Inland Norway University

Faculty of Educational Sciences

Tale Gabriella Vesterlid

Master's Thesis

Assessing oral English in secondary school: challenges for novice English teachers.

Lektorutdanning i engelsk

2019

Consent to lending by University College Library	YES ☒	NO ☐
Consent to accessibility in digital archive Brage	YES ☒	NO ☐

Acknowledgements

When working on my master's thesis, my thought process has been similar to what Charles Bukowski once wrote: *"what matters most is how well you walk through the fire"*. This year, my thesis has been the biggest source to my frustrations but it is also a confirmation that I can overcome challenges and learn from it. It is important to note that I could not have done it alone:

First, I want to thank the people who made it possible to write this thesis: the teachers who took part in this study.

I want to thank my supervisor Juliet Munden. Thank you being available in moments of despair, for encouraging me and finding ways to motivate me. Thank you for all your detailed comments and suggestions – it is impressive how knowledgeable you are.

A big thanks to my family and friends for all your support, and for helping me take my mind off the thesis once in a while.

Hamar, May 2019

Tale Gabriella Vesterlid

Table of contents

List of Figures

List of Tables

Abstract

This study investigates how novice teachers of English, at lower and upper secondary level, work with the assessment of oral English. In addition, the study focuses on how the novice teachers perceive their own competence in assessing speaking skills, and how their teacher education has prepared them for assessing oral English.

The research findings show that the teachers find it challenging to assess oral English. According to the teachers, there is a lack of a shared understanding of how to assess language abilities. The participants report that the assessment is largely based on the teachers' subjective opinions. The participants problematize vague and broad competence aims in the curricula that open up for a range of different interpretations. The participants are concerned that the lack of a shared understanding of assessment in oral English can lead to unfair assessment of the students.

The teachers report that the training they received during the teacher education was not sufficient for working with assessing oral English. They express uncertainty regarding interpretations of competence aims and criteria, and are generally uncertain of whether or not their assessments are reliable and valid.

Keywords

Novice teachers, secondary school, language assessment, oral English, rating scales, reliability, validity, subjective assessment

Norsk sammendrag

Denne studien undersøker hvordan nyutdannede engelsklærere i ungdomskolen og videregående skole jobber med muntlig vurdering i engelsk. I tillegg fokuserer studien på hvordan lærerne anser egen vurderingskompetanse, og hvordan utdanningen deres har forberedt dem til å vurdere muntlig engelsk.

Forskningsfunnene viser at lærerne finner vurdering av muntlig engelsk utfordrende. Lærerne oppfatter at det ikke er noen felles forståelse av hvordan man skal vurdere muntlig engelsk på tvers av skoler i Norge og at vurdering i stor grad er basert på læreres subjektive meninger. Lærerne trekker frem at kompetansemålene i læreplanen er lite konkrete, og at dette åpner for mange ulike tolkninger av kompetansemålene. Urettferdig vurderingspraksis blir trukket frem som den ytterste konsekvensen av manglende felles forståelse for vurdering av muntlig engelsk.

Lærerne rapporterer at utdanningen deres i svært liten grad har forberedt dem til å vurdere muntlig engelsk. De forteller om usikkerhet rundt tolkning av kompetansemål og kriterier for kjennetegn på måloppnåelse, og er generelt usikre på om deres vurderinger pålitelige.

Nøkkelord

Nyutdannede lærere, ungdomsskole, videregående skole, vurdering, muntlig engelsk, reliabilitet, validitet, subjektiv vurdering

1. Introduction

In this thesis I will discuss assessment of oral English in lower and upper secondary school. Speaking skills are an important part of the curriculum in language teaching, thus it is an important objective to assess in an accurate, just and appropriate way (Luoma, 2004, p. 1). However, there are many factors that influence how speaking skills are assessed, and one factor is the teacher. Studies indicate that variances in how a speaking score is reached is not irrespective of the rater. In an OECD report by Nusche, Maxwell and Shewbridge (2011, p. 52), it is stated that in Norway there does not seem to be a shared understanding of what constitutes the competencies required to receive an adequate, good and excellent performance in the different subject areas. A lack of a shared understanding can lead to unfair assessment of the students. Thus, how teachers work with oral skills and the assessment of oral English is of importance. The present study indicates that how the teacher interprets the curriculum, grading criteria, and constructs connected to speech differs and that this can affect the scoring of a student.

1.1 Research aim and purpose

The present study seeks to investigate novice English teachers' thoughts about assessing oral English, how they work with oral assessment in the English subject, and how their education has prepared them for this work. The current national Norwegian curriculum, the Knowledge Promotion (LK06), leaves much of the operationalization up to the local schools and individual teachers. Seeing as teachers working in Norwegian schools have much autonomy in their profession, how they interpret the curriculum is of interest. The assessment of students should be reliable and valid – a shared understanding of what to assess is necessary to secure this. Thus, the present study sets out to find if there is a shared understanding of what to assess. To explore this, the novice teachers participating in the study provide information about how they work with assessing oral English: what and how they assess, how they operationalize the competence aims, and challenges and issues related to assessing oral English.

The novice teachers in this study have recently finished their teacher education, an education that to some extent should prepare them for the teacher profession. The present study aims to

find out if/how their education has prepared them for assessing oral English and how they perceive their competence as raters.

1.2 Thesis structure

The present chapter presents theory and terminology necessary to understand assessment practice in Norwegian schools, and therefore, information from relevant official educational documents is included. In Chapter 2, previous research on rater behavior, cognition, beliefs and practices, is accounted for and connected to the issues of reliability and validity in assessment. Chapter 3 outlines how the research has been conducted: the method used to sample information is presented, and possible issues and limitations discussed. In Chapter 4, I present and analyze the research findings and connect the findings to the theory and previous research. The research findings and emerging issues will be discussed further in Chapter 5. Finally, I will make my concluding remarks in Chapter 6.

1.3 Formative and summative assessment

Classroom assessment based on teacher judgement has long been the primary form of assessment in Norway (Nusche et al., 2011, p. 56). This is why it is important to study what the teachers base their judgements on: how they think, what they believe and how they carry out their practice. Teachers working in Norwegian schools have to attend to two concepts when assessing: summative assessment, which is assessment of learning, and formative assessment, which is assessment for learning (Munden & Sandhaug, 2017, p. 122). The present study investigates both formative and summative assessment but the main focus is on summative assessment. The reason being that during the interviews, it became evident that the participants mostly understood assessment as grading and summative assessment, and not as feedback and formative assessment. It is important to note that the teachers do not directly comment upon the two forms of assessment but rather they present issues concerning assessment.

Summative assessment is the overall achievement grade in the subject. The overall achievement grade should be based on a broad foundation of assessment, as it is meant to indicate the student's overall competence in the subject, also it is emphasized that the student's effort should not affect the grade (Norwegian Directorate for Education and Training, 2015, p. 13). In

Norway, students receive an overall achievement grade in the English subject after year 10, and after finishing the first or second year in upper secondary school (Norwegian Directorate for Education and Training, 2015, p. 14).

According to Bøhn (2016, p. 7), summative assessment in upper secondary school is usually based on various forms of classroom assessment, and the assessment is given in the form of overall achievement grades decided by the subject teacher. The same is true for lower secondary. The marks awarded are decisive for further education: students at lower secondary can apply to upper secondary schools, and students at upper secondary can apply to higher education. Thus, the different forms of summative assessment must be regarded as high-stakes (Bøhn, 2016, p. 7), and seeing as the summative assessment is primarily decided by the teacher, it calls for a shared understanding of what to assess.

In recent years, formative assessment has gained increasing prominence in both policy and practice in Norway (Nusche et al., 2011, p. 50). Summative and formative assessment can be carried out in the same way but the intention behind the assessment is different (Dysthe, 2008, p. 17). Formative assessment intends to promote the students' learning (Norwegian Directorate for Education and Training, 2015, p. 1), and it is used to gather information that can better the teaching and guidance of the students (Dysthe, 2008, p. 17). Four principles are central to achieve assessment for learning:

1. The student has to understand what is expected.
2. The feedback given should provide information about the quality of the student's work or performance.
3. The student should be given advice on how to improve.
4. The student should be involved in their own learning process and in self-assessment.

(Norwegian Directorate for Education and Training, 2015, p. 1-2).

Traditionally, teachers have not been trained in formative assessment but with the reformed teacher education implemented in 2010, it is one of the competences that graduating teachers are expected to have (Nusche et al., 2011, p. 50-51). Formative assessment should be covered as part of the subject of didactics and be embedded into the different subjects in teacher education (Nusche et al., 2011, p. 50-51).

1.4 The Norwegian national curriculum

In order to understand how teachers work with assessment of oral English, it is necessary to have knowledge about the frameworks in the Norwegian educational system. The Norwegian national curriculum, the Knowledge Promotion, and the English subject curriculum are legally binding mandates for teachers, schools and local authorities. The current national curriculum was brought into Norwegian schools in 2006 and revised in 2013.[1] It covers 10 years of mandatory schooling, in addition to upper secondary education, and provides curricula for each subject. The Knowledge promotion does not present detailed plans telling teachers what to teach, rather it is goal-driven: it provides goals for all the subjects but how to reach them is up to the teacher (Munden & Sandhaug, 2017, p. 49).

The national curriculum describes the competence that the students are to achieve in each subject, thus, the understanding of the term competence is of importance (Norwegian Directorate for Education and Training, 2016, p. 1). The Knowledge Promotion understands the term competence in the following way: "competence is the ability to solve and master complex challenges. The students show competence in specific situations by using knowledge and skills to solve the tasks at hand" (Norwegian Directorate for Education and Training, 2016, p. 1, own translation). To what extent the student reaches the competence is, partially, up to the teacher, hence how the teachers interpret the definition of the different levels of competence is important.

Even though the Knowledge Promotion gives teachers a lot of freedom and responsibility, lack of specificity in the national curriculum and English subject curriculum, leaves much interpretation to the local schools and individual teachers. Knowing that assessment is largely based on teacher judgement, it is important to have knowledge about how teachers interpret the curricula as their interpretations about what to teach and how to assess may differ.

[1] The national curriculum is under revision for the second time and the changes are predicted to be implemented in year 2020 (Norwegian Directorate for Education and Research, 2019).

1.4.1 Operationalization of the curriculum

Accoring to the Ministry of Education and Research (2004, p. 40), the competence aims in the curriculum must be formulated in a way that makes it possible to assess the students using the aims as reference, and it is necessary that schools develop assessment criteria locally. At the same time, all assessment with grades should be based on standards and be measurable (Ministry of Education and Research, 2004, p. 39). The operationalization of the competence aims involves formulating criteria for assessment, which is done by the local school and the individual teacher. Developing local learning objectives can be beneficial to reach the competence expressed in the competence aims (Norwegian Directorate for Education and Training, 2016, p. 4). Munden and Sandhaug (2017, p. 51) argue that the competence aims are not intended to communicate directly with pupils, and that teachers have to break the aims down into more specific learning objectives for the pupils to work with. However, it is emphasized that even though dividing the competence aims into learning objectives can be constructive, it is the competence aims the students are going to work towards and be assessed in (Norwegian Directorate for Education and Training, 2016, p. 2).

Even though the development of the assessment criteria is largely left to the local level, the Regulations to the Education Act (§3-4, 2009) does equip the teachers with general definitions of different levels of achievement. In lower and upper secondary school, grades ranging from one to six is used to set the overall achievement grade in the subject (The Regulations to the Education Act §3-4, 2009). What constitutes the different grades is described in few details:

a) Grade 6 expresses that the student has excellent competence in the subject.
b) Grade 5 expresses that the student has very good competence in the subject
c) Grade 4 expresses that the student has good competence in the subject.
d) Grade 3 expresses that the student has relatively good competence in the subject.
e) Grade 2 expresses that the student has low competence in the subject
f) Grade 1 expresses that the student has very low competence in the subject.

(own translation) (The Regulations to the Education Act §3-4, 2009).

To illustrate the challenge with these definitions of the grades: achieving the grade 3 means that the student has “relatively good competence in the subject”. However, what constitutes

relatively good degree of competence is up to the local school or individual teacher, and thus how the schools and teachers interpret the descriptions is an issue of concern.

It is not only the general definitions of different levels of achievement that lacks specificity. The competence aims are not perceived by teachers as specific enough to guide teaching and assessment: there is a lack of clear statements of learning goals and expectations (Nusche et al., 2011, p. 52). As pointed out in the Knowledge Promotion, local operationalization of the competence aims is required. The intermediate learning goals and the more specific teaching content, methods and grading criteria are expected to be developed at the local level (Nusche et al., 2011, p. 52). When criteria are developed at local and individual level, it is reasonable to assume that there may be local differences in what content is taught, what methods the teachers use and the grading criteria they develop. Nusche et al. (2011, p. 52) comment upon the localized nature of Norwegian education and states that: the broad competence aims are meant to give teachers ownership in establishing their teaching program. However, this is challenging as many teachers find it difficult to make concrete lesson plans, objectives and assessment activities based on the broad competence aims (Nusche et al., 2011, p. 52).

The tension between the nationally developed competence aims and the locally developed learning objectives and/or criteria illustrates a paradox in the Knowledge Promotion. Johannessen (2018, p. 22) captures the paradox in one simple sentence: “the competence aims are to be general and possible to assess at the same time”. Dividing the competence aims into smaller learning objectives should make it clearer for the students what is expected of them. However, learning objectives should not be too detailed as this can lead to a loss of the connection to the competence aims in the curriculum (Hartberg, Dobson & Gran, 2012, p. 31). At the same time, Bøhn (2016, p. 3) states that criteria designed and implemented on the local level can be legitimate for assessment purposes, seeing as the main intention of assessment is to *promote* learning, a process where the teacher has a prominent role. However, much of the assessment is used to measure learning and therefore, a shared understanding of assessment is necessary for accurate measuring.

1.4.2 The lack of a shared understanding

The localized nature of the curriculum mixed with little specificity, leads to assessment that is largely based on teacher judgement. This calls for a shared understanding of how to assess between schools and teachers. Nusche et al. (2011, p. 52) writes that the reference points and criteria for assessment need further clarification, and refers to the OECD review team who argue that clearer rubrics that detail assessment criteria would be beneficial for assessment. The need for clarification became evident as there did not seem to be a shared understanding regarding the competencies required to receive an adequate, good and excellent performance in the different subject areas, and the potentially resulting unfairness in teacher grading of students because of the lack of a shared understanding (Nusche et al., 2011, p. 52).

The lack of a shared understanding raises issues about the consistency and fairness of student assessment, reporting and grading: Norwegian research indicates that there are large variations in the ways teachers set overall achievement marks (Nusche et al., 2011, p. 53-54). Nusche et al. (2011, p. 54) state that there is no guarantee that teachers discuss grading criteria within or across schools. In relation to the ethical dimension of language testing, Weir (2005, p. 1) writes that the scores given affect people's lives, and therefore getting it right and ensuring test fairness is a necessity. A lack of shared understanding among teachers can threaten the fairness of assessment.

1.5 The English subject curriculum

So far I have presented general information about assessment in Norwegian secondary education while the following will be subject-specific. I will The English subject also faces the challenges presented above: the competence aims are vague and the operationalization is largely left to the local school or individual teacher, which might contribute to differences in assessment. The subject curriculum is a document that the teachers can use to plan their teaching and to assess their students. Students are assessed in the competence aims stated in the English subject curriculum (Ministry of Education and Research, 2004, p. 40).

A subject curriculum is provided in both lower and upper secondary school. In addition to the subject curriculum, lower secondary schools have a national common frame of reference for

assessing oral English. This common frame of reference is arranged in rubrics and consists of criteria for the different levels of achievement, ranging from grade two to six (Norwegian Directorate for Education and Training, 2017, p. 3). The criteria is developed from the competence aims, and are meant to be a guide when setting the students' final grade in year 10, also the criteria are meant to serve as a national common guide for how to assess (Norwegian Directorate for Education and Training, 2017, p. 1).

Teacher collaboration about the competence aims and the criteria should contribute to a common understanding about what the students are supposed to learn, and what constitutes the different levels of achievement in a subject, which again is meant to lead to fair assessment of the students (Norwegian Directorate for Education and Training, 2017, p. 1). Despite the need for a nationally developed frame of reference on lower secondary level, this is not provided in the mandatory English subject in upper secondary.

1.5.1 Relevant competence aims

The English subject curriculum is structured into four main subject areas that supplement each other and thus should be considered together: language learning, oral communication, written communication, and culture, society and literature (Norwegian Directorate for Education and Research, 2013, p. 2-3). Under each of the four main subject areas a list of competence aims are grouped together. The competence aims specify what students are expected to master at the end of instruction at different levels.

The four main areas are the same in the English subject in lower secondary school and for the mandatory English subject in upper secondary secondary. Even though the main areas are the same, the competence aims differ. For the purpose of the present study, I consider the main areas language learning, oral communication, and culture, society and literature as relevant to focus on because these areas address different aspects of language learning, how to communicate and the content that is to be communicated. Under the main areas, 20 competence aims are listed for lower secondary level, and 18 for upper secondary level. Seeing as the lists of competence aims are long, it is likely that the competence aims in focus vary depending on the oral assessment situation. Instead of presenting all of the competence aims, in the following

paragraphs I will present the three main areas identified as relevant for assessing oral English in both lower and upper secondary level.

The focus in the main area language learning, is to know about different aspects of learning a new language, and make connections between English, one's native language and other languages (Norwegian Directorate for Education and Training, 2013, p. 3). In addition to knowledge about the English language, knowledge about one's own language learning is important, which is why self-assessment is emphasized as a useful skill. This includes being able to assess one's use of the language, own learning needs, and to choose suitable strategies and methods to learn and use the English language (Norwegian Directorate for Education and Training, 2013, p. 3).

