Kine Brattland

Threats to validity of oral assessment in English

Master's thesis in English and Foreign Language Education

Trondheim, May 2018

Norwegian University of Science and Technology
Faculty of Social and Educational Sciences
Department of Teacher Education

NTNU
Norwegian University of
Science and Technology

Abstract

Despite the increased attention directed towards the quality of educational assessment, research display that few studies have addressed assessment of oral competence in specific. This sheds light on the urgency of devoting attention to assessment of oral competence, to gain more knowledge about oral assessment in general, in addition to developing knowledge of the validity aspect of oral assessment, which is argued to be the most significant quality of assessment. Through a qualitative case study, this research project investigates what threats to assessment of oral competence in English that are revealed through teachers' reasoning about own assessment literacy. The findings in this study identify five major threats to validity: teacher insecurity, contextual dilemmas, variance in teacher assessment literacy, unclear guidelines for teacher classroom assessment, and poorly defined constructs. From a system approach, this thesis discusses the influence of these threats on current understanding of assessment of oral competence in English, and what the threats imply about the curriculum's and teachers' influence on assessment. The thesis suggests an expansion of the current theoretical framework of the chain of validation, and advocates for directing the attention towards teachers' role in the assessment process, and to the aspects that influence teachers' interpretations and decisions.

Sammendrag

Til tross for den økte oppmerksomheten på kvaliteten av vurdering i utdanning viser forskning at få studer har tatt for seg vurdering av muntlig kompetanse. Dette belyser nødvendigheten av å rette oppmerksomhet mot muntlig vurdering, for å få mer kunnskap om vurdering av muntlig kompetanse generelt, i tillegg til å utvikle kunnskap om aspekter relatert til validitet i muntlig vurdering, som er hevdet å være det viktigste kvalitetsaspektet i vurdering. Gjennom en kvalitativ case-studie undersøker dette forskningsprosjektet hvilke trusler mot validitet i vurdering av muntlig kompetanse i Engelsk som kommer til syne i læreres resonnering om egen vurderingskyndighet. Funnene i studien identifiserer fem hovedtrusler mot validitet: læreres usikkerhet, kontekstuelle dilemmaer, variasjon i læreres vurderingskompetanse, uklare retningslinjer for læreres vurderingsarbeid i klasserommet og dårlig definerte konstrukter. Fra et systemperspektiv diskuterer denne masteroppgaven innflytelsen av disse trusler på nåværende forståelse av vurdering av muntlig kompetanse i Engelsk, og hva truslene innebærer for læreplanens- og lærerens innflytelse på vurdering. Denne masteroppgaven foreslår en utvidelse av det nåværende teoretiske rammeverket om validitetskjeden, og argumenterer for å rette fokus mot lærerens rolle i vurderingsprosessen og mot aspektene som påvirker læreres tolkninger og beslutninger.

Acknowledgments

Writing this master's thesis has given me valuable experiences and knowledge, not just about assessment of oral competence, but also about myself and my own limitations. Being a full-time master student in addition to working as a teacher has had its challenges, but this combination has also taught me what I can achieve if I challenge myself to go one step further.

I want to take this opportunity to thank everyone that has motivated me and given me feedback and support throughout the process. Firstly, I want to give a special thanks to my main supervisor, Henning Fjørtoft, for all your motivation, help and guidance during the research process. Secondly, I am also grateful for the constructive feedback provided by my co-supervisor, Lise Vikan Sandvik, and for inviting me to be a part of the SKUV-project. Thirdly, I want to thank the research participants, for taking the time to be a part of this study, and for their contribution. Last, but not least, I want to thank my husband and my daughter for their patience and support.

Trondheim, May 2018,

Kine Brattland

Table of contents

Index of tables and figures

1. Introduction

The increased attention devoted to the quality of the educational system in Norway has led to a focus on the role of assessment as a means for learning and development. Assessment of oral skills has especially gained focus in the Norwegian curriculum during the last 20 years and is defined as one of the five basic skills in the curriculum (Fjørtoft, 2016, p. 119; Norwegian Directorate of Education and Training, 2006). However, research on the field displays that assessment of oral competence has received less attention from researchers than for example assessment of written competence (Svenkerud, Klette & Hertzberg, 2012, p. 36). This underrepresentation is also visible in new international handbooks of assessment, where oral assessment is not addressed in specific (Andrae & Cizek, 2010; Fjørtoft, 2016; McMillan, 2013). This display both the relevance of gaining more knowledge on oral assessment and the urgency for research that directs attention to assessment of oral competence in specific, to develop current knowledge on the field.

The consequence of the inadequacy in current research is reflected in teachers' assessment of students' oral competence where research displays a limited coherence between aims, activities, and assessment of oral skills (Fjørtoft, 2016, p. 119). Some of these challenges are addressed in a research project that investigated the teaching of oral skills in Norwegian primary schools, which demonstrated that the participating teachers mainly assessed oral skills through presentations and that students received little formative feedback on how to improve their performance. Instead, when assessing oral competence, the teachers mainly focused on creating good learning environments and removing pressure from students, by motivating them through positive encouragement (Svenkerud, Klette & Hertzberg, 2012, p. 44). Despite that these factors are important baselines for the ability to learn, it reveals how contextual challenges can influence teachers' prioritization, and lead to trade-offs that can weaken the validity of the assessment (McMillan, 2013, p.102).

