delrapport 1 fra evaluering av forsøk med halvårsvurdering med én eller to karakterer i norsk. Litteraturstudie*. NIFU-rapport 33/2015. Oslo: NIFU.

Seland, A., Hovdhaugen, E., Lødding, B., Prøitz, T. & Rønsen, E. (2018). *Sluttrapport fra evaluering av forsøk med halvårsvurdering med én eller to karakterer i norsk*. Oslo.

Sjaastad, J., Carlsten, T.C. & Wollscheid, S. (2016). *Får elevene den opplæringen de har krav på? Kartlegging av undervisningstimer med kvalifiserte lærere i videregående opplæring. Rapport 26/2016*. Oslo: NIFU.

Smestad, B. & Fossum, A. (2019). *Primary school exams in calculations/mathematics in Norway 1946–2017: Content and form. CERME 2019.*

Solheim, R. & Matre, S. (2014). *Lærarsamtalar om elevtekstar: Mot eit felles fagspråk om skriving og vurdering.*

Steffensen, K. & Ziade, S.E. (2009). *Skoleresultater 2008. En kartlegging av karakterer fra grunnskoler og videregående skoler i Norge*. Rapporter 2009/23. Statistisk sentralbyrå.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51(2)* (pp. 161–179).

Tobias, S. (1985). Test anxiety: Interference, defective skills, and cognitive capacity. *Educational Psychologist, 20* (pp. 135–142).

Tveit, S. (2007). Elevvurdering i Kunnskapsløftet. In H. Hølleland (Ed.), *På vei mot Kunnskapsløftet*. Oslo: Cappelen akademisk forlag.

Tveit, S. (2018). (Trans)national Trends and Cultures of Educational Assessment: Reception and Resistance of National Testing in Norway and Sweden during the Twentieth Century. In C. Alarcon & M. Lawn (Eds.), *Assessment Cultures*, *(Studia Educationis Historica series).* Berlin: Peter Lang.

Tveit, S. & Olsen, R.V. (2018). Eksamens mange roller i sertifisering, styring og støtte av læring og undervisning i norsk grunnopplæring. *Acta Didactica Norge, 12(4)*.

Universitets- og Høgskolerådet. (2011). *En helhetlig tilnærming til lærerutdanning*: *Rapport fra en arbeidsgruppe nedsatt av Nasjonalt råd for lærerutdanning.* Oslo: UHR.

Universitets- og Høgskolerådet – Lærerutdanning (UHL-LU). (2017). *Felleskapittel – Nasjonale Retningslinjer for Lærarutdanningane.* Retrieved from https://www.uhr.no/_f/p1/i4fbd09e0-6a5f-4a13-9e89-3971c57cfa5d/fellestekst-for-retningslinjene-for-alle-typer-av-larerutdanning.pdf (last accessed: 12 Feb 2019)

Utdanningsdirektoratet. (2009). *Utdanningsspeilet 2008*. Retrieved from https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Utdanningsspeilet-2008-ei-analyse-av-grunnopplaringa-2009/ (last accessed: 06 Feb 2019)

Utdanningsdirektoratet. (2013). *Utdanningsspeilet 2013*. Retrieved from https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/utdanningsspeilet_2013/us2013.pdf (last accessed: 06 Feb 2019)

Utdanningsdirektoratet. (2015). *Rapport om utviklingen i klager på standpunktkarakterer fra 2010 til 2015*. Oslo.

Utdanningsdirektoratet. (2016). *Erfaringer og vurderinger av eksamen våren 2012 og 2013*. Oslo. Last accessed: 14 Nov 2018 Retrieved from https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Erfaringer-og-vurderinger-av-eksamen-varen-2012-og-2013/ (last accessed: 06 Feb 2019)

Utdanningsdirektoratet. (2017). *Utdanningsspeilet 2017*. Retrieved from http://utdanningsspeilet.udir.no/2017/ (last accessed: 06 Feb 2019)

Utdanningsdirektoratet. (2018). *Utdanningsspeilet 2018*. Retrieved from https://www.udir.no/tall-og-forskning/finn-forskning/tema/utdanningsspeilet/ (last accessed: 16 Dec 2018)

Utdanningsdirektoratet. (2018a). *Rammeverk for lærerens profesjonsfaglige digitale kompetanse (PfDK)*. Oslo. Last accessed: 29 Nov 2018 Retrieved from https://www.udir.no/kvalitet-og-kompetanse/profesjonsfaglig-digital-kompetanse/rammeverk-larerens-profesjonsfaglige-digitale-komp/ (last accessed: 06 Feb 2019)

Utdanningsdirektoratet. (2018b). *Trekkordning ved eksamen for grunnskolen og videregående opplæring Udir-2-2018*. Retrieved from

https://www.udir.no/regelverkstolkninger/opplaring/eksamen/trekkordning-ved-eksamen-for-grunnskole-og-videregaende-opplaring-udir-2-2018/ (last accessed: 17 Dec 2018)

Utdanningsdirektoratet. (2018c). *Rammeverk for eksamen*. Retrieved from https://www.udir.no/eksamen-og-prover/eksamen/rammeverk-eksamen/5.-analyse-av-eksamen-og-bruk-av-resultatene/ (last accessed: 16 Dec 2018)

Utdanningsdirektoratet. (2018d). *Eksamensundersøkelse engelsk 10. trinn*. Utdanningsdirektoratet. (31.5.2018).

Utdanningsdirektoratet. (2019). *Sluttrapport vurdering for læring*.
Retrieved from https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Erfaringer-og-vurderinger-av-eksamen-varen-2012-og-2013/

Utdanningsdirektoratet. (2018f). *Retningslinjer for læreplanutvikling* (unpublished).

Utdannings- og forskningsdepartementet. (2004). *Kultur for læring.* (Meld. St. 30 (2003–2004)).

van de Watering, G., Gijbels, D., Dochy, F. & van der Rijt, J. (2008). Students' assessment preferences, perceptions of assessment and their relationships to study results. *High Educ, 56* (pp. 645–658).

van der Vleuten, C. & Schuwirth, L.W.T. (2005). Assessing professional competence: from methods to programmes. *Medical Education (39)* (pp. 309–317).

Waagene, E., Larsen, E., Vaagland, K. & Federici, R.A. (2018). *Spørsmål til Skole-Norge høsten 2017: Analyser og resultater fra Utdanningsdirektoratets spørreundersøkelse til skoler og skoleeiere*. NIFU.

Wass, V., Van der Vleuten, C., Shatzer, J. & Jones, R. (2001). Assessment of clinical competence. *The Lancet (357)* (pp. 945–949).

Wiliam, D. (1996). National Curriculum Assessments and Programmes of Study: validity and impact. *British Educational Research Journal, 22(1)* (pp. 129–141).

Wilson, M. (2005). Constructing Measures: An Item Response Modeling Approach, Volume 2. Lawrence Erlbaum Associates.

Wollscheid, S., Hjetland, H.N., Rogde, K. & Skjelbred, S.V. (2018). *Årsaker til og tiltak mot kjønnsforskjeller i skoleprestasjoner. En kunnskapsoversikt*. NIFU.