The main area about oral communication targets a wide range of skills. The students should be able to use suitable strategies for communication: to listen, speak and converse using the English language (Norwegian Directorate for Education and Training, 2013, p. 3). Developing vocabulary, using idiomatic structures and grammatical patterns, and learning to speak clearly and to use the correct intonation, are listed under oral communication (Norwegian Directorate for Education and Training, 2013, p. 3). However, the variety of the English language is not specified, thus, what is perceived as correct intonation may vary depending on the teacher. The students are expected to be able to use the English language in different situations: to adjust the language to the purpose and the recipient, which includes the skill to distinguish between formal and informal oral language (Norwegian Directorate for Education and Training, 2013, p. 3). Also, different media and resources, and developing language skills relating to different subject areas are listed under oral communication (Norwegian Directorate for Education and Training, 2013, p. 3).

The main area culture, society and literature focuses on cultural understanding, mainly in English-speaking countries (Norwegian Directorate for Education and Training, 2013, p. 3). A main concern is to have knowledge about the usage of the English language as a world language, in addition to promoting understanding and respect for other cultures (Norwegian Directorate for Education and Training, 2013, p. 3-4).

The main areas present broad competence aims in both lower and upper secondary level, and it is up to the local school and/or individual teachers to interpret what information is important to

include and how to work with different topics to reach the aims. One competence aim in lower secondary is that the students are supposed to "discuss and elaborate on different types of English literature form English speaking countries" (Norwegian Directorate for Education and Training, 2013, s. 9). There is a similar competence aim in the mandatory English subject in upper secondary: "discuss and elaborate on different types of English language literary texts from different parts of the world" (Norwegian Directorate for Education and Training, 2013, p. 11). The two competence aims referred to both include the phrase "discuss and elaborate on", however, what the students are supposed to discuss and elaborate on is not clearly specified, which is in alignment with the ideal of teacher autonomy in the Knowledge Promotion. The downside is that the lack of specificity in the competence aims makes it difficult for the teachers to use them to guide the teaching and assessment (Nusche et al., 2011, p. 52). The broad competence aims open up for discrepancies in operationalization between schools and the individual teachers. Different interpretations might be problematic as the understanding of what to assess can vary.

1.5.2 Oral skills: a basic skill

In the Knowledge Promotion, five basic skills fundamental to learning, are defined and implemented in all subjects in compulsory and secondary education (Norwegian Directorate for Education and Training, 2012, p. 5). The priority of oral competence in the Norwegian school is evident as 'oral skills' is listed as one of five basic skills. The English subject curriculum has implemented and defined oral skills the following way:

> *Oral skills in English means being able to listen, speak and interact using the English language. It means evaluating and adapting ways of expression to the purpose of the conversation, the recipient and the situation. This further involves learning about social conventions and customs in English-speaking countries and in international contexts. The development of oral skills in English involves using oral language in gradually using more precise and nuanced language in conversation and in other kinds of oral communication. It also involves listening to, understanding and discussing topics and issues to acquire more specialized knowledge. This also involves being able to understand variations in spoken English from different parts of the world.*

(Norwegian Directorate for Education and Training, 2013, p. 4-5).

How the teachers interpret this definition, will affect the practice of oral English. According the definition above, the students must have a range of oral skills and know how to use them in different situations. Conversations and discussions are explicitly mentioned, which implies that they should be in focus. The students are to have an active role as they must listen, speak and interact using the English language. From the definition of oral skills, the teachers have to provide a variety of settings for the students to practice their oral skills. However, basic skills are not something to be learnt in addition to the competence aims but they are fully integrated in the competence aims (Munden & Sandhaug, 2017, p. 52).

1.6 Defining terminology

As stated in the previous paragraphs, the Norwegian educational system has a localized nature. This implies that there might be a range of different interpretations of the competence aims and how to assess them, which raises the question of 'what' to assess.

What to assess is stated in the competence aims but is not perceived as specific enough to guide teaching and assessment. Lower secondary level has a nationally developed frame of reference consisting of rubrics with criteria for the different levels of achievement. This serves as a guide for what to assess, and seeks to develop a shared understanding to secure reliable and valid assessment. However, no such frame of reference is available in the mandatory English subject in upper secondary. A lack of a common frame of reference and a shared understanding of what to assess may be problematic, especially in oral English: as we will see in the chapter presenting previous research (Chapter 2), speaking is a difficult skill to assess reliably. In attempting to provide clarity about what to assess, one often comes across the terms 'construct' and 'criterion'. In the following, I will explain how these terms are defined in the present study.

1.6.1 'Construct' and 'criterion'

Weir (2005, p. 1) describes constructs as the "underlying abilities we wish to measure in students, the *what* of language testing". Constructs are often based on a frame of reference such

as a course syllabus (Bachman & Palmer, 2010, p. 211). In Norway, the frame of reference is the subject curriculum, which forms the basis for the operationalization (Bøhn, 2015, p. 2). In the present study, the following constructs have been identified as underlying abilities to measure the students' oral competence: communication, fluency, vocabulary, grammar, pronunciation, intonation and content. The constructs are based on those identified in research conducted by Bøhn (2016) and Johannessen (2018).

A competence aim listed under oral communication in lower secondary after year 10 is to "express oneself fluently and coherently, suited to the purpose and situation". Thus, one construct is fluency; still, there is a need for further clarification in order to assess this construct. Fluency is an abstract noun that cannot be directly observed, hence, one has to identify observable indicators of this construct (Fulcher & Davidson, 2007, p. 370). A way to interpret fluency is to operationalize it as Brown, Iwashita and McNamara (2003, p. 23) give examples of in their report from their exploratory study about English-for-academic-purposes, where features such as 'hesitation' and 'fillers' serve as observable indicators for the construct fluency.

To assess whether or not a student has expressed him or herself fluently, the teachers need to identify indicators or key aspects for the construct. The operationalization of the competence aims is largely left to local level, which means that the local schools or individual teachers have to develop criteria for assessment. For the purpose of the present study, the term 'criterion' is best described by Brindley (1991, p. 140), who exemplifies it this way: someone doing an oral interview is given a score based on a rating scale containing key aspects of the performance to be assessed. These key aspects are the criteria (Brindley, 1991, p. 140). Key aspects have, to a certain extent, been identified nationally for the English subject in lower secondary school. Here, teachers can refer to a common document consisting of rubrics with criteria for what constitutes the different levels of achievement in oral English, which functions as a rating form/scale.

To clarify the understanding of constructs and criterion in this thesis, I reserve the right to use the term construct in relation to different aspects in language such as: fluency, grammar, vocabulary, pronunciation, content and communication. Criterion will be used when talking about key aspects of oral communication in the English subject, defined nationally, locally, or individually by the teacher.

1.7 Summary

The Knowledge Promotion and the English subject curriculum are legally binding mandates for teachers, schools and local authorities in Norway. The documents function as a common frame of reference as these are the documents teachers turn to for information about the competence the students are to attain. The curricula consists of broad and vague competence aims that are difficult to use as a guide in teaching and assessment (Nusche et al., 2011). It is unusual to have a competence plan like the Knowledge Promotion without giving guidelines and criteria at the same time (Munden & Sandhaug, 2017, p. 49). However, the idea is that working with the curricula locally, developing plans and creating own criteria for each competence aim, will give teachers ownership of the plan, ensure that they adopt it and put it into practice (Munden & Sandhaug, 2017, p. 49).

The lack of specificity in the competence aims can lead to a range of different understandings of what is important to teach and assess. The national common frame of reference in English in lower secondary, consisting of rubrics with criteria describing the different levels of achievement, is meant to secure a shared understanding of what and how to assess. It functions as a rating form/scale. Such rubrics with criteria is not developed for upper secondary level, where the operationalization of the competence aims are left to the local school and individual teacher. However, general definitions of the different levels of achievement is described by the Regulations to the Education Act (§3, 2009), the definitions have few details which leaves much of the interpretation to the teacher.

The localized nature of the Norwegian education system is problematic as it is challenging to have a shared understanding of assessment when it relies heavily on individual teacher judgement. However, the Knowledge Promotion advocates local school's development of learning objectives and criteria. At the same time, dividing the competence aims into smaller units might lead to a loss of focus on the overarching competence aims that the students are to be assessed in.

2. Literature Review

2.1 Introduction to the chapter

As stated in Chapter 1, teachers working in Norwegian schools have much autonomy: they have to interpret the national curriculum, the subject curriculum, and decide what and how they are going to teach and assess. Thus, how teachers think and behave is of interest. The present chapter will focus on rater cognition, behavior and practices: the concept of teacher cognition is introduced, and previous research on rater behavior presented. As rater behavior affects the reliability and validity of assessment, these terms are accounted for in this chapter.

As will become evident in the presentation of empirical research findings (chapter 4), the participants in the present study mostly understood and talked about assessment as summative. In order to provide theory that can be connected to and discussed in relation to the findings in the present study, the empirical studies presented in the current chapter is concerned with summative assessment of speaking skills.

The chapter focuses on the following objectives:

1. Identify factors that can have an impact on the assessment of oral performances in the English subject.
2. Evaluate critically possible implications.

2.2 Reliable and valid assessment

Two central terms connected to assessing speaking skills is reliability and validity, and thus it is necessary to provide an understanding of these terms before presenting previous research on rater behaviour. Luoma (2004, p. 170) argues that because the rating process involves human raters, which will lead to some variability, special procedures are needed to ensure the reliability and validity of speaking assessment. Reliability is related to the consistency of the scores and validity to the scores' meaningfulness (Luoma, 2004, p. 175). Thus, to ensure reliable and valid assessment in the English subject a shared understanding of what is to be assessed is central.

2.2.1 Reliability in assessment

Fulcher and Davidson (2007, p. 375) define reliability in assessment as the consistency of measurement: the test taker should receive the same score on a test taken several times during a reasonable period of time, and should receive the same score irrespective of whichever rater is used. Any variation in scoring should be due to differences relevant to the construct of interest, not irrelevant factors such as who did the scoring, the particular items used for the day, or whether the student was having a 'good' or a 'bad' day (Black and William, 2012, p. 244). Luoma (2004, p. 178) recognizes the importance of consistency for reliable scores and points out the most common way for providing reliability in assessment: through rating forms. According to Luoma (2004, p. 172) raters are, even when making an overall score, often asked to give detailed information about how he or she arrived at the score. In order to gather the detailed information, a rating form functions as a concrete form that the raters use to record their ratings: it is meant to help the rater compare an examinee's performance to the criteria rather than the other examinees' performances (Luoma, 2004, p. 172).

Luoma (2004, p. 172) advocates rating forms because they help to structure the rating process, speed it up and make it consistent. In addition the forms define what the raters pay attention to during the rating (Luoma, p. 172). Another benefit of using a well-structured rating form is that it can increase the efficiency of feedback sessions and combining the feedback with advice for further learning can make the feedback more useful for students and teachers (Luoma, 2004, p. 175).

In addition to rating forms, Luoma (2004, p. 177) mentions rater training as a way to ensure score reliability. Rater training has been criticized as a form of indoctrination where novice raters are taught to evaluate performances in the system's terms, changing their perception of the world without there being proof that the scoring criteria is valid, however, it is argued that providing evidence about the validity of the criteria will solve this issue (Luoma, 2004, p. 177). Luoma (2004) continues writing that test developers recognize that it is impossible to give comparable ratings without training, and comparability is considered important and so rater training is one way to ensure this.

Subjectivity in assessment is a bigger concern in more informal classroom settings and assessment as the rater has to reflect upon how they assess, attempt to be just and use the

assessment criteria consistently (Luoma, 2004, p. 179). Johannessen (2018, p. 44) addresses what she describes as the complicated nature of reliability in classroom assessment in Norway: the assessment is meant to be used as a tool to enhance instruction and learning but is also used by the teacher to decide a grade in the subject. Seeing as, in nearly all educational systems, admission to higher education is based on the sum of the student's course grades in different subjects (The Norwegian Universities and Colleges Admission Service, 2013), reliability in classroom assessment is crucial as it affects the possibilities and limitations a student has when applying to higher education.

Luoma (2004, p. 179) proposes some examples of how to ensure reliable scoring in the classroom: the rater can focus on concrete features of a performance, identifying strengths and weaknesses and then compare this to the assessment criteria. In addition, the rater can do a self-check. The rater can revisit a rated performance after finishing rating the last performance of a group: this will help the rater to discover if internal standards have changed in the course of the rating work (Luoma, 2004, p. 179).

2.2.2 Validity in assessment

When describing the term validity in regard to assessment, Bøhn (2016, p. 14) refers to it as the quality or 'soundness' of an assessment procedure. Green (2014, p. 75) states that validity is often seen as *the* essential quality of good assessment. The definition of the concept of validity has shifted from whether or not a test measures what it is purports to measure, to the inferences that are made from assessment results (Bøhn, 2016, p. 14). The definition of validity that holds consensus today is the one given in the Standards for Educational and Psychological Testing:

> *Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests … The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself.*

> (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014).

This definition is highly relevant in the present study as the aim is to investigate teachers' interpretations of speaking performances, how they collect evidence to support assessment, what this evidence is and if the evidence is affected by rater bias and/or different interpretations of the criteria being measured. Green (2014, p. 75) emphasizes that the definition above recognizes that there can be no such thing as 'a valid assessment':

> *Validity is not properly thought of as a quality of assessments at all, but is a quality of the interpretations that users make of assessment results: an assessment can only be considered valid for certain purposes.*

The first step in assessment should be to define and agree on what knowledge, skills or abilities are to be assessed, however, knowledge, skills and abilities can not be directly observed and measured but they can be described and they are variable: some people have more of these attributes, others have less (Green, 2014, p. 76). When assessing language ability, how do we know that we have the same understanding of the concept, how can one prove the truth of their claim that one person has a better language ability than another (Green, 2014, p. 76 – 78). In order to achieve similar understandings of what is to be assessed and how to assess it, these topics have to be discussed amongst schools and teachers.

Sandvik (2013, p. 38) writes that there are several types of validity in an educational setting: content-related evidence of validity, criterion-related evidence of validity, and construct-related evidence of validity. These kinds of validity evidence can be used to shed light on the value of assessment results (Green, 2014, p. 78). The three aspects relates to whether or not the content of a test can be said to be representative for a given subject, and if the teacher can draw conclusions about a student's performance based on results from a test (Sandvik, 2013, p. 38). Green (2014, p. 78 - 79) writes that, when it comes to content-validity, tasks should be carefully chosen to ensure that a sufficient range of material is included, and content-validity should be carried out before an assessment is put into operation. Criterion-related validity, on the other hand, cannot be collected before the assessment and refers to the results of the

assessment and some alternative indicator, such as teacher judgements or results from another assessment of recognized validity (Green, 2014, p. 79). Green (2014, p. 81) argues that it is fundamental that everybody involved in interpreting assessment results shares at least a basic understanding of the constructs involved. Construct-validity is seen as embracing all forms of validity evidence, as this is the evidence that the theory underpinning the assessment provides a sound basis for the decision (Green, 2014, p. 81).

Sandvik (2013, p. 39) recognizes that outlining validity in this way has its limitations: it does not consider technical aspects of the test, nor the context and intention of the assessment, or the effect it has on student learning and motivation. The teachers' education, experiences, and how they view learning is all part on the assessment context, and affects how evidence of learning is collected and interpreted (Sandvik, 2013, p. 39). Sandvik (2013, p. 41) proposes looking at validity as a chain of interpretations that attempts to provide meaning to the aims in the curriculum.

As stated in section 2.2, ratings that involve human raters will lead to some variability. In the following section, the concept of teacher cognition will be presented and connected to variability in rater behavior. Thereafter, previous research on variability in rater behavior is presented.

2.3 Teacher cognition

The present study seeks to investigate how novice teachers of English work with oral assessment: teacher cognition is relevant in this respect because what the teachers think, know and believe will affect their classroom practices. In a Norwegian setting, where the subject curriculum is broad and open for interpretation, a study of teacher cognition can contribute to explain why teachers have different teaching and assessment practices.

Teacher cognition is concerned with what language teachers think, know and believe – and the relationship to teachers' classroom practices (Borg, 2006, p. 1). Borg (p. 7) refers to a report by the National Institute of Education (1975) as the start of a tradition of studying teacher cognition, the report argued that:

> *It is obvious that what teachers do is directed in no small measure by what they think (...) If, however, teaching is done and, in all likelihood, will continue to be done by human teachers, the question of relationships between thought and action becomes crucial.*
>
> (National Institute of Education, 1975).

The studies on teacher cognition recognize that teachers play a central role in shaping classroom events: their knowledge and beliefs affects how they act, thus understanding teacher cognition is central to the process of understanding teaching (Borg, p. 1). Borg (p. 283) argues that language teachers have cognitions about all aspects of their work and that these can be described using various psychological constructs. In the figure below, Borg (p. 283) illustrates these psychological constructs and the relationships among teacher cognition, teacher learning and classroom practice:

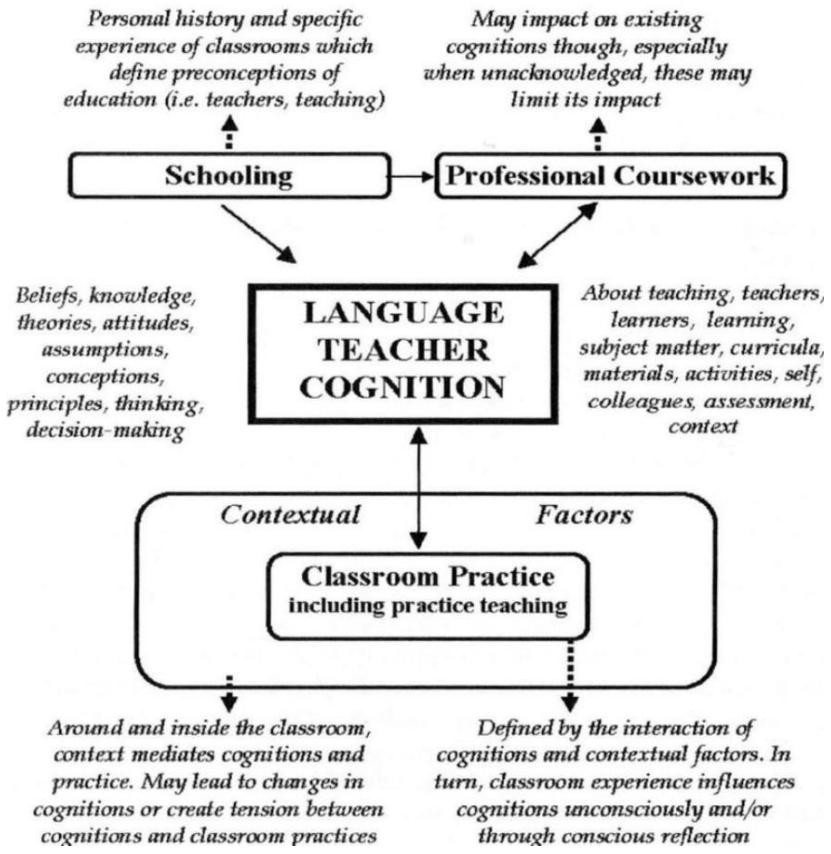

Figure 1: Elements and processes in language teacher

As the figure shows, the teacher's schooling affect the teacher cognition. Borg (p. 283 – 284) states that teachers' experience as learners is part of informing cognitions about teaching and learning, these cognitions may continue to influence teachers throughout their careers. In addition, professional coursework, and contextual factors, affect the teacher cognition, and vice versa. The teachers' experiences, and their beliefs, attitudes and assumptions about teaching, teachers, learning, learners and assessment will vary depending on the individual teacher. Therefore, rater behavior and variability in assessment can, to some extent, be explained by the concept of teacher cognition.

In relation to the cognition and practices of novice teachers, Borg (p. 81) writes that the transfer of knowledge and beliefs from teacher education to classroom practice is not linear. Contextual factors such as professional relationships with colleagues and immediate concerns with managing learners, influence the cognitions and practices of novice teachers, and may outweigh principles learned during teacher education (Borg, p. 81). Practicing language teachers hold cognitions about all aspects of their work, the cognitions are shaped by the teachers' lived experiences, such as their education (Borg, p. 107). Changes in the cognitions happen over time, and Borg (p. 107) argues that there are differences in the thinking and knowledge of more and less experienced and expert teachers. When Borg provides indications that there are differences in teacher cognition based on the teachers' experience, this implies that the teacher's experience will affect the assessment of the student. In high-stakes assessment, such as oral exams and setting the overall achievement grade in the subject, one can assume that teacher cognition can affect the reliability and validity of the assessment.

2.4 Rater behavior

In his doctoral thesis 'What is to be assessed? Teachers' understanding of constructs in an oral English examination in Norway' (2016), Bøhn presents an article (2015) aiming to explore how EFL teachers in Norway understand the constructs to be tested in an oral English exam at upper secondary level, where no common rating scale has been provided.

Bøhn (2015) comments upon the difference between the written exam and oral exam in English in Norwegian upper secondary schools. While the Directorate for Education and Training administers the written exam, the oral exam is left to local educational authorities and in many

cases the individual schools (Bøhn, p. 3). Consequently, the oral exam lacks the standardization that the written exam has with the common exam format, common exam tasks, and a common written rating scale (Bøhn, p. 3).

When studying how teachers score exam performances, Bøhn (p. 8) found variability in scoring behavior: the teachers vary in their understanding of the constructs and criteria to be tested, and what kind of criteria they view as salient. Bøhn (p. 2) points out that studies investigating assessment practices more generally found that even though criterion-referenced assessment is required by the Education Act, teachers might find such assessment difficult. This may be one explanation for why the teachers' rating behavior differed.

Further, the study found that the teachers at large focus on the same overall features of performance with some variation in the way they attend to more narrow features (Bøhn, p. 8). Also, Bøhn (p. 9) found that raters apply non-criterion relevant information when scoring performance, such as effort. There were indications that the teachers rate the students attending the vocational study program (VSP), especially the weaker students who risk failing, more leniently than they do students on the general study program (GSP) (Bøhn, p. 6). Bøhn (p. 6) argues that this shows that the teachers may give credit to students who "try their best" in order to compensate for lack of language or content knowledge. At the same time, Bøhn (p. 6) emphasizes that there was a balance to be seen as some of the other VSP teachers take the opposite stance: denying that they would give extra credit for effort as they are not allowed to do so. The teachers in the study reported that they score performance holistically (Bøhn, p. 4).

Bøhn (p. 9) concludes that the study points to the problem of not having a common rating scale as a common rating scale is believed to strengthen the validity of the score interpretations. Another implication of the study is the problem of introducing comprehensive content construct at the intermediate to upper-intermediate level, as the results indicate that teachers working with students with lower proficiency in the subject pay less attention to the content construct (Bøhn, p. 9). Finally, as the examiners in the oral exam seem to focus on the students' ability to reflect on content, the study points to the importance of including topical knowledge in classroom practices to prepare the students for oral examination (Bøhn, p. 9).

Variations in rater behavior are also evident in Ang-Aw and Goh's (2011) study. The study, aiming to examine rater behavior at high-stakes 'O' Level oral examinations in Singapore, was

conducted using Concurrent Verbal Protocols, questionnaires and scores (Ang-Aw and Goh, p. 33). Their findings show that exam scores were qualitatively determined to be due to four differences: emphases of factors assessed, constructs of oral proficiency, rater interpretations and approaches in assessment (Ang-Aw & Goh, p. 31). The researchers concluded that even when given similar training, assessment was not entirely reliable: the raters followed the marking scheme to varying extent, they were preoccupied with different aspects of the candidates' performances, and assessed in a dissimilar way (Ang-Aw & Goh, p. 44). Thus, the researchers pointed out the challenge of subjectivity in assessment:

> *Language assessment is a complex process where the raters are often required to carry out subjective assessment of a persons' language ability.*
>
> (Ang-Aw & Goh, p. 31).

Ang-Aw & Goh (2011) believe that their findings suggest that the raters' behavior threatens the validity and reliability of the assessment. Firstly, the raters focused on non-criterion factors, which implies that they look at different aspects when they judge the performances and do not stick strictly to the marking scheme (Ang-Aw & Goh, p. 37). Secondly, the raters operationalized the construct of oral proficiency differently, which Ang-Aw and Goh (p. 38) consider to have serious implications on the overall reliability of a wide-scale oral examination. In addition, the study reports that raters have different interpretations of candidates' performances and the candidates' scores (Ang-Aw & Goh, p. 38-39).

> *Raters may award different scores to the same performance or the same score to different performances.*
>
> (Ang-Aw & Goh, p. 39).

A reason why one might find such differences in the assessment of oral performance is that raters assess with different levels of severity/leniency (Ang-Aw & Goh, p. 42). Three types of raters were identified in this study: a consistently lenient rater, a consistently strict rater and a rater who was lenient towards the stronger candidate but strict towards the weaker candidate (Ang-Aw & Goh, p. 42).

Even though the exam was a criterion-references test, all raters except one, compared the candidates' performances: one of the raters even adjusted the marks after comparing the performance of two candidates (Ang-Aw & Goh, p. 43). Ang-Aw and Goh (p. 43) believe that such inter-candidate comparisons can change the test from criterion-references to a norm references test where the candidates' scores would be affected by the relative oral proficiency of the group taking the test. According to Ang-Aw and Goh (p. 43-44), the ambiguity of the descriptors as a whole might cause the inter-candidate comparison, and when raters are not able to clearly match the candidates' performances to the descriptors, they can end up feeling that a particular score is too high or too low for a candidate.

From the results, Ang-Aw and Goh (p. 44) advocate that raters need to assess characteristics of candidates' performances that correspond to the descriptors in the marking scheme, they need to be clear about the aspects of candidates' performances that should be in focus and which aspects should not. In addition, rater's feedback should be elicited and qualitative feedback should be given to them during training, and lastly, for raters identified as more unreliable during training sessions, statistical adjustments for rater characteristics or double ratings could be carried out (Ang-Aw & Goh, p. 44).

Another indication that there is variability in rater behavior is Kim's (2015, p. 239) study, which point out that rater effects can be a potential source for variance that interferes with the accurate measurement of examinee language ability. Kim (p. 242) looked at three groups of raters with different backgrounds to study how they used a given analytic scoring rubric while rating speaking performance. The participants were pre- and in-service language teachers divided into groups based on their experience and expertise in rating L2 speaking assessment (Kim, p. 241-243). This resulted in three groups: novice (entirely new), developing (some background: two or three years), and experiences raters (over five years), all of the raters either were native speakers of English or had a fluency that was native-like (Kim, p. 241-243). Kim (p. 242) aimed to find out how the different rater groups interpret the rating scales and the performance level descriptors: they had three sessions where they rated performance, and the study describes how the raters change their interpretation of the rating scales and performance level descriptors over time.

Kim (p. 248) found that the three groups had different levels of understanding in regard to the rating scale and descriptors: novice raters often confused the rating scales, the developing raters also misunderstood parts of the scale but less frequently, as opposed to the experienced raters who generally understood the rating scale correctly. Kim (2015, p. 249) argue that one of the reasons why the novice teachers might have difficulty is their lack of experience in assessing speech and their limited understanding of language concepts.

In between the sessions of rating, the groups received training, and the novice raters showed improvement in their understanding of the scale (Kim, p. 250). The developing rating group seemed to have even better rater training effect than the novice group, while in the case of the experienced raters there was no major difference in their scoring behavior (Kim, p. 252). All three groups displayed different levels of rating performance during each rating session (Kim, p. 254). Kim (p. 256) argue that the results suggest that the raters' background should be taken into consideration: group level training would make it possible to differentiate rater training which will lead to more dependable raters and, more likely, provide reliable ratings.

Orr's (2001) article also supports the claim about variation in rater perceptions. Orr (p. 152) describes Fist Certificate in English (FCE) raters' thoughts while assessing oral performance. The findings show that raters did not heed the same aspects of the assessment criteria, and that they paid attention to a range of non-criterion relevant information (Orr, p. 143). Consequently, Orr (p. 143) says that the raters provided a range of scores, and in addition the raters giving the same score perceived the performance differently.

Orr (p. 143) argues that with the varied nature of raters' perceptions it would be impossible to say how any one speaking score had been reached, which has implications for the validity of the raters' interpretations of the performance.

> *In many oral performance tests, these scores are reached subjectively using rating scale descriptors to guide the examiner towards a number.*
>
> (Orr, p. 143).

Orr (p. 151) reports that especially three aspects of the performances were commented on: the candidate's presentation of her/himself (effort, body language, preparedness for the test), how

the candidate compared with another learner, and the global impression of the performance. Seeing as there were a number of comments about these aspects, Orr (p. 151) does not seem to think that they are incidental but rather an integral part of the raters' thought processes. Orr (p. 151) emphasizes that the point of having a rating scale for assessing language abilities is that the candidate's performance is compared with the scale and not, as in this study, with other learners' performances. The raters do, however, report that comparisons between learners was necessary, as the scale was unclear (Orr, p. 151). According to Orr (p. 153) many raters show difficulty adhering to the assessment criteria and have difficulty understanding the model of communicative language ability that the rating scales are based on.

In his concluding remarks, Orr (p. 152-153) argues that one must interpret the FCE Speaking test scores with caution. Orr (p. 152 – 153) refers to McNamara (1996) to support his conclusion about the importance of being skeptical to the meaning of test scores, and improve the understanding of scores to increase the fairness to the test candidates. Orr (p. 153) proposes that raters should see examples of the process of how expert judges reach and justify their scores, and that the usability and usefulness of the Speaking test rating scales should be questioned.

2.5 Summary

From the research presented in this chapter, and in relation to the first objective mentioned above, it becomes evident that variability in the raters' scoring behavior affects the assessment of oral performances. The variability in scoring behavior is identified as the overarching challenge, with the following sub-challenges:

1. Challenges related to the lack of a shared understanding of what to assess.
2. Subjectivity in assessment.
3. The effects teacher cognition has on assessment.

These factors are central to objective number two: evaluate critically these possible implications, as the factors have implications on the reliability and validity of oral assessment.

With basis in the research presented in this chapter, I argue that raters lack a shared understanding of what to assess in oral English. Several empirical studies have shown variability in the raters' scoring behavior: that they include non-criterion relevant information when assessing, use inter-candidate comparison, and that they report difficulties using rating forms/scales provided for the assessment.

Inter-candidate comparison was found in Orr's (2001) research as well as in Ang-Aw and Goh's (2011) research. Ang-Aw and Goh's (p. 43) research showed that inter-candidate comparison even lead to adjusting one candidate's mark. Orr (p. 151) mention that the raters believed the comparisons to be necessary, as the rating scale was unclear. Bøhn's (p. 8) study also show tendencies that the teachers find criterion-referenced assessment difficult. Both Orr (2001) and Ang-Aw and Goh (2011) report that a rating scale to assess language abilities is useful to reduce inter-candidate comparison.

There is no common frame of reference to assess oral English in upper secondary school in Norway, however, there is one for assessing written English. The operationalization of the constructs happen at local level, and in many cases it is up to the individual teacher. Thus, the operationalization of competence aims and what constructs and criteria are salient to assess, will largely be based on the individual teacher's opinions.

The subjectivity in oral assessment has been pointed out by Ang-Aw and Goh (p. 44), who state that when assessing language, the raters are often required to carry out subjective assessment. Ang-Aw and Goh (p. 42) problematize subjectivity even more as they report that they identified three types of raters: a consistently lenient rater, a consistently strict rater and a rater who was lenient towards the stronger candidate but strict towards the weaker candidate. Bøhn's research also indicate differences in severity/leniency among the raters. Bøhn (p. 6) reports differences in how raters assess students with different levels of proficiency: VSP students was rated more leniently than GSP students.

Not only does it seem to be differences in how lenient or strict the raters are. Kim's (p. 248) research show that the novice raters confuse the rating scales more often than the developing and experienced raters do. Kim (p. 256) argue that the rater's background should be taken into consideration and that group level training can lead to more dependable raters and reliable scores.

The research point out the challenges with raters attending to non-criterion relevant information when assessing, inter-candidate comparison, difficulty using the rating scales, no common rating scale available, differences in severity/leniency depending on the individual teacher and the student being assessed, and differences in assessment based on rater background. The research presented mention these factors when they give possible reasons for why there is variability in the raters' scoring behavior. The variability in assessment affects the reliability and validity of assessment.

To ensure reliable and valid assessment of oral English, a shared understanding of what is to be assessed is necessary. Black and William (2012, p. 244) point out that there should not be any variation in scoring due to irrelevant factors, but as the research presented in this chapter show: non-relevant information is included in the assessment of oral English. Black and William (2012, p. 244) continue elaborating on irrelevant factors saying that factors such as who did the scoring should not be a source of variation in scoring, however, the research presented shows that what the raters view as important to assess vary, and that the raters' subjective opinion plays a part in the scoring of performances. Fulcher and Davidson (2007, p. 375) point out that the test taker should receive the same score irrespective of whichever rater is used. When looking at the research provided in this chapter, there is clear evidence that consistency of a score irrespective of the rater is not certain. Sandvik (2013, p. 39) argues that the teachers' education, experiences, and how they view learning is all part on the assessment context, and that this affects how evidence of learning is collected and interpreted, which supports the reporting of subjectivity in assessment.

3. Method

3.1 Introduction to the chapter

In the present chapter I outline the chosen research design and how it has contributed to answering the research questions. I also present the different phases of the research process, including how the material was analyzed. In addition, I discuss research validity, and some of the strengths and weaknesses of the research design.

3.2 The phases of the research process

The purpose of this study is to investigate how novice teachers of English work with assessment of oral English in lower and upper secondary school, and how they view their competence in assessing oral English. In the planning phase of the project, after settling on the overarching research question, I decided to conduct interviews to collect expansive information from the participants. The main goal of this study is not to present a truth on the matter but rather, as stated by Brinkmann and Kvale (2015, p. 65), to contribute with "useful" knowledge. Hence, as the qualitative research interview attempts to understand the world from the subjects' points of view, describe their experiences or articulate reasons for action (Brinkmann & Kvale, 2015, p. 3), semi-structured interviews are suitable to shed light on the research question. According to Brinkmann and Kvale (2015, p. 150), the use of semi structured interviews can provide descriptions of how the interviewees interpret themes in the study. A semi-structured interview gives the researcher the opportunity to have some suggested questions ready that can guide the conversation, at the same a semi-structured interview opens up to changes of sequence and follow up questions (Brinkmann & Kvale, 2015, p. 150). I concluded that for the present study, using semi structured interviews would be beneficial because it invites the participants to speak freely about the topics under investigation, and it gives the researcher the chance to ask questions for further exploration or clarification.

3.2.1 The first interview guide

Two interview guides were designed in this study. One was created with the purpose of answering what challenges novice teachers of English face when assessing oral English and how they view their competence in regard to assessing oral English, which was the initial research question of the dissertation. The interview guide made for this purpose was used when conducting the first interview. However, the formulation of the research question and the design of the interview guide proved to be challenging. The thesis question was based on the challenges that novice English teachers face when assessing oral English without there being any proof that such challenges existed. The assumptions of challenges when assessing oral English came from my own work experience as a teacher, conversations with colleagues and fellow teacher training students. Naturally, the interview guide aimed to investigate these challenges among the teachers, which resulted in an interview that did not provide a nuanced picture of assessing oral English. Thus when going forward, the study changed the main focus from challenges when assessing oral English to how novice English teachers work with assessing oral English and how they view their competence in in regard to assessing oral English. The research question was no longer based on existing beliefs about challenges with assessment but opened up for potential challenges to be revealed and discussed. In a way, the first interview guide and interview functioned as a pilot in that the research question and interview guide was revised after finishing the first interview. Even though the questions in the revised interview guide is formulated in a more open way which allows for looking at other aspects, my existing beliefs, knowledge and reading of other studies, still affects the revised interview guide.

The information gathered from this interview is useful when answering the research question and therefore is included in the present study. Limitations with be discussed further in section 3.4.

3.2.2 The revised interview guide

The revised interview guide consisted of twelve questions constructed to give information that would help to answer the overarching research question: how do novice teachers of English work with assessing oral English, and how to they view their competence in assessing oral

English (Appendix 3). I identified topics that I believed would be relevant to shed light on the subject and formulated the questions accordingly. When developing the themes and questions for the interview I looked to earlier research by Bøhn (2016) presented in chapter 2 of the present study, and a masters' thesis by Johannessen (2018) who explored teachers' understandings of what to assess. The initial six questions aimed to provide relevant information about the participants, and to get the interviewees talking and to feel comfortable in the interview situation. Thus, they were asked questions about their workplace, teaching background and experience. The guide was written in Norwegian, and the purpose of the interview was transparent to the participants from the start.

I designed the interview guide so that the participants would answer what, how and why. Kvale (2007, p. 58) problematizes the 'why' questions in an interview, claiming that these questions might lead to an over-reflected intellectualized interview, however such questions about the subjects' own reasons for their actions may be important in their own right, which I believed they were in the present study. Kvale (2007, p. 58) suggests that the 'why' questions should be posed towards the end of the interview, in this case the questions were posed after the participants had answered questions of what and how about a topic before they were asked to answer why. Seeing as the interviews were semi-structured there was an, as Kvale (p. 65) puts it, openness to changes of sequence and question forms in order to follow up the answers given and the stories told by the interviewees. The initial questions were broad and open, allowing the participants to speak freely. The possibility of asking follow up questions gave the researcher a chance to probe and ask for clarification if necessary.

According to Kvale (2007, p. 57) the questions in an interview should be easy to understand, short, and devoid of academic language. In the present study, it was necessary to include some academic language: formulations in the Norwegian curriculum, assessment for learning (AFL), and assessment of learning (AOL). In the beginning of the interview, the academic language and the terms were discussed with the participants so that they were able to give their own definition of these terms and their own understanding of the curriculum.

Going forward the participants were asked how they work with oral assessment throughout the schoolyear, what competence aims they deem as important when assessing oral English, and

how they interpret and operationalize the competence aims. In addition, the participants were probed to give reasons for their choices and interpretations.

To find out what constructs the participants believed to be important and their understanding of the constructs when assessing oral English they were given a sheet that presented the constructs: communication, content, vocabulary, grammar, pronunciation, and fluency. Going forward the participants ranked the importance of these constructs and gave their reasoning behind this ranking.

Seeing as there is a common national rating scale in lower secondary school but not in upper secondary, I wanted to explore if and how the teachers use this and their opinions about it. Thus, the participants were asked if such a rating scale, local or national, existed for them to use, and whether or not they used this as a tool in assessment. In addition, I asked the participants to give their thoughts about such a common frame of reference.

The present study wanted to investigate not only how the participants work with oral assessment in English but also how they view their own competence when assessing oral English. Therefore, I inquired about the participants' education, if and how it had prepared them for assessing oral English, and asked the participants to describe their own competence when it comes to assessing oral English.

3.2.3 Recruitment of the interviewees

When recruiting participants, I had to define the term 'novice teacher'. In the present study, a novice teacher is a teacher who has been formally qualified as a teacher for three years or less. The sampling technique used to collect data is a non-probability approach called convenience sampling that allows the researcher to choose participants that are available; e.g. fellow students or colleagues (Biggam, 2015, p. 165). The approach was chosen because of the challenges with recruitment of participants: some of the people that were asked to participate declined. I believe that the present study would have benefited from having more than four informants as this would have contributed to a more nuanced picture of how novice English teachers work with assessment.

As Biggam (2015, p. 165) points out, there are some limitations to using convenience sampling: seeing as the sample has not been selected randomly, one can not claim that the results are representative for a larger population. However, seeing as the present study does not aim to provide a truth about the topic but rather useful knowledge, I would argue that the information gathered by convenience sampling is valuable. The possible limitation of convenience sampling will be discussed further in section 3.4.

Four novice teachers of English in the eastern part of Norway were asked to participate in the study and they all accepted. Two of the teachers work in lower secondary school while the two other teachers work in upper secondary school. Below is a short introduction of the participants, providing details about their education and teaching experience. For anonymity purposes pseudonyms have been used, namely Lynn, Sara, Mark and Tia.

Name	**Lynn**
Year of graduation	2017
Formal qualifications	Teacher education Year 5 – 10 160 credits in English
Teaching experience after being fully qualified	Lower secondary teacher (8th grade) for one school year Upper secondary teacher (year 1 – 3) for one and a half school year

Name	**Sarah**
Year of graduation	2018
Formal qualifications	Teacher education Year 8 – 13 160 credits in English

Teaching experience after being fully qualified	Less than six months as a upper secondary teacher (year 1-3)

Name	**Mark**
Year of graduation	2016
Formal qualifications	Teacher education Year 5 – 10 60 credits in English
Teaching experience after being fully qualified	Less than six months as a lower secondary teacher (8th grade)

Name	**Tia**
Year of graduation	2018
Formal qualifications	Teacher education Year 8 – 13 160 credits in English
Teaching experience after being fully qualified	Less than six months as a lower secondary teacher (9th grade)

Table 1: The participants' formal qualifications and teaching experience

3.2.4 Conducting the interviews

The interviews were conducted at a time and a location that was convenient for the participant. A few weeks before the interviews, the participants were sent an e-mail which contained a formal document that stated the purpose of the project, and provided them with information

about approximately how long the interview would take, and that the interview would be audio recorded and later transcribed if the participant consented (Appendix 5). The e-mail also provided the participants with information about approval from The Norwegian Centre for Research (NSD) in regard to the research question and interview guide (Appendix 6). To make the situation as comfortable as possible for the interviewees they could choose to conduct the interview in Norwegian or English, all of them chose Norwegian. They were given the choice of seeing the interview guide before the interview, and Tia was the only participant who decided to see the interview guide beforehand. The participants were encouraged to contact the researcher if they had any questions.

All four of the participants signed the consent form before starting the interview and were informed that they could withdraw their consent at any time. The interviews were scheduled to last from 20 to 30 minutes but it turned out that the interviews varied in length: the first interview was the shortest, lasting for 12 minutes, and the longest lasting for 39 minutes. The varying length of the interviews will be discussed as a possible limitation in section 3.4, in addition to using two different interview guides in the study, and only one participant, Tia, viewing the interview guide prior to the interview.

3.3 Data analysis

When starting the data analysis I looked to Bøhn's doctoral thesis (2016) for inspiration. Bøhn (2016, p. 42) made a point out of all the factors that could influence the interview 'output': he mentions the interviewers' bias, expectations of the interviewee, and the possibility that the interviewee wants to express a certain identity. I realize that the interviews conducted were complex situations, both socially and linguistically.

3.3.1 Transcription

All four of the audio-recorded interviews were transcribed within a week after completing the individual interviews. I chose to transcribe all of the material from the interviews because I did not know yet which sections would be relevant to include in the final report. Because the interviews were conducted in Norwegian and were therefore transcribed in Norwegian.

Passages that were relevant to include in the report were translated to English. Roberts (1997, p. 168) points out that transcribed data is already interpreted data, all transcription is representation and there is no natural or objective way in which talk can be written. Keeping this in mind the transcribers have to develop a transcription system that can best represent the interactions they have recorded: the transcriptions should be accurate and readable but also make it clear to the reader that this is constructed, which is problematic for accuracy and readability (Roberts, 1997, p. 168). According to Kvale (2007, p. 98) the question of what a correct valid transcription is, cannot be answered. Kvale (2007, p. 98) argues that since there is no true, objective transformation from oral to written form, a more constructive question would be 'what is a useful transcription for my research purposes?'. For the purposes of the present study I chose to transcribe the interviews only by writing down what the participants said, I did not include intonation or emphasis, but pauses were included in order to make sense of the text.

3.3.2 The process of analyzing

Biggam (2015, p. 191) proposes a data analysis process used for a case study at Inverclyde University. This way of analyzing the data (see figure 2) has been used in the present study.

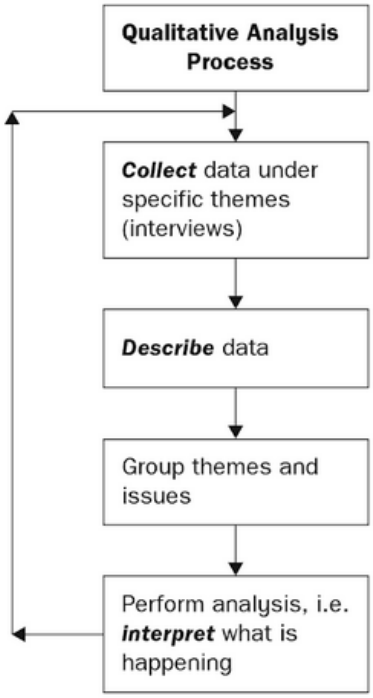

Figure 2: Qualitative data analysis

The researcher begun collecting data using interviews, the themes for these interviews were oral assessment in English and perceived competence when assessing. After collecting the data, the data had to be transcribed before themes and issues were identified. Finally, the data was interpreted. However, Biggam (2015, p. 190) refers to Wolcott (1994), Miles and Huberman (1984), and Crewell (1997) who emphasize that the process of analyzing qualitative data is an iterative process that makes it possible to capture and understand themes and patterns. In the present study, even during the collection of the data, the data has to some extent, been described, themes and issues have been identified and interpreted throughout the research process.

The process of analyzing the data was not linear but a movement back and forth between the different phases. Nevertheless, there has been a systematic analysis of the data, grouping of themes and issues, and interpretation of the data. Each of the questions from the interview guide has its own table and the participants' responses were sorted into these tables. Table 3 below shows an example of a table for systematically analyzing the data.

Question	When assessing oral English, describe how you would rank the following constructs from the most important to the least important: communication, content, vocabulary, grammar, pronunciation, and fluency.		
Lynn	**Sara**	**Mark**	**Tia**
1) Communication/ content 2) Vocabulary/ pronunciation 3) Fluency 4) Grammar	1) Fluency 2) Communication 3) Content 4) Vocabulary 5) Pronunciation 6) Grammar	1) Communication 2) Vocabulary 3) Pronunciation/ fluency 4) Grammar 5) Content	Tia was not asked this question because she was interviewed using a different interview guide.

Table 2: Systematically sorting of the data

The table made it easy for the researcher to gain an overview of the answers given by the participants. Looking at the table, one can easily spot that the participants have different opinions of the importance of the constructs. The participants also gave reasons for their ranking, thus in the empirical research findings chapter, relevant excerpts from the interview transcriptions with the participants explanations for the ranking is included and discussed. Figure 2 allows me to describe the data, group themes and issues, and this then form the basis for an analysis.

3.4 Possible limitations

The present study does not seek to provide a truth about a topic, and it is important to note that the participants are not necessarily representative for a larger population. There are only four participants in the study and they have been sampled by convenience. However, the participants provide information about how they work with assessing oral English and how they view their own competence. This information can lead to interesting discussions and can possibly contribute to further studies regarding assessment. It is important to note that, rather than universal knowledge, interview knowledge is situated knowledge – therefore transferring this knowledge to other situations is problematic (Kvale, 2007, p. 143).

Another possible limitation with the sampling technique is that the researcher already had a relation to the participants in the study, which can influence the research. Nevertheless, the interview regarded a professional subject and was carried out in a professional manner to ensure that the results are as fair and free from bias as possible.

Using interviews as a way of collecting material is an issue that needs addressing. Kvale (2007, p. 24) states that one has to take into account the interview situation: the interaction between the subjects, stress during the interview and self-understanding. These factors might influence the results in the present study. To avoid that the participants would give answers that they thought the researcher wanted to hear, and to avoid leading the participants in one way or the other, the questions were open and wide, and the participants were encouraged to speak freely. In an attempt to reduce stress during the interview, the researcher started the interview with questions about experience, education and place of work that were easy to answer.

At the same time, using interviews as a sample technique is problematic in terms of personal opinion. One must consider ethical issues in the process of analyzing interviews: of how penetratingly the interviews can be analyzed and whether or not the participants should be included in the interpretation of their statements (Kvale, 2007, p. 24). The present study did not open up for the participants to have a say in the interpretation of their statements because this was regarded as too time consuming. Not giving the participants this opportunity is a possible limitation: the researcher may be interpreting the statements in a different way than they were intended.

As mentioned in section 3.2.1, the present study has applied two interview guides. One of the participants, Tia, was interviewed using an interview guide and with a research question that was later revised before conducting interviews with the other participants. Furthermore, Tia was the only participant who wanted to look at the questions before conducting the interview. Hence, it is problematic to analyze Tia's interview and the other participants' interviews the same way. As the present study does not seek to provide any generalizations but rather provide valuable information, Tia's interviews is included because the researcher considers this as valuable information that will be beneficial for providing nuance in the study.

4. Empirical research findings

4.1 Introduction to the chapter

This chapter is concerned with the results of the empirical research findings of the present study. The research concentrates on four novice English teachers and their thoughts about assessing oral English. Mark and Tia work in lower secondary school, and Sarah and Lynn work in upper secondary school in the eastern part of Norway. To begin with, the aim of the present study was to investigate both formative and summative assessment. However, during the interviews it became apparent that, for the most part, the participants understood assessment to mean grading. Therefore, even though aspects of formative assessment is mentioned, the research findings are mostly concerned with summative assessment in oral English.

The chapter is structured in the following way: identified main topics from the interviews serve as headings. Under each heading relevant questions and excerpts from the transcribed interviews will be described and connected to the theory from chapter 1 and previous research described in chapter 2. The headings were formulated after identifying essential areas that the participants had in common and expressed during the interviews. Parts of the transcripts from the interviews can be found in Appendix 4, with the purpose of demonstrating how the interviews have been transcribed.

Section 4.2 will present findings about how the teachers plan oral assessment so that the learners are able to demonstrate their competence. In the following sections, the findings concerning operationalization of the competence aims in the English subject, what is assessed, and issues related to a lack of shared understanding will be presented. Finally, the novice teachers' thoughts about their own competence in assessing oral English and their view on their education will be addressed.

4.2 Learners demonstrating competence in oral English

The teachers were asked how they plan and work with oral assessment in English. It was emphasized from the researcher that both formative and summative assessment were included

in this question. The teachers report two ways of assessing oral English: student presentations or having conversations with the students.

Lynn said that usually her students have some kind of presentation, either in front of the whole class or in small groups, or they can hand in a digital presentation. Lynn conveyed that she usually lets the students choose how they want to give the presentation, and states that the reason why is that the students should be allowed to join in on decisions about how to work in the subject. The hope is that they will gain some sort of ownership to the learning process:

> *I think that the students should be given the opportunity to decide how they are assessed, they know themselves and know in what situations they are best able to show they competence in English. In addition, it has to do with the student's personality and relations to the rest of the class. In a group of students where everyone trusts each other, many students choose to do the presentation in front of the class. The students who are insecure usually choose to hand in a digital presentation.*

Involving the students in the learning process is in line with the concept of formative assessment where two of the principles for success are that the students understand what is expected of them and that they should be involved in their own learning process (Norwegian Directorate for Education and Training, 2015, p. 1-2).

Mark had been working as an English teacher for three months and at the time of the interview, he had only conducted one oral assessment in English with his eighth graders: a video presentation. Mark says that some of his students are anxious about presenting in front of the whole class. Therefore, as this was the first presentation in the eighth grade, he chose a digital presentation that they could hand in. Mark states that he will continue letting the students have a say in the oral assessment situations. At the same time, student presentations is something the class will continue with, Mark gives the following reason:

> *As the curriculum is now, in the case of an oral exam in English after year 10, the students have to be able to give a presentation of some sort. I look at it as practice for the oral exam.*

Mark argues that presentations are a traditional way to assess oral English in schools, and given that the oral exam in year 10 takes the form of a presentation, it is sensible to practice what is to come. Sarah on the other hand, reports that at the school where she works, they do not use student presentations as a means for assessing oral English. Instead, they have conversations with the students and believe that this will allow them to achieve a higher level of competence:

> *The students show much higher competence when conversing than when presenting. Listening is part of oral competence, and by using conversations the students show that they are able to participate in a conversation, and communication is in focus.*

The basic skill 'oral skills' defines that the students must be able to listen, speak and interact using the English language, and use the language in conversations and discussions (Norwegian Directorate for Education and Training, 2013, p. 4-5). This is in line with what Sarah advocates. When asked to elaborate on why she chose to use conversations instead of presentations, Sarah said the following:

> *The students at my school usually have a high competence in the English subject, I would say that around 90 per cent of the students master the details of the language, so there is no need to address this. Using conversations in oral assessment allows them to show reflection, which is one of the skills that are of high value now and aligns with the socio cultural view of learning. The old form of presentation is very passive and has little room for spontaneity. The students have to be able to adjust to the situation, if they get a question that they are not prepared for they have to show that they are able to master that as well, then, they will show a higher degree of competence in my opinion.*

What Sarah says aligns with how competence is defined in the Knowledge Promotion: "competence is the ability to solve and master complex challenges. The students show competence in specific situations by using knowledge and skills to solve the tasks at hand" (Norwegian Directorate for Education and Training, 2016, p. 1, own translation). Using conversations is a more complex communicative situation than giving a presentation: the student cannot fully prepare for what is to come in a conversation.

It is worth noting that Mark and Sarah do not assess with grades at the time of the interviews. Mark says that they do not have summative assessment until the end of the term. At Sarah's

school, an upper secondary school, they only use grades at the end of the term and when setting the grade for the student's overall achievement in the subject, a decision that is meant to help the students focus on the learning process instead of the grade. As mentioned in Chapter 1 (section 1.3), formative assessment has gained increasing prominence in Norwegian schools (Nusche et al., 2011, p. 50). Focusing on the learning process is evident in Mark and Sarah's practice of assessment without grades. In addition, a part of formative assessment is to include the students in the learning process: which Lynn and Mark does by letting the students take part in deciding how to perform in oral English.

4.3 Operationalization of relevant competence aims

The competence aims in the English subject are broad and vague, and with so many schools and teachers interpreting the curriculum, it is reasonable to assume that there are differences in how the competence aims are operationalized. Thus, a critical issue in the present study is how the teachers conduct this operationalization and their thoughts on the matter.

From interpreting the competence aims, Lynn usually formulates learning objectives that she provides for her students. Lynn problematizes how the interpretation of competence aims is largely left to the local school and at times the individual teacher. Lynn says that at the school where she works, they cooperate with other schools within their organization. She perceives that, within their organization, they have a shared understanding of the competence aims. However, she is not sure that their organization's understanding is shared by other schools and teachers:

> *I am always uncertain of whether or not I have given my students the right focus because I believe that it varies. In the case of an oral exam, I am not sure that the external rater and I have interpreted the competence aims in the same way.*

Lynn argues that different schools and teachers will have different interpretations of the competence aims. As she understands it, many teachers base their teaching on different textbooks and other learning material provided by a range of publishing houses. Lynn believes that these textbooks and learning materials advocate various content, thus it is reasonable to assume that the focus differs. If the understandings differ, the assessment can differ as well:

Nusche et al. (2011, p. 52) argue that a lack of shared understanding can result in unfair assessment of the students.

Mark reports that he has to interpret and operationalize the competence aims by himself. At the school where he works they do not have a team of English teachers working together but he still feels that he can ask his colleagues if he needs help. Mark thinks it is challenging to work individually with the competence aims:

> *It is difficult. The interpretation and operationalization is characterized by my own subjective opinion. I have to say that it would be nice to discuss it with someone.*

Mark's thoughts about the subjective nature of assessment is evident in the literature as well: Ang-Aw and Goh (2011, p. 31) state that assessment is a complex process and a subjective assessment of a person's language abilities. One way to attempt to reduce the subjectivity in assessment is by trying to form a shared understanding as they do at Sarah's school. Here, the teachers work together to interpret and operationalize the competence aims. According to her, they have a shared understanding of these aims:

> *At the time, we are working on making a model conversation or presentation so that the students know what is necessary to achieve a high level of competence in oral English. We talk to the student about key words like what is a wide range of vocabulary, what does fluency mean, what is good intonation, what is good pronunciation, is it important to sound American or English.*

When working with the competence aims, Sarah, unlike Lynn, does not formulate learning objectives from the competence aims found in the English curriculum. Instead, Sarah connects the different competence aims. When giving her reasons why she says that:

> *There are many aspects of oral competence, I think it is "old school" to break the competence aims into smaller learning objectives, apart from when a student is struggling with a specific sound and wishes to change it, then we can formulate smaller objectives. But it is complex so we use complex situations as well.*

Nevertheless, Sarah reports that she formulates criteria for assessment with the students. This is not necessarily in relation to the competence aims but rather in connection to communication.

In order to break down the concept of communication, Sarah makes bullet points. Sarah argues that the competence aims are too detailed for the students, and this makes them inaccessible to them:

> *If you create more learning aims from the competence aims, it will be even more aims to attend to and it will be too much for the students.*

The Norwegian Directorate for Education and Training (2016, p. 4) states that it is beneficial to develop local learning aims and criteria in order to reach the competence expressed in the competence aims. However, the students are going to be assessed in the competence aims, thus the competence aims must not be divided in a way that loses perspective of the overarching aims (Norwegian Directorate for Education and Training, 2016, p. 2). At the same time, Nusche et al. (2011, p. 52) write that the competence aims are not specific enough to guide teaching and assessment. In addition, a principle of formative assessment is that the students understand what is expected of them (Norwegian Directorate for Education and Training, 2015, p. 1-2), a understanding that might be difficult to comprehend using only the competence aims. Sarah, however, believes that creating more learning objectives will not lead to clarification for the students, but rather more confusion.

Tia states that at her school they work in teams to interpret and operationalize the competence aims: they usually spend time discussing the competence aims. Even though Lynn, Sarah and Tia cooperate with their colleagues to understand the competence aims, it is important to note that there is no guarantee that teachers discuss grading criteria within or across schools (Nusche et al. 2011, p. 54). For instance, Mark does not have a team to cooperate with, yet he still feels like he can ask his colleagues for help if he needs it. Differences in the understanding of criteria for assessment will be discussed further in chapter five (section 5.1.2).

4.3.1 Relevant competence aims in oral English

The teachers were asked what competence aims they view as relevant when assessing oral English, and if they focus on some aims more than others. On this question, the four teachers had some common thoughts: they all viewed the competence aims listed under oral

communication as the most important aims when assessing oral English, more important than competence aims listed under language learning and culture, society and literature.

Lynn said that she focuses on different competence aims throughout the school year but mostly the students are assessed on the aims listed under oral communication along with one aim from culture, society, and literature. Lynn believes that the aims listed under oral communication are the most important ones to assess. Lynn mentions one competence aim: "introduce, maintain and terminate conversations and discussions about general and academic topics related to one's education program", and claims that this competence aim is easy to focus on as it is specific.

Mark also believes that the aims listed under the category oral communication are the most important ones to assess. Mark mentions the competence aim "introduce, maintain and terminate conversations on different topics by asking questions and following up on input" as important. In addition, Mark points out some specific competence aims he usually focuses on: "choose and use different listening and speaking strategies that are suitable for the purpose", "express oneself fluently and coherently, suited to the purpose and situation", and "express and justify own opinions about different topics". In order to communicate in a good way, Mark argues that it is beneficial to be able to use the central patterns for pronunciation, intonation, word inflection and different types of sentences in communication

When giving his reasons for why he believes the communicative competence aims are the most important to assess Mark argues that:

> *Later in life, many of my students will not need to be able to discuss how they live in Britain and in the USA. They will learn about history and geography in other subjects at school, like in social sciences class. Personally, I would say that, apart from working as a teacher, I do not have much use for a high competence in many of the aims. It is not as if I discuss these things with my friends.*

Mark uses experiences from his own life in his understanding of the purpose of education, and this affects which competence aims he focuses on. This relates to the concept of teacher cognition: what teachers think, know and believe. Borg (2006) argues that teachers' lived experiences, such as their education and their experience as learners, informs their cognitions

about teaching and learning. This might lead to differences in the students' education depending on their teacher's thoughts, knowledge and beliefs.

Mark continues his explanation saying that the competence aims in culture, society and literature will to some extent be covered in the other subjects at school, history and politics and such. However, Mark seems to be a bit inconsistent in his statements as he says that the competence aims relating to culture, society and literature are important as those who do not learn history are doomed to repeat history.

As with Lynn and Mark, Sarah believes the oral communication aims are more important to focus on when assessing oral English than the competence aims listed under language learning and culture, society and literature. She argues that the competence aim "express oneself fluently and coherently in a detailed and precise manner suited to the purpose and situation", is the most important one as this is important in all of the areas; language learning, oral communication, written communication, and culture, society and literature. At the same time, Sarah highlights that the students have to talk about *something*, hence the content is also of value. This aligns with Bøhn's (2015, p. 9) findings that examiners focus on students' ability to reflect on content, hence, including topical knowledge in the classroom practices is necessary.

When giving her reasons for why oral communication aims are more important, Sarah states that:

> *In the new curriculum, one can see that communication in the English subject is central, the students have to be aware of the situations they are in: adjust the communication to the recipient, and they have to be able to connect this in light of different cultural factors and the society.*

Competence aims listed under oral communication were the most salient for Tia as well:

> *I think the most important thing is the students' ability to communicate. Thus, every competence aim that explicitly mentions communication are the most essential.*

Tia said that in addition to oral communication she assesses the structure of the student's presentation: that they have a clear introduction, main part and conclusion. Tia continues saying

that she focuses on language, that the students have decent range of vocabulary, intonation, pronunciation, and that they have few grammatical errors.

4.4 What is assessed?

The teachers were asked to rank the importance of the following constructs when assessing oral English: communication, content, vocabulary grammar, pronunciation, and fluency[2].

Lynn	Sara	Mark
Communication/content	Fluency	Communication
Vocabulary/pronunciation	Communication	Vocabulary
Fluency	Content	Pronunciation/fluency
Grammar	Vocabulary	Grammar
	Pronunciation	Content
	Grammar	

Table 3: Ranking of constructs

Lynn had trouble ranking some of the constructs, especially communication and content, as content is a crucial part of communication. After debating back and forth, she said the following about communication and content:

> *I want to put communication over content but I think that is unrealistic. Because I assess to what extent the student communicates, and something has to be communicated. So, content and communication is equal.*

Communication and content was not the only constructs Lynn found problematic to rank: vocabulary and pronunciation came second. Fluency came next and then grammar. Lynn

[2] Tia was interviewed using a different interview guide, therefore Tia has not answered the question and is not presented in the table.

believes that fluency is important for the rhythm of the language but she also points out that oral communication often lacks fluency:

> *If you listen to the conversation we are having now, you will find that the fluency is not good. If the oral communication is to be as authentic as possible, I teach the students to take some breaks with natural 'ums' and 'ehs' and everything else you will find in an oral conversation.*

Lynn ranked grammar last. However, she does state that grammar is important for how the student communicates but as long as they can make themselves understood, grammar is less critical. According to Lynn, it can be a bit unfair to assess the students' grammar: she believes that a student can receive a high level of achievement in the subject even with grammatical errors. Lynn points out that grammatical errors are not unusual even for native speakers of English.

Mark advocates that communication is the most important construct as this is what the students will need in English. Further, he says that content in itself is not as important, and puts it least in the ranking. He gives the following explanation:

> *It depends on which competence aim you are working with but to be able to have a functional English for the rest of your life, the content in itself is not the most important thing but rather how you communicate this content.*

According to Bøhn's (2015, p. 9) findings, examiners will focus on the student's ability to reflect on content. Even though Mark states that what is in focus depends on the competence aim, he says that content is not the most important thing, and in section 4.3.1 he argues that the students can learn much of the content in other subjects at school. Thus, it seems that content in oral English is not Mark's main priority. There are differences in how Mark and Lynn assess oral English as Lynn ranked content first and Mark last. This might have implications for their students in a possible oral exam situation if the external rater attends to the constructs in different ways than they have done and the students are used to.

Sarah says that she is having difficulties distinguishing the constructs. She ranks the constructs in the following order: fluency, communication, content, vocabulary, pronunciation and grammar. Sarah does not perceive fluency as an own category and says that:

> *I rank fluency first because for me fluency is not purely language or content but the combination of it all. It is the fluency that, if you have bad fluency it will hinder communication, if you have good fluency it will better the communication, so it is part of everything. It is not a single category for me.*

According to Sarah, communication of the content is most central, so she finds it challenging to choose which one of these constructs are more important, and adds that the combination of communication and content is essential for fluency.

When it comes to grammar, vocabulary and pronunciation, Sarah believes that if the level is sufficient it does not hinder communication, whereas if the student has many errors in grammar, a small range of vocabulary and difficulties with pronunciation, this will hinder communication.

> *If the pronunciation is not native-like that is fine. The vocabulary does not have to be advanced and you would have to make big grammatical errors not to be understood.*

At times, the teachers struggled with the ranking, which can be an indication that the ranking of the constructs can vary. Variation in the understanding of which construct and criteria are viewed as salient, was evident in Bøhn's (2015, p. 8) study as well. A consequence of teachers weighing constructs differently can be that what they focus on different aspects in an assessment situation, which might lead to dissimilar scores.

Tia, as stated previously, was not asked to rank the constructs. However, Tia reported that what is important to assess depends on the student's level of competence in the subject. Her utterance implies that the importance of the constructs vary:

> *For students with low competence in the English subject, I focus more on content and their ability to put forward their ideas when I assess. When I assess students that have medium to high level of achievement in the subject, I can focus more on pronunciation, fluency, structure and other things.*

The tendency of assessing differently depending on the level of the student was also reported by Bøhn (2015, p. 6): VSP students, and especially weaker students, were assessed more leniently than GSP students were. Tia reports that she pays more attention to content when she assesses students with lower levels of proficiency in English, and more linguistic features with

students with higher levels of proficiency. However, Bøhn (2016, p. 59) refers to Brown et al., (2005), Pollitt and Murray (1996) and Sato (2012) who found the opposite: raters pay attention to linguistic features with students with lower levels of proficiency, and content at the higher levels.

4.5 A common frame of reference

In lower secondary school, a national frame of reference for oral assessment in the English subject is provided as rubrics consisting of criteria for the different levels of achievement. This rating form is meant to guide assessment and teaching. The same cannot be found in the mandatory English subject in upper secondary school where the teachers have to rely on the subject curriculum consisting of the competence aims as a common document for assessing students' competence in oral English.

Lynn, who works as a teacher in upper secondary, teaches the mandatory English subject and international English, a subject that students can choose in year two in upper secondary. Lynn reports that there is no frame of reference in the mandatory English subject but there is one in international English, however it is difficult to use as a tool:

> *The common frame of reference in international English is very hard to read and especially to use as a tool for the students. In the beginning, I tried to use it without success. I think it is very vague and the level is too high for the students. And it is difficult for me to make sense of it.*

When asked if she believed that this common frame of reference was beneficial, Lynn answered no. Lynn did say that they use the criteria for assessment as a starting point but that they are difficult to interpret:

> *It is difficult to know what weight that should be given to each criteria. In the Norwegian subject, raters have gotten a lot of training, it is not the same in English. I would like to be a rater at an exam but have not yet been picked for the task and I have an understanding that more experienced raters get picked for this job. I will continue to volunteer as an exam rater as it would be very useful to see how others weigh and*

interpret the same criteria. Because I am not certain that we interpret it in the same way.

The broad competence aims open up for discrepancies in the operationalization between schools and individual teachers. Different understandings can result in varying assessments. It is peculiar then, that a guide for how to assess is only provided for English in lower secondary and the subject international English in upper secondary, and not for the mandatory English subject in upper secondary. This will be discussed further in chapter five (section 5.2.1).

Lynn has doubts about whether or not teachers and schools interpret the competence aims and the common frame of reference in the same way and points at some of the implications this might have:

Students will receive different training. It will lead to differences in the final grade. If I assess the importance of grammar and fluency in a different way than others, my students might get an advantage and a higher final score. Maybe I am more strict than others, I mean, we interpret it so differently and that will give different results on the students' diplomas which will give them different benefits when applying to higher education.

Lynn continues saying that she would like rater training to reduce differences in rater perceptions and behavior. Rater training as a way to ensure reliability is advocated by Luoma (2004, p. 177). However, Ang-Aw and Goh (2011, p. 44) found that even when given similar training assessment was not entirely reliable. The raters followed the marking scheme to varying extent, were preoccupied with different aspects of the candidates' performances, and assessed in a dissimilar way (Ang-Aw and Goh, 2011, p. 44).

When asked if there is a need for a common frame of reference in the English common subject, Lynn answers that there is a need for it but that the criteria for what to assess must be much more clear:

I have not yet had a group of students picked out for an oral exam but we had one group at the school last year and it was clear that the external rater had a totally different understanding than us. The external rater emphasized content a lot more. We had a different understanding of how to conduct an oral exam. It might as well have been me

who had students up for oral examination and I would have done the same as the teacher I work with. That proves that we have a similar understanding within our organization but when external raters come there is a deviation.

Lynn is under the impression that more experienced raters are picked for assessing oral exams. Even though this means that less experienced teachers will not get much practice, it can be advantageous for the assessment: as Kim's (2015, p. 248) research showed, novice teachers were often confused by the rating scale, while experienced teachers generally understood the rating scale correctly. Thus, using more experienced teachers as raters might improve the reliability and validity of the assessment.

Lynn states that the measuring of students' level of achievement is not accurate, and that this is a weakness in the educational system. As it is now, Lynn does not believe that the schools and the teachers provide assurance that the assessment is fair and that the students are applying to higher education with the same terms. Lynn's view can find support in Green's (2014, p. 76-78) questions that when assessing language ability, how do we know that we have the same understanding of the concept, how can one prove the truth of their claim that one person has a better language ability than another. Lynn's concluding thoughts about a possible common frame of reference in the English subject is as follows:

It would be nice with a common frame of reference but it would have to be specific. Clear on what criteria should be important, and there should be more training.

Mark, who is a teacher in the 8th grade, can refer to the national common frame of reference in lower secondary school when assessing oral English. Mark states that he does not use this tool much but he thinks he will start to use it more. Mark believes it can be a good tool but points out that even with a common frame of reference there will still be a degree of subjectivity in oral assessment:

Everyone makes it into their own. It can help but if you look at the criteria for grade 3 and 4, it will be up to the listener to decide how good it is. Right? It is not specific, it is very vague. And it is my subjective opinion of what is good that counts.

As has been pointed out, the competence aims and the assessment criteria are perceived as unclear. General descriptions of how to assess seems to be a tendency in the Knowledge

Promotion: in the descriptions of the different levels of achievement provided by the Regulations to the Education Act (§3-4, 2009), the definition of what constitutes the different grades is largely up to the individual teacher. This is exactly the point that Mark problematizes: the grade given to the student depends on the teacher's subjective opinion of what constitutes the different levels of competence.

Mark believes that even with a rating scale there will be subjectivity in the assessment. Orr (2001, p. 143) argues that scores are reached subjectively using rating scale descriptors to guide the examiner towards a number, which cannot be said to secure reliable and valid assessment. When asked to elaborate on the subjectivity in oral assessment, Mark said the following:

> *It leads to a lot of differences. I am sure that many teachers have different opinions than me when looking at the constructs you asked us to rank, then they will give other grades I guess.*

Such differences should not occur, as Black and William (2012, p. 244) states: any variation in scoring should be due to differences relevant to the construct of interest, not irrelevant factors such as who did the scoring.

Sarah, who works in upper secondary, says that they do not have a national common frame of reference but that they have made an informal one at the school. In addition, the teachers focus on making the assessment reliable and valid: sometimes two teachers assess the same student. Sarah believes that it would be beneficial to have a common frame of reference developed nationally mainly because it would help develop a shared understanding and make the assessment more valid nationally. She points out that she does not have much faith in using scoring rubrics as a common frame of reference but rather recordings or videos demonstrating the different levels of competence. When asked to give reasons why she did not have much faith in scoring rubrics, Sarah said the following:

> *It is difficult because, in my experience, it is too much for the students to relate to. So many details, ranking the students from low, medium to high. We are taking the focus away from that and into: are you communicating, yes or no? What are you not communicating and how can we improve this? It removes some of the subjective aspect of me as a rater. Assessment is spontaneous, it is subjective and very immediate. I think*

> *it is problematic to rank the students in detailed scoring rubrics because if a performance feels like a grade four or five can depend on the day that I am having. So, for the assessment to be more valid I do not use this.*

One point that Sarah makes is that not using scoring rubrics will, to some extent, remove the subjective aspect of the rater doing the scoring. However, Louma (2004, p. 172) advocates rating forms because they help structure the rating process, make it consistent and define what the raters pay attention to when assessing. In addition, rating forms is meant to help the rater compare an examinee's performance to the criteria rather than the other examinees' performances (Luoma, 2004, p. 172). Bøhn (2016, p. 8) is also in favor of rating scales and states that in order to focus on those aspects of the performance which the test is intended to measure, rating scales are considered invaluable tools. Rating forms/scales or scoring rubrics are meant to guide the rater and decrease subjectivity in assessment but, according to Sarah, it leads to more subjectivity in scoring performances. Rater scales/forms as tools for assessment in oral English will be discussed further in chapter five (section 5.2.1).

4.6 Working with assessment in teacher education

Seeing as the present study focus on novice teachers, the participants where asked about their education and if/how it had prepared them for the job of assessing oral English. In addition, I wanted to find out how competent they feel when assessing oral English.

According to Lynn, her pre-service teacher education has not prepared her for assessing oral English. Lynn reports that they did work with assessment during her education but she did not feel like she could transfer and use the information in the real world. During her practice period, Lynn got to practice working with assessment but states that she did not get much out of it:

> *We had assessment in our practice periods but with the guidance of other souls trying to interpret the same material as us. So I do not think it was sufficient.*

Another factor Lynn found unsatisfactory in her education was that many of the lecturers at the university had never been working in schools themselves. Lynn believes that this affects the quality of the lecturers teaching, and the output the students are left with:

In the classes we had at the university, I did not find it trustworthy as many of the lecturers had not been working in schools themselves, I do not have any confidence in what they are saying. So, it is very, one gets left all alone. Trying to make sense of the cryptic content and criteria in the curriculum. So no.

Lynn believes that it would have been helpful to get more specific training in assessment: a visit from the department of education or getting the same training as raters do.

Mark also reports that his pre-service teacher education did not prepare him for assessing oral English. Mark did some oral assessment during one of his practice periods but says that it was a coincidence that he was given the opportunity to do this. He explains that practicing assessing oral English was not a planned part of the practice period. According to Mark, there is a need for more discussions about oral assessment in teacher education:

You could have discussed a video of an assessment situation and maybe created some tasks collectively. Because then it will not be as subjective anymore. It is very vague. I am not sure that any teachers experience a student's performance the same way. I do not have any facts to back it up but I believe that the student's grade depends on the teacher, so the students do not have the same opportunities to succeed. Many teachers may have similar interpretations but this can go either way.

During her pre-service teacher education, Sarah worked with a project concerning oral assessment, so she thinks that this part of her education prepared her for the task of assessing oral English. Other than her dissertation, the education did not prepare her for scoring student performance. Sarah believes that more practical work and discussions about assessment in oral English should be included in the education:

I think more practical didactics with different cases is the way to go. And that the constructs: content, fluency, grammar, vocabulary, pronunciation and communication are discussed. A common frame of reference for assessment should be formulated during the education, how to understand the constructs. And yeah, the education should take part in creating a common frame of reference, then much of the work with novice teachers will already be done.

Even though Sarah did not think that her education prepared her for assessing oral English, she points out that she does not think anyone will ever feel one hundred per cent confident when assessing. This may be true but Kim's (2015, p. 248) research indicate that the rater's competence improves with experience: novice teachers were often confused, the developing raters misunderstood parts of the rating scale but less frequently, and the experienced raters generally understood the rating scale correctly.

Like the others, Tia does not think that her education has prepared her for assessing oral English. Tia says that there has been much focus on theory about assessment but that they have not had useful practice:

> *There has been much theory about formative and summative assessment, what it is and so on. I do not think it has been much practice in how to assess. We did some in our practice period but who gets this opportunity in the practice period is a hit or miss. I have been very lucky while others have only been observing.*

4.7 Competence in assessing oral English

At the time of the interviews, Mark, Sarah and Tia had only been working as teachers for a few months, while Lynn had been working for two years. From the previous section (section 4.6), it is evident that the participants did not find their education satisfactory in regard to oral assessment in English. Having gained some experience from working as teachers, the participants were asked how they view their competence in assessing oral English.

Lynn does not feel confident that her assessment practice is correct but she thinks that she does it to the best of her abilities. As last year, she is anxious about sending in her report with oral assessment criteria:

> *I am anxious to send my report with the oral criteria that I have made, or my focus area, to see if the rater in the other end approves it, if not my focus for the year…we might have lost something important. Cause we have to agree locally: the subject teacher and the external rater. So you are in the hands of others. I have done what I believe is the best for my students and it is up to an external rater to agree on this focus or not. I feel*

confident that I do everything in my power and that I am thorough. I do not feel confident that I have done it correct.

Lynn worries that there will be even more confusion with the new curriculum that is coming.

I fear that the first two years after the new curriculum comes, no one is going to know what they are doing until the first rounds of exams where we will see what is being weighted. Once again, it will be interpretations.

Mark reports a low belief in his own abilities when assessing oral English.

I do not have confidence in myself when assessing oral English, no. I have an opinion about what I think is good in respect to the documents that we have looked at today. So I am competent but I am new at this, I am not sure that what I do is correct. I am not sure that I share a common understanding with other teachers. I think a lot about whether or not I am doing the right things, if I have enough knowledge, if other teachers agree with me. I stress a lot with assessment.

Sarah's experience is that her competence in assessing oral has changed because they work with conversations when they assess oral English. She says that she does not worry as much about her competence being good enough because:

I do not have a rigid understanding that every single assessment has to be valid. Not giving grades throughout the year kind of removes some of the responsibility of validity, as we are supposed to point them forward all the time, so I am not so worried about my competence not being good enough.

Sarah seems to focus on assessment for learning: the students are supposed to move forward all the time. All assessment with grades should be based on standards and be measurable (Ministry of Education and Research, 2004, p. 39), but when removing the grades, the assessment might not have to be based on standards to the same extent, nor can formative assessment be measured in the same way. This might be one reason why Sarah does not have a rigid understanding of assessment.

Tia believes that her competence is improving. She perceives that in the team that she works in, they have similar opinions about performances and assessment. However, she does report

that she quite often feels insecure about assessment. A shared understanding of what to assess is important to Tia but she does point out that one has to go with one's gut feeling at times:

> *I think a common understanding is important but sometimes you have to go with your gut and make a decision if there is disagreement, or get more teachers to assess. But it is important. We assess each others students to enhance objectivity. So yeah. It is important that we have the same understandings when assessing each others students, if not the assessment might not be fair.*

I interpret Tia's statement about one's gut feeling as a factor in assessment problematized by Ang-Aw and Goh (2011, p. 43-44): that the raters can end up feeling if a particular score is too high or too low for a candidate. The teacher's subjective opinion can threaten the reliability and validity of the score. However, getting more teachers to assess the same candidate performance will enhance the reliability of the assessment, as possible differences in scoring can be uncovered and discussed.

Tia was very insecure when she started working as a teacher and she did not know what she was meant to assess. Now, Tia has more confidence and is able to say exactly what she focuses on when assessing oral English.

> *I am more skillful now and feel more confident about what is important when the students give oral performances.*

4.8 Summary of empirical research findings

To assess oral English, presentations and conversations were mentioned as means to do so. Lynn and Mark reported that they used presentations to prepare the students for a possible oral exam. In addition, they wanted the students to take part in their own learning process and did that by allowing them to choose what way to perform e.g.: digital presentation, presentations in front of the whole class or conversations. Sarah, on the other hand, uses conversations when the students give oral performances. Sarah believes that conversations gives the students the opportunity to show a higher level of competence. Lynn believes that the students themselves know best how to show their level of competence, and gives this as another reason for letting them choose. This is in line with a central principle to achieve assessment for learning: that

students are involved in their own learning process (Norwegian Directorate for Education and Training, 2015, p. 2). Mark and Lynn only gives grades at the end of term in an attempt to have focus on the learning process, this is also in alignment with the principle of formative assessment.

Competence aims regarding oral communication are reported by the teachers as most important. At the same time the teachers point out that the students have to communicate *something*, so content is important as well. Like the others, Tia found communication aims superior, but she also said that what is important to assess varies depending on the student's level of achievement in the subject.

The teachers rank the constructs fluency, communication, vocabulary, grammar, content, and pronunciation differently, and they do not have the same understanding of these constructs. All of the teachers view the construct communication as the most important one, while the ranking of the other constructs varies. This supports Bøhn's (2015, p. 1) findings that there are differences in how raters perceive the importance of the constructs. However, his study showed that the raters had the same general ideas of the constructs to be assessed (Bøhn, 2015, p. 5) while in the present study, there are indications that the teachers do not have the same understanding of the constructs. Bøhn (2016, p. 59) found that there were notable differences in the importance attributed to the construct 'content': the teachers weighted this construct differently. The same is evident in the present study as Lynn ranked communication/content as the most important constructs to assess while Mark ranked content last. In addition, the differences in the understanding of the constructs was clear in the case of the construct 'fluency': Lynn stated that oral communication often lacks fluency and that this is ok. Sarah, however, gives fluency a much more central role saying that it is not purely language or content but a combination of it all.

All of the teachers said that within their school there is a shared understanding of what to assess in oral English. They do not think that there exists a shared understanding between schools and emphasizes that there is a lot of subjectivity when assessing oral English. The assessment is largely based on teacher judgement. The teacher's thoughts, knowledge and beliefs affects the classroom practice and the assessment. As shown in studies conducted by Bøhn (2015), Ang-

Aw and Goh (2011), Kim (2015) and Orr (2001), the subjective opinion of the rater influences the scoring of the students. This can be problematic for the reliability and validity of assessment.

A common frame of reference for assessing oral English is provided in lower secondary school and in the subject international English in upper secondary. The common frame of reference is developed as scoring rubrics with criteria for the different levels of achievement. However, such scoring rubrics are not provided for the mandatory English subject in upper secondary school. Lynn, who teaches both of these classes in upper secondary, does not think the frame of reference has been beneficial as it is difficult to interpret: how the different criteria are weighed is an issue of concern. Mark does not use the common frame of reference much. Still, he believes that it can be a good tool but that there is much subjective opinion involved in assessment anyway. Sarah thinks that a common frame of reference in upper secondary can make assessment more valid but is not sure how do implement it in a good way.

All the teachers but Mark report that they cooperate with their colleagues with the interpretation and operationalization of the competence aims. Mark says that he does much of this work on his own and that this is very challenging. The teachers report that the competence aims are vague and that it is easier to focus on the competence aims that are more specific. Lynn problematizes that the interpretation and operationalization of competence aims happen at local level, as this does not advocate a shared understanding across schools.

None of the teachers believes that their education has prepared them for assessing oral English. The training in assessment they have received during their practice periods have seemed random, and they report that one can not take for granted that assessment will be included in the practice periods. According to the teachers, a shared understanding of what to assess should be developed during the education.

The teachers do not feel confident that they are assessing correctly in oral English. Mark reports that he is not sure that other teachers agree with how he does assessment. Sarah says that assessment is spontaneous and subjective. Lynn believes that there is a need for clarity about what to assess, and that more rater training is needed to reduce differences in rater behavior.

5. Discussion

In the following, the emerging issues from the research findings will be discussed further. I argue that this can lead to differences in operationalization of the curriculum and that assessment of speaking skills is affected by the teacher's subjective opinions. Also, variability in assessment as a consequence of the rater/teacher will be discussed, and the challenges of using rating scales/forms and rater training to reduce the variability in scoring. The role of the teacher education will also be commented upon.

5.1 The localized nature of assessment in Norway

From the research findings in the present study, the level of autonomy that the teachers have through the Knowledge Promotion is problematized. Because of broad and vague competence aims and a lack of a shared understanding nationally of what to assess, the participants state that the teachers' subjective opinions is what counts the most when assessing oral English. However, they are quite sure that they have a shared understanding of what to assess within their school but emphasize that from one school to another, the understanding of what to assess differs. The differences in what teachers perceive as important to assess can lead to unfair assessment of the students. This is a downside of working with the curriculum and assessment on local and individual level.

5.1.2 Different operationalizations of the curriculum

The Knowledge Promotion is goal-driven and it is up to the teacher how to reach the competence aims (Munden & Sandhaug, 2017, p. 49). This gives the teacher much autonomy, which is an ideal in the current national curriculum. However, the participants find it difficult to use the competence aims as guides for assessing, and state that what teachers focus on and how they operationalize the competence aims may differ.

Differences in how the participants operationalize the curriculum was evident in their reflections about student performance in oral English. Lynn and Mark believed it was important that the students took part in deciding how to give an oral performance. In addition, Mark meant

it was important to prepare the students for an oral exam in year 10, where they are to give a presentation. Sarah, on the other hand, used conversations to assess oral English, as she believed that this would allow the students to show a higher level of competence.

Lynn usually formulates learning objectives from the broad competence aims, Sarah does not, as she believes that dividing the aims into smaller objectives is out of date. Even though the competence aims should not be divided in a way that loses the perspective of the competence aims (Norwegian Directorate for Education and Training, 2016, p. 2), they should be divided into smaller objetives as the competence aims are not intended to be communicated directly with students (Munden and Sandhaug, 2017, p. 51). The findings demonstrate that teachers have different opinions about whether or not the competence aims should be divided into learning objectives, and how it should be done. However, the novice teachers provide sound arguments for their choices. Differences in how the competence aims are operationalized is in itself not necessarily an issue because students can achieve competence in many different ways. The issue is that the participants find it challenging to work with the competence aims on a local and individual level, and that they express insecurity about how to use the competence aims as a guide for assessment. In addition, I argue that to what extent the teachers formulate learning objectives should be up to the teacher as this can vary depending on the class, a smaller group of students or individual students. However, what the findings indicate is that teachers interpret the competence aims in different ways, and that it is challenging to work with. It might be especially challenging for novice teachers as they lack experience.

A part of working with operationalization of the competence aims is to formulate criteria for the different levels of achievement. Even though Lynn, Mark, Sarah and Tia cooperate with their colleagues to understand the competence aims, it is important to note that there is no guarantee that teachers discuss grading criteria within or across schools (Nusche et al. 2011, p. 54). Lynn explicitly problematizes how the operationalization of the competence aims is left to the local level and individual teacher, Mark states that it is challenging to operationalize the competence aims individually as it relies heavily on his subjective opinions.

A common frame of reference consisting of criteria is available for English in lower secondary school but not for the mandatory English subject in upper secondary school. The criteria for assessing oral English in lower secondary school have vague formulations. When assessing the

students' ability to express themselves, in relation to their intonation and pronunciation, the teachers have to interpret the descriptions in the criteria. A low level of achievement is described as follows: "expresses oneself with a certain intonation and understandable pronunciation". A medium level of achievement is described as "expresses oneself with clear intonation and pronunciation", and finally to achieve a high level of achievement the student must "express oneself with good intonation and pronunciation" (Norwegian Directorate for Education and Training, 2017, p. 2, own translation). The teachers have to interpret the descriptions of intonation and pronunciation, and the question arises: what constitutes a *certain* intonation and understandable pronunciation, a *clear* intonation and pronunciation, and a *good* intonation and pronunciation? As Mark pointed out, the criteria are not specific, it is the listeners subjective opinion of what is good that counts. Thus, a student's intonation and pronunciation might be assessed differently depending on the teacher doing the assessment, which is problematic for the reliability and validity of the scores.

Vague descriptions of what constitutes different grades and levels of achievement seems to be a tendency in the curriculum: the Regulations to the Education Act (§3-4, 2009) uses "quite good, good and very good" to describe grade 3, 4 and 5 in lower and upper secondary school. It is up to the teacher to interpret and decide what these descriptions mean in practice. Seeing as the criteria and descriptions provided nationally are vague, one can assume that the criteria the local schools and individual teachers develop tend to look somewhat similar to the ones developed nationally as these are what they can look to as a reference. Again, the lack of experience might make it especially challenging for novice teachers to interpret the vague descriptions.

5.1.3 Subjective assessment

When looking at the research provided in this thesis, it is reasonable to assume that consistency of a score irrespective of the rater is not certain. This was problematized by the participants in the study as well. The localized nature of assessment in Norway leaves much room for subjective assessment, the participants in the present study believe that much of the assessment relies on the teacher's subjective opinion. The participants made the same point as Green (2014, p. 76-78): when assessing language ability, how do we know that we have the same

understanding of the concept, how can one prove the truth of their claim that one person has a better language ability than another. Luoma (2004, p. 1) also recognizes the challenges of assessing speaking as there are so many factors that influence the way we evaluate someone's oral proficiency.

Lynn argues that the measuring of students' level of achievement is not accurate: the Norwegian educational system cannot provide assurance that the assessment is fair and that the students are applying to higher education on the same terms. Hence, Lynn is not confident that the assessment is justified. Bachman and Palmer (2010, p 94) writes that the first person that needs to be convinced that the assessment is justified is the person doing the assessment:

> *If we lack the confidence that we can justify the consequences of the assessment use, the decisions to be made, how we will interpret the assessment records, or how we will analyze or score test takers' performances, then we are in no position to be able to convince other stakeholders.*

Assessment is a task that should be taken seriously by the teachers, schools and local authorities as it affects the students' lives. Backman and Palmer (2010, p. 92) writes that the uses of any given assessment will affect the lives of individual stakeholders, and as decision makers, teachers are accountable for the uses of a particular assessment. Further they argue that accountability involves being able to justify the decisions and consequences of the assessment (Bachman & Palmer, 2010, p. 92-93). As pointed out, the participants find it problematic to justify assessment of oral English because it is largely based on the teacher's subjective opinions. Even though an assessment is affected by the subjective opinions of the teacher, it does not necessarily mean that it is not a justified assessment: the teachers provide scores based on their professional judgements and they are trained to do so. Nevertheless, the participants in the present study seem to believe it is challenging to justify assessment of oral English as they are uncertain if they are doing it correctly because their own subjective opinions are largely involved.

5.1.4 Differences in what is assessed

The participants in the present study ranked the constructs differently and they did not have the same understanding of the constructs. In addition, Tia reported that what is assessed depends on the student's level of achievement in the subject. These findings show similar tendencies to Bøhn's (2016) study where the teachers' understanding of the constructs varied and some teachers stated that they assess students with different degrees of leniency and severity.

Differences in what aspects that are assessed can lead to raters providing a range of scores, and that they perceive the same performance in different ways (Orr, p. 143). Thus, it can threaten reliable and valid assessment. Not having clear guidelines for what to assess can be problematic. However, when given rating scales and descriptors of how to assess Kim's (2015) findings indicate that raters have different understandings of these guidelines. Confusion with the rating scales was especially evident with the novice raters, while the experienced raters usually got it right (Kim, 2015, p. 249).

Research also indicate that raters assess with different levels of severity and leniency (Bøhn 2016, Ang-Aw & Goh, 2011). Differences in severity/leniency when assessing is pointed out in the present study as well, as Lynn contemplates whether she is stricter than other teachers are. Such differences may lead to discrepancies both in classroom assessment and in oral exams. It is important to note that it is the teacher who decides the overall grade in the subject, a grade that is put on the student's diploma and is part of the admission to further education. Thus, just as oral exams, setting the overall achievement grade in the subject can be regarded as high-stakes and must be reliable and valid.

5.1.5 Teacher cognition

The present study indicates that there are differences in how the curriculum is operationalized and what is assessed. In addition, the findings indicate that the teacher's subjective opinion affects the assessment of oral English. A possible explanation for these differences is the localized nature of assessment in Norway. Another factor that should be taken into consideration is teacher cognition.

Language teacher cognition is complex: personal history and contextual factors are part of defining a teacher's conception of education (Borg, 2006, p. 283). Borg (2006, p. 283) states that language teachers have beliefs, knowledge, attitudes and assumptions about teaching, learning, subject matter, curricula, assessment and so on. Teacher cognition affects classroom practice and vice versa: contextual factors around and inside the classroom affect the teacher's cognition (Borg, 2006, p. 283). Therefore, teacher cognition can partly explain why teachers have different ways of interpreting the curriculum and different assessment practices. Sandvik (2013, p. 39) argues that the teachers' education, experiences and how they view learning is all part of the assessment context. As these factors vary from individual to individual, one can argue that the assessment context will vary depending on the teacher.

Both Borg and Sandvik explicitly mention the teachers' education. The teacher education should provide assurance of a certain competence. Education is a personal and contextual factor that, to some extent, can provide the pre-service language teachers with similar beliefs, knowledge and attitudes to subject matter, curricula and assessment. The participants in the present study reported that they believed it would be beneficial to develop a shared understanding of assessment during the teacher education. One of the participants even suggested a concrete example of how to get training that is more specific towards assessment during the teacher education: a visit from the department of education, or receiving similar training as raters. However, with similar training and education, and even with a curriculum with more detailed guidelines, the individual teacher will be a part of the assessment context.

5.2 Variability in assessment

Who did the scoring should not lead to variations in scoring (Black and William, 2012, p. 244). As the previous research presented in this thesis and my research findings suggest, this is not the case. The participants report uncertainty of the reliability and validity of the students' scores because they believe that much of the assessment of oral English relies on the teacher's subjective opinions. The participants pointed out that this might have to do with a lack of a shared understanding of what to assess. The previous research presented in this thesis shows that raters attend to non-criterion relevant information, compare candidates, and assess with different degrees of severity/leniency when assessing oral performances. In addition, the rater's background affects the assessment. This is especially interesting in the present study as it

focuses on novice teachers of English. Kim's (2015) studies indicate that novice raters are confused by rating scales more often than more experienced raters are. Teacher cognition plays a part in the teachers' practice and their cognitions are shaped by their lived experiences (Borg, 2006, p. 107), therefore, based on experience, it is reasonable to assume that the teacher's experience will affect the assessment.

Even though assessment of oral English might be especially challenging for novice teachers, as the previous research depicts, experienced raters face the same challenges. Variability in assessment caused by the rater is a threat to reliability and validity. Rater scales and rater training has been proposed as means to increase reliable and valid assessment, and will be discussed in the following paragraphs.

5.2.1 Rating scales/forms as tools for assessment

Munden and Sandhaug (2017, p. 49) state that it is unusual to have a competence plan like the Knowledge Promotion without giving guidelines and criteria at the same time. The participants in the present study report that they believe a common frame of reference is necessary to secure more reliable and valid assessment. However, they also said that they were uncertain how this could be done in a good way. Sarah stated that she did not have much faith in scoring rubrics and that video clips demonstrating the different levels of achievement in oral English might work better.

Rubrics with criteria for the different levels of achievement is available for the English subject in lower secondary school. The common frame of reference is meant to serve as a guide for assessment but such a guide is not provided for the mandatory English subject in upper secondary school. As the purpose of the common frame of reference is to guide assessment and provide fair assessment nationally, it is peculiar that one does not exist for the mandatory English subject in upper secondary school.

However, as the findings in the present study show, Mark does not use this tool much, and Lynn states that even when provided with criteria for assessment, it is difficult to know what weight should be given to each criteria and if teachers interpret the criteria in the same way. Sadler (2010, p. 545) claims that the majority of criteria are abstract concepts without sharp

boundaries, thus they have to become known and formed by individuals, and shared in social or professional contexts. Even though Sadler (2010) focuses on criteria in higher education, I believe that some of the points he makes are transferable to a Norwegian educational context because what Sadler says about criteria can apply to criteria in every level of education.

According to Sadler (2010, p. 545) a challenge with criteria as concepts is that particular terms can mean different things to different teachers, even with a fixed list of criteria there can still be differences in the teachers' interpretations of the same criteria, which can affect the consistency of assessment. Nevertheless, Sadler (2010, p. 541) argues that in order to provide explanations for their judgements, teachers invariably make use of criteria and invoke which criteria are salient to a given judgement. Even if teachers focus on slightly different criteria when assessing oral English and interpret the criteria differently, they will have to explain their judgements, and through the Knowledge Promotion, the teachers are trusted to be able to conduct assessment in a responsible and fair manner.

Sarah argues that ranking students in detailed scoring rubrics is problematic and will not lead to more reliable and valid assessment. She states that assessment is subjective and immediate, and whether or not a student performance feels like a grade four or five depends on her as a rater. Ang-Aw and Goh (2011, p. 43 – 44) pose similar challenges: if the descriptors do not clearly match a candidate's performance, raters can end up feeling if a score is too high or too low for a candidate. Nevertheless, Luoma (2004), Bøhn (2016), and Orr (2001) advocate rating scales in the assessment of oral skills, and argue that it will strengthen the reliability and validity of the assessment, reduce inter-candidate comparison and subjectivity in assessment.

Sadler (2010, p. 548) looks at it from a different point of view, arguing that because of the focus on specific criteria rather than quality, the use of rubrics and criteria might inhibit a full-bodied concept of quality. Bøhn (2016, p. 69) also makes the point that rating scales are not the universal solution to increased quality in educational contexts. Rating scales may not be able to capture the complexities of what is to be tested, and a less formalized structure opens up for integrating valuable learning outcomes not necessarily specified in the English subject curriculum but in the Core Curriculum (Bøhn, 2016, p. 69). To back up his statement Bøhn (2016, p. 69) refers to Baird et al. (2014, p. 82), they argue that assessment is intrinsically linked to teaching and learning and should not be treated in isolation. However, based on his findings,

Bøhn (2016, p. 70) concludes that considering the threats to reliability and validity, a national rating scale that can be locally adapted should be included in the scoring of performance in English exams in upper secondary.

5.3 Teacher education and rater training

The present study does not argue that teacher education should prepare pre-service teachers fully for the profession. As pointed out by Sarah, one may never feel one hundred percent confident when assessing oral English. However, it is troubling that none of the teachers participating in the present study found their education even slightly close to satisfactory in regard to assessing oral English. In addition, it is concerning that the novice teachers report that who gets to conduct assessment during their practice periods is random.

Even though the teacher education is a large part of what provides the theoretical and practical knowledge base for how teachers assess oral English, Borg (2006, p. 81) claims that the transfer of knowledge and beliefs from teacher education to classroom practice is not linear. A range of contextual factors can outweigh principles learned during teacher education (Borg, 2006, p. 81). Nevertheless, as both formative and summative assessment are crucial parts of the teacher profession, the teacher education should be able to assure a certain level of competence in assessment. Tia reports that during her education there was much focus on theory about formative and summative assessment but that they did not get many opportunities to practice how to assess. The findings indicate that the focus on theory about assessment is not enough to prepare pre-service teachers for assessing oral English. I argue that efforts should be made to ensure that pre-service teachers get the opportunities to work with assessment in oral English during their practice periods.

Seeing as subjectivity is an issue of concern according to the novice teachers, the wish for a shared understanding of assessment in oral English is understandable. Nonetheless, it does not seem likely that a shared understanding will become a certainty as a result of rating scales, because the teachers have to interpret the scales as well. The autonomy given to teachers through the Knowledge Promotion might be especially challenging for novice teachers, they do not have much experience to base their teaching and assessment on. Also, they do not have much practice operationalizing the national curriculum and English subject curriculum.

Lynn suggested that, during their education, they could have received the same training in assessment as raters do. However, if they were to receive this it would only be rater training in relation to the written English exam. A rater scale and rater training are provided in written English exams in upper secondary school but not in oral English exams (Bøhn, 2016, p. 69). Knowing how challenging it is to assess oral performances it is, as Bøhn (2016, p. 69) puts it puzzling that rating scales and rater training is not provided for the oral English exam. Even with similar rater training, entirely reliable assessment is not certain (Ang-Aw & Goh, 2011, p. 44). Nevertheless, Kim's (2015, p. 254 – 256) study showed that the raters displayed different levels of rating performance after receiving rater training, and that group level training will lead to more dependable raters and more reliable ratings.

6. Concluding remarks

In this thesis I have discussed challenges with assessing oral English in lower and upper secondary school. Through the Knowledge Promotion, Norwegian teachers have autonomy to base their teaching and assessment on their interpretations of the national curriculum and English subject curriculum. Thus, teachers in Norway are trusted to teach and assess in a way that helps the students reach the competence aims stated in the curriculum. However, as the present study shows, novice teachers find it challenging to base their assessment on the competence aims as they are uncertain that teachers interpret the competence aims in similar ways, and uncertain if their interpretations are correct.

As the previous research presented in the thesis demonstrate, it is challenging to assess oral skills as there are so many factors that can influence our impression. In the thesis, the focus has mainly been on how the rater influence the scoring of students' oral performances. From the research findings, I argue that differences in what teachers perceive as salient to assess and different interpretations of the competence aims and scoring criteria, can threaten reliable and valid assessment in oral English. The teacher's subjective opinions seem to affect the scoring of students' oral performances to a large extent, and this should be given serious attention by teachers and national educational authorities.

The participants in the present study ranked the constructs pronunciation, intonation, communication, content, vocabulary and grammar differently, and the findings indicate that they do not have the same understanding of the constructs. However, all of the novice teachers ranked communication as the most important construct to assess. The importance of the construct content, on the other hand, varied among the novice teachers. I argue that differences in how the constructs are ranked can affect the assessment of oral English: the same student performance can be scored differently depending on the teacher and how the teacher weigh the various constructs.

The novice teachers report a range of challenges with assessing oral English. They argue that there is no shared understanding of assessment in oral English across schools, and pointed to the consequences this might have: differences in teaching and assessment, and variances in the student's overall achievement grade which may affect their admission to upper secondary school or higher education. The role of the teacher's subjective opinions in assessing oral

English seems to be the biggest concern for the novice teachers and they do not seem to be comfortable that their subjective opinions play such a big part in the assessment of their students. Overall, the novice teachers are not confident that the oral assessment in English is reliable and valid.

Rater training and rater forms/scales have been discussed as means to secure more reliable and valid assessment. The novice teachers believe that there is a need for a common frame of reference for assessing oral English but they are not sure how this can be done in a good way. Scoring rubrics, as they are formulated in lower secondary, was not perceived to be a sufficient solution by the novice teachers as they did not believe that this would help reduce the effects of the teacher's subjective opinion. However, based on the theory and previous research presented, I argue that there is a need for a common frame of reference for assessing oral English in upper secondary school as a contribution to developing a shared understanding of assessment. The common frame of reference provided for lower secondary is meant to secure fair assessment and I find it puzzling that one is not provided for the mandatory English subject in upper secondary. At the same time, I argue that there is a need for sufficient training and discussions about rater scales to ensure similar interpretations of the criteria for the different levels of achievement, and the weight given to each criteria.

None of the novice teachers believed that their teacher education had prepared them for assessing oral English. It is my belief that most of the skills needed in the teacher profession comes from experience from working as a teacher, and not the teacher education. However, knowing how challenging it is to assess speaking skills, it is troubling that none of the teachers found their education somewhat close to sufficient in regard to assessing oral English.

As long as speaking skills are assessed by human raters there will be a certain degree of subjectivity in assessment. The teachers are part of the assessment context and their assessment is affected by them as individuals. I believe that it is important that teachers are aware of how they affect the assessment of oral English, and that teachers strive to ensure reliable and valid assessment. The present study shows that novice teachers are very much aware of the challenges with subjectivity in assessing oral English, and that they are concerned with oral assessment being as reliable and valid as possible.

My motivation and interest for writing a master's thesis about oral assessment in English, has been based on my own experiences from working as an English teacher. I find oral assessment in the English subject challenging, and I wanted to explore how other novice English teachers experience assessing oral English. From reading literature and previous research, and from conduction my own research on the topic, I have gained knowledge about how complex the task of assessing speaking skills is. In addition, I have learnt from doing my own research. If I were to do it again, I would have narrowed the focus area down to summative *or* formative assessment, as I believe that this would have made the topic of the interview more precise for the participants and myself.

Based on the theory and empirical research findings presented in this thesis, I suggest that further research should address ways to secure a shared understanding of how and what to assess in oral English. Furthermore, the assessment of oral English is part of setting the students' overall achievement grades, and therefore how the teachers assess oral English in the classroom should be given attention in future research. Another relevant research area that should be investigated further, is how the teacher training education work with assessment of speaking skills, and prepare the pre-service teachers for assessing oral English.

Bibliography

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Ang-Aw, H. T., & Goh, C.C.M. (2011). Understanding Discrepancies in Rater Judgement on National-Level Oral Examination Tasks. *RELF Journal, 42,* 31-51. Retrieved from https://journals-sagepub-com.ezproxy.inn.no/doi/pdf/10.1177/0033688210390226

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice.* Oxford, UK: Oxford University Press.

Baird, J.-A., Hopfenbeck, T. N., Newton, P., Stobart, G., & Steen-Utheim, A. T. (2014). *State of The Field Review: Assessment and Learning.* (Case number 13/4697). Oslo: Knowledge Centre for Education.

Biggam, J. (2015). *Succeeding with your Master's Dissertation. A step-by-step handbook.* New York: Open University Press.

Black, P., & Wiliam, D. (2012). The Reliability of Assessments. In J. Gardner (Ed.), *Assessment and Learning* (pp. 243-263). London: Sage.

Borg, S. (2006). *Teacher Cognition and Language Education. Research and Practice.* London: Continuum.

Brindley, G. (1991). Defining Language Ability: The Criteria for Criteria. In S. Anivan (Ed.), *Current Developments in Language Testing: Anthology series 25.* Singapore: Southeast Asian Ministers of Education Organization (SEAMEO) Regional Language Centre. Retrieved from: https://eric.ed.gov/?id=ED365150

Brinkmann, S., & Kvale, S. (2015). *Interviews: Learning the craft of qualitative research interviewing.*Thousand Oaks: Sage.

Brown, A., Iwashita, N., & McNamara, T. (2005) *An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks* (TOEFL Monograph Series, MS-29). Princeton, NJ: Educational Testing Service.

Bøhn, H. (2015). Assessing Spoken EFR Without a Common Rating Scale: Norwegian EFL Teachers' Conceptions of Construct. *Sage Open*. October-December, 1-12. Retrieved from https://www.duo.uio.no/handle/10852/53230

Bøhn, H. (2016). *What is to be assessed? Teachers' understanding of constructs in an oral English examination in Norway.* (Doctoral Thesis). Retrieved from https://www.duo.uio.no/bitstream/handle/10852/53229/PhD-Boehn-DUO.pdf?sequence=1&isAllowed=y

Creswell, J. (1997). *Qualitative Enquiry and Research Design.* London: Sage.

Dysthe, O. (2008). *Klasseromsvurdering og læring.* Retrieved from https://www.udir.no/globalassets/filer/vurdering/vfl/andre-dokumenter/felles/olga-dyste-bedre-skole-08.pdf

Fulcher, G., & Davidson, F. (2007). Language Testing and Assessment. Oxford: Routledge.

Green, A. (2014). *Exploring Language Assessment and Testing. Language in Action.* New York: Routledge.

Hartberg, E.W., Dobson, S. & Gran. L. (2012). *Feedback i skolen.* Oslo: Gyldendal Akademisk.

Johannessen, S. L. (2018). *Oral assessment in the English subject. Teachers' understandings*

of what to assess. (Master's thesis, Inland University). Retrieved from http://hdl.handle.net/11250/2560214

Kim, H. J. (2015). A Qualitative Analysis of Rater Behavior on an L2 Speaking Assessment. *Language Assessment Quarterly, 12:3, 239-261.* Retrieved from https://www-tandfonline-com.ezproxy.inn.no/doi/pdf/10.1080/15434303.2015.1049353?needAccess=true

Kvale, S. (2007). *Doing Interviews.* London: Sage Publications Ltd.

Luoma, S. (2004). *Assessing Speaking.* New York: Cambridge University Press.

McNamara, T.F. (1996). *Measuring Second Language Performance.* London: Longman.

Miles, M. & Huberman, A. (1984). *Qualitative Data Analysis: An Expanded Source Book.* London: Sage.

Ministry of Education and Research. (2004). *Kultur for læring.* (Meld. St. 31, 2003-2004). Retrieved from https://www.regjeringen.no/contentassets/988cdb018ac24eb0a0cf95943e6cdb61/no/pdfs/stm200320040030000dddpdfs.pdf

Munden, J. & Sandhaug, C. (2017). *Engelsk for secondary school.* Oslo: Gyldendal Akademisk

National Institute of Education. (1975). *Teaching as Clinical Information Processing.* Washington DC: National Institute of Education.

Norwegian Directorate for Education and Training. (2012). *Framework for Basic Skills.* Retrieved from https://www.udir.no/contentassets/fd2d6bfbf2364e1c98b73e030119bd38/framework_for_basic_skills.pdf

Norwegian Directorate for Education and Training. (2013). *English subject curriculum.* Retrieved from http://data.udir.no/kl06/rest_/ENG1-03.pdf?lang=eng

Norwegian Directorate for Education and Training. (2015). Fremmedspråk – veiledning til læreplanen. Retrieved from https://www.udir.no/laring-og-trivsel/lareplanverket/veiledning-til-lp/fremmedsprak---veiledning-til-lareplanen/3-praktiske-eksempler/#vurdering

Norwegian Directorate for Education and Training. (2015). *Fire prinsipper for god underveisvurdering.* Retrieved from https://www.udir.no/laring-og-trivsel/vurdering/om-vurdering/underveisvurdering/

Norwegian Directorate for Education and Training. (2016). *Å forstå kompetanse*. Retrieved from https://www.udir.no/laring-og-trivsel/lareplanverket/forsta-kompetanse/

Norwegian Directorate for Education and Training. (2017). *Engelsk: kjenneteikn på måloppnåing. Rettleiande nasjonale kjenneteikn på måloppnåing for standpunktvurdering etter 10.trinn.* Retrieved from https://www.udir.no/laring-og-trivsel/vurdering/sluttvurdering/Engelsk-kjenneteikn-pa-maloppnaing/

Norwegian Directorate for Education and Training. (2019). *Nye læreplaner i grunnskolen og gjennomgående fag i vgo – hva skjer når?* Retrieved from https://www.udir.no/laring-og-trivsel/lareplanverket/fagfornyelsen/hva-skjer-nar-i-fornyelsen-av-fagene/

Nusche, D., Earl, L., Maxwell, W.,& Shewbridge, C. (2011). *OECD Reviews of Evaluation and Assessment in Education: Norway*. Retrieved from https://www.oecd.org/norway/48632032.pdf

Orr, M. (2002). The FCE speaking test: Using Rater Reports to Help Interpret Test Scores. *System, 30,* 143-154. Retrieved from https://ac-els-cdn-com.ezproxy.inn.no/S0346251X02000027/1-s2.0-S0346251X02000027-

main.pdf?_tid=c16cdc5c-7e82-455c-bb29-777413a51839&acdnat=1555501455_819e4873fc71ff75b92331ed6bd77ede

Pollitt A., & Murray, N.L. (1996). What Raters Really Pay Attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th language research testing colloquium, Cambridge.* Cambridge: Cambridge University Press.

Regulations to the Education Act, FOR-2009-07-01-964. (2009). Retrieved from https://lovdata.no/dokument/SF/forskrift/2006-06-23-724

Roberts, C. (1997). Transcribing Talk: Issues of Representation. *TESOL Quarterly 31*(1), 167-172. Retrieved from https://www.jstor.org/stable/3587983?seq=1#page_scan_tab_contents

Sadler, D. R. (2010). Beyond Feedback: Developing Student Capability in Complex Appraisal. *Assessment & Evaluation in Higher Education, 35*(5), 535-550. Retrieved from https://www-tandfonline-com.ezproxy.inn.no/doi/pdf/10.1080/02602930903541015?needAccess=true

Sandvik, L.V. (2013) Perspektiver på individuell vurdering i skolen. In Sandvik, L. V., Buland, T. (Eds.), *Vurdering i skolen. Operasjonaliseringer og praksiser.* Retrieved from https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/2013/fivis2.pdf

Sato, T. (2012). The Contribution of Test Takers' Speech Content to Scores on an English Oral Proficiency Test. *Language Testing, 29*(2), 223-241.doi:10.1177/0265532211421162.

The Norwegian Universities and Colleges Admission Service (2013). *Ordinær kvote.* Retrieved from https://www.samordnaopptak.no/info/opptak/opptakskvoter/ordinerkvote.html

Weir, C. (2005). *Language Testing and Validation: An Evidence-Based Approach.*

Basingstoke: Palgrave MacMillan.

Wolcott, H. (1994). *Transforming Qualitative Data: Description, Analysis, Interpretation.* Thousand Oaks, CA: Sage.

Appendices

Appendix 1

Competence aims after Year 10 in the English subject curriculum identified as relevant to oral assessment:

Language learning

- use different situations, working methods and learning strategies to develop one's English-language skills
- comment on own work in learning English
- identify significant linguistic similarities and differences between English and one's native language and use this knowledge in one's own language learning
- select different digital resources and other aids and use them in an independent manner in own language learning

Oral communication

- choose and use different listening and speaking strategies that are suitable for the purpose
- understand and use a general vocabulary related to different topics
- demonstrate the ability to distinguish positively and negatively loaded expressions referring to individuals and groups
- understand the main content and details of different types of oral texts on different topics
- listen to and understand variations of English from different authentic situations
- express oneself fluently and coherently, suited to the purpose and situation
- express and justify own opinions about different topics
- introduce, maintain and terminate conversations on different topics by asking questions and following up on input
- use the central patterns for pronunciation, intonation, word inflection and different types of sentences in communication
- understand and use different numerical expressions and other kinds of data in communication

Culture, society and literature

- discuss and elaborate on the way people live and how they socialise in Great Britain, USA and other English-speaking countries and Norway
- explain features of history and geography in Great Britain and the USA
- discuss and elaborate on different types of English literature from English-speaking countries
- describe and reflect on the situation of indigenous peoples in English-speaking countries
- create, communicate and converse about own texts inspired by English literature, films and cultural forms of expression
- communicate and converse about contemporary and academic topics

(Norwegian Directorate for Education and Training, 2013, p. 8-9).

Appendix 2

Competence aims, after Vg1 – programmes for general studies and Vg2 – vocational education programmes, in the English subject curriculum identified as relevant to oral assessment:

Language learning

- evaluate and use different situations, working methods and learning strategies to further develop one's English-language skills
- evaluate own progress in learning English
- evaluate different digital resources and other aids critically and independently, and use them in own language learning

Oral communication

- evaluate and use suitable listening and speaking strategies adapted for the purpose and the situation
- understand and use a wide general vocabulary and an academic vocabulary related to his/her own education programme
- understand the main content and details of different types of oral texts about general and academic topics related to one's education programme
- listen to and understand social and geographic variations of English from authentic situations
- express oneself fluently and coherently in a detailed and precise manner suited to the purpose and situation
- introduce, maintain and terminate conversations and discussions about general and academic topics related to one's education programme
- use patterns for pronunciation, intonation, word inflection and various types of sentences in communication
- interpret and use technical and mathematical information in communication

Culture, society and literature

- discuss and elaborate on culture and social conditions in several English-speaking countries
- present and discuss current news items from English language sources
- discuss and elaborate on the growth of English as a universal language
- discuss and elaborate on different types of English language literary texts from different parts of the world
- discuss and elaborate on English language films and other forms of cultural expressions from different media
- discuss and elaborate on texts by and about indigenous peoples in English-speaking countries
- select an in-depth study topic within one's education programme and present this (Norwegian Directorate for Education and Training, 2013, p. 9-10).

Appendix 3

Intervjuguide

Utdanning:

Ferdigutdannet:

Studiepoeng i engelsk:

Arbeider på klassetrinn:

Har arbeidet på klassetrinn:

Erfaring med eksamen i engelsk:

Hvordan legger du opp til muntlige vurderinger gjennom skoleåret? (både formativ og summativ vurdering)

- Be om begrunnelse for hvordan muntlige vurderinger blir lagt opp.

Hvilke kompetansemål vil du si er relevante når du vurderer elevens kompetanse i muntlig engelsk?

Har du mer fokus på visse kompetansemål?

- Hvorfor?

Hvordan jobber du med kompetansemålene i planleggingen og gjennomføringen av muntlig vurdering?

Hvordan arbeider du med tolkning og operasjonalisering av kompetansemålene i engelskfaget?

Oppfølgingsspørsmål til spm. 2. og 3: Hvorfor velger du å legge opp til muntlig vurdering på denne måten? Og hvorfor velger du å jobbe med kompetansemålene på denne måten?

Står du fritt til å legge opp vurderingen slik du selv ønsker, eller er det et samarbeid/føringer fra skolen sin side?

Deltakeren blir gitt en liste med disse komponentene. Deltaker blir spurt om å rangere hva de anser som viktig i vurderingen av elevens muntlige kompetanse. Deretter må deltaker begrunne hvorfor. Hva anser du som viktig i vurderingen av elevens muntlige kompetanse?

- Innhold
- Flyt
- Grammatikk
- Ordforråd
- Uttale
- Kommunikasjon

Har dere en felles referanseramme for vurdering av muntlig engelsk enten utviklet nasjonalt eller på/av din arbeidsplass?

Hva er dine tanker om å ha en felles nasjonal referanseramme for vurdering av muntlig engelsk?

Kan du beskrive om/hvordan lektor/lærerutdanningen har forberedt de til å jobbe med vurdering av muntlig kompetanse i engelskfaget?

Hvordan opplever du din kompetanse når det kommer til vurdering av muntlig engelsk?

Appendix 4

Interview

ME: Kan du beskrive om/hvordan lektor/lærerutdanningen har forberedt de til å jobbe med vurdering av muntlig kompetanse i engelskfaget?

L: Jeg vil si at det ikke har forberedt meg til å jobbe. Vi hadde noen øvelser, men det blir ikke … er ikke virkelighetsnært i det hele tatt. Og uten å ha god kjennskap til kompetansemålene og forstå hva som ligger i kompetansemålene … uten å sitte med akkurat det arbeidet som man gjør som ny lærer og må prøve å nøste opp hva er det som egentlig ligger her, så kan man heller ikke vite hva som skal vektlegges i muntlig vurdering.

ME: var praksis med på å hjelpe deg på veien?

L: Nei. Det vil jeg jo ikke si. Jeg begynte på grunnskolelærerutdanning som var 5-10, vi hadde jo vurdering når vi var i praksis, men det blir jo også med veiledning av andre tolkende sjeler da som sitter og prøver å tolke det samme materialet. Så jeg syns jo at det blir tynn suppe, sånn til slutt. Og i undervisning eller i forelesning så … vet ikke, syns ikke det blir så troverdig … det er jo mange som foreleser som aldri har vært ute i skolen sjøl, så da har jeg ikke noe tillit til de de sier. Så det blir veldig … jeg syns man blir veldig overlatt til seg sjøl. Å prøve å finne ut av det kryptiske innholdet og de kryptiske kriteriene man får. Så … nei.

ME: er det noe spesifikt du teker kunne ha vært med i utdanninga som kunne ha hjulpet deg?

L: Ja, jeg tenker jo at på ett eller annet tidspunkt så kunne jeg vel ønske at det du måtte velge enda mere retning. Nå har det jo for så vidt blitt det i de nye utdanningene, men å ha en 5-10 tilnærming det blir for stort sprik. Så jeg tenker at på et eller annet tidspunkt så burde man ha valgt enda mer retning og fått en mer konkret opplæring i vurdering. Det burde ha vært noen fra sentralt som kom og ja … vi kunne ha fått samme skolering som sensorer gjør, samme kurset for eksempel.

Appendix 5

Vil du delta i forskningsprosjektet

”How novice English teachers assess their competence in the English subject with regard to assessment, and how they work with assessment in oral English.”

Dette er et spørsmål til deg om å delta i et forskningsprosjekt hvor formålet er å undersøke nyutdannede engelsklæreres opplevde kompetanse når det kommer til vurdering i engelskfaget, og hvordan lærere arbeider med vurdering i muntlig engelsk. I dette skriver gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg.

Formål

Formålet med prosjektet er å undersøke nyutdannede engelsklæreres opplevde kompetanse i vurderingen av muntlig engelsk, og hvordan de arbeider med muntlig vurdering i engelskfaget. Prosjektet er en masteroppgave.

Hvem er ansvarlig for forskningsprosjektet?

Tale Gabriella Vesterlid (student ved høgskolen i Innlandet) er ansvarlig for forskningsprosjektet.

Hvorfor får du spørsmål om å delta?

Prosjektet bruker et bekvemmelighetsutvalgt. Det vil si at det har blitt valgt personer som vil være nyttige i undersøkelsen. Jeg spør om du kan delta fordi du er nyutdannet engelsklærer og jobber som engelsklærer.

Hva innebærer det for deg å delta?

Du vil bli bedt om å stille til en intervju sammen med forskeren. Intervjuet vil bli tatt opp og senere transkribert. Etter at intervjuet har blitt transkribert vil lydopptaket bli slettet. Intervjuet vil ta mellom 15 - 30 minutter.

Det er frivillig å delta

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

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger

Vi vil bare bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket. Juliet Munden: dosent ved høgskolen i Innlandet og veileder til dette prosjektet vil ha tilgang til de transkriberte intervjuene. Det vil ikke være behov for å samle inn navn eller kontaktopplysninger til dette prosjektet. Du som deltaker vil ikke kunne gjenkjennes i publikasjonen.

Hva skjer med opplysningene dine når vi avslutter forskningsprosjektet?

Prosjektet skal etter planen avsluttes 15.05.2019. Lydopptakene av intervjuene sletter umiddelbart etter at intervjuene er transkribert.

Hva gir oss rett til å behandle personopplysninger om deg?

Vi behandler opplysninger om deg basert på ditt samtykke. På oppdrag fra Tale Gabriella Vesterlid har NSD – Norsk senter for forskningsdata AS vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

Hvor kan jeg finne ut mer?

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

- Høgskolen i Innlandet ved Tale Gabriella Vesterlid
  Telefon: [redacted] e-mail: [redacted]

- Høgskolen i Innlandet ved Juliet Munden
  Telefon: [redacted] e-mail: [redacted]

- Vårt personvernombud: NSD – Norsk senter for forskningsdata AS, på epost (personverntjenester@nsd.no) eller telefon: 55 58 21 17.

Appendix 6

Det innsendte meldeskjemaet med referansekode 214423 er nå vurdert av NSD.

Følgende vurdering er gitt:

Det er vår vurdering at behandlingen av personopplysninger i prosjektet vil være i samsvar med personvernlovgivningen så fremt den gjennomføres i tråd med det som er dokumentert i meldeskjemaet med vedlegg, samt i meldingsdialogen mellom innmelder og NSD, den 22.10.18. Behandlingen kan starte.

MELD ENDRINGER
Dersom behandlingen av personopplysninger endrer seg, kan det være nødvendig å melde dette til NSD ved å oppdatere meldeskjemaet. På våre nettsider informerer vi om hvilke endringer som må meldes. Vent på svar før endringer gjennomføres.

TYPE OPPLYSNINGER OG VARIGHET
Prosjektet vil behandle alminnelige kategorier av personopplysninger frem til 15.05.19.

LOVLIG GRUNNLAG
Prosjektet vil innhente samtykke fra de registrerte til behandlingen av personopplysninger. Vår vurdering er at prosjektet legger opp til et samtykke i samsvar med kravene i art. 4 og 7, ved at det er en frivillig, spesifikk, informert og utvetydig bekreftelse som kan dokumenteres, og som den registrerte kan trekke tilbake. Lovlig grunnlag for behandlingen vil dermed være den registrertes samtykke, jf. personvernforordningen art. 6 nr. 1 bokstav a.
NSD ber om at samtykkeslippen revideres slik at den kun inneholder avkrysningsalternativer som er aktuelle for studien.

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

DE REGISTRERTES RETTIGHETER
De registrerte vil ha følgende rettigheter i prosjektet: åpenhet (art. 12), informasjon (art. 13), innsyn (art. 15), retting (art. 16), sletting (art. 17), begrensning (art. 18), underretning (art. 19), dataportabilitet (art. 20). Rettighetene etter art. 15-20 gjelder så lenge den registrerte er mulig å identifisere i datamaterialet.
NSD vurderer at informasjonen om behandlingen som de registrerte vil motta oppfyller lovens krav til form og innhold, jf. art. 12.1 og art. 13.
Vi minner om at hvis en registrert tar kontakt om sine rettigheter, har behandlingsansvarlig institusjon plikt til å svare innen en måned.

FØLG DIN INSTITUSJONS RETNINGSLINJER
NSD legger til grunn at behandlingen oppfyller kravene i personvernforordningen om riktighet (art. 5.1 d), integritet og konfidensialitet (art. 5.1. f) og sikkerhet (art. 32).
For å forsikre dere om at kravene oppfylles, må dere følge interne retningslinjer og/eller rådføre dere med behandlingsansvarlig institusjon.

OPPFØLGING AV PROSJEKTET
NSD vil følge opp behandlingen ved planlagt prosjektslutt for å avklare om behandlingen av personopplysningene er avsluttet.

Lykke til med prosjektet!

Kontaktperson hos NSD: Kjersti Haugstvedt
Tlf. Personverntjenester: 55 58 21 17 (tast 1)