The influence of the social context on oral assessment also demonstrates how factors related to human influence are specifically influential on the assessment (Pøitz & Borgen, 2010, p. 115). Nevertheless, research on the field have not yet adequately addressed what happens in practice when teachers assess students' competence (Brookhart, 2005, p.11; Stiggins & Conclin, 1992, p. 20). The lack of focus on these areas is suggested as a reason for the

inconsistency between educational principles and how classroom assessment is conducted. Even though the curriculum defines what counts as knowledge, teachers are interpreters of the curriculum, mediators of knowledge and assessors of that knowledge.

When teachers discuss evidence of student learning, they evaluate the validity of their assessment. This kind of validation involves a step by step justification of all aspects related to the assessment process (Crooks, Kane & Cohen, 1996, p. 265). Moreover, the evaluation of validity also involves locating features that can reduce the validity of assessment, namely threats to the validity (Crooks, Kane & Cohen, 1996, p. 267; Messick, 1987, p. 44). This highlight the importance of also directing attention to the validity aspect of oral competence, because validity is considered "the most important quality of an assessment" (Crooks, Kane & Cohen, 1996, p. 265). This means that increasing the knowledge of oral assessment also involves an investigation of the aspects that influence the validity of oral assessment.

A search for literature on validity of assessment display that also research on the validity aspect of assessment is limited. One of the reasons for the knowledge gap in the field of validity of assessment is argued to be because it relates on human justification as evidence of the validity argument. This makes assessment more challenging to carry out, assess and defend (Gipps, 2015). These challenges address the need for a "more systematic inquiry […] into the needs of teachers as assessment developers and validators, and what works best to meet those needs" (McMillan, 2013, p.103). In addition to highlighting the complexity of teachers' job as assessors of students' performance, the knowledge gaps presented above sheds light on the importance of gaining more knowledge on the validity aspect of oral assessment, to develop current knowledge and practices. Therefore, with an aim of addressing these knowledge gaps, this thesis investigates the validity of oral assessment in English.

1.1 Purpose

The purpose of this study is to gain knowledge of factors that threaten the validity of oral assessment in English. Through insight into teacher's reasoning about oral assessment, this thesis examines what factors that potentially reduces the validity of assessment of oral competence in English. In addition, by investigating how the interplay of educational policies

and teachers' practice can influence the validity in their oral assessment, an aim of this study is also to contribute to expanding our knowledge of teachers' role in- and teachers' influence on oral assessment and thus also student learning. The purpose of this study is not to criticize the practice of the teachers involved, but to use their reasoning to gain insight into the field. A consequence of the many factors that influence assessment, is that validation of oral assessment is considered a challenging process for teachers in practice. Therefore, one can assume that there probably is a wide discrepancy between teachers' practice on the field.

1.2 Research question

The aspects presented above demonstrates the need for more research on validity of oral assessment, to strengthen current assessment practices and our understanding of oral assessment. Thus, in the context of English as a foreign language, this study analyzes teachers' reasoning on own practice of oral assessment, and aims to answer the following research question:

What threats to assessment of oral competence in English are revealed through teachers' reasoning about own assessment literacy?

1.3 Definitions

The section above sheds light on four specific terms that are central to this thesis: validity, validation, threats, and oral competence. In this part of the introduction, I define and specify these terms.

1.3.1 Validity

Validity is defined as "an evaluative summary of both the evidence for and the actual—as well as potential—consequences of score interpretation and use." (Messick, 1995, p. 742). In other words, validity is connected to all aspects of the assessment, where the validity argument is based on an evaluation of the entire assessment, through an investigation of the evidential justification presented by the teacher. The aspects that are highlighted as influential for the validity of assessment are "considerations of content, criteria, and consequences" (Messick, 1995, p. 742). In other words, the validity addresses the quality aspect of the assessment, which is based on the relevance of the evidence the teacher present as a

justification of his or her interpretation of students' performance in correlation to the intended purpose of the assessment and the consequences of the assessment.

1.3.2 Validation

Validation is the term used to address the process of investigating the trustworthiness of the score interpretations. An interpretation is valid if "the claims inherent in the interpretation are supported by appropriate evidence" (Kane, 2016, p. 199). This display both the subjectivity of the assessor and the role of evidence through the justification of interpretations, which also points out the need to investigate if the justification presented are valid. The core concept of validation is to analyze the credibility of the arguments that are presented as justifications for interpretations (Kane, 2016, p. 198). The following description defines validation of assessment:

> Score validation is an empirical evaluation of the meaning and consequences of measurement. As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use (Messick, 1995, p. 742).

Firstly, this means that teachers need to evaluate if their interpretations of students' performance are according to the standards and definitions in the curriculum. Secondly, it also means that teachers should evaluate if the evidence used to justify students' competence presents a realistic image of students' knowledge on the field. Thirdly, the teacher needs to evaluate if the consequences of the assessment on the student correlate with the actual performance and needs of the student.

1.3.3 Threats to validity

Validity theories shed light on the relationship between assessment scores and social consequences, where it is advocated that "evaluation of intended and unintended consequences of any testing is integral to the validation of test interpretation and use" (Messick, 1994, p. 13). This addresses the tradition of identifying possible aspects that can reduce the validity of the assessment, to remove such unintended features from the assessment. In the field of validity research, the term *threats* are used when addressing such negative factors that can reduce validity (Crooks, Kane & Cohen, 1996, p. 267; Messick, 1987, p. 44). Hence, when evaluating the validity of assessment, one needs to locate and identify possible threats to the validity of the assessment. This means that validation also

requires teachers to reflect upon possible factors that may have negatively influenced the accuracy of their interpretations or lead to misleading the correctness of their interpretation. In other words, it can involve locating threats that have led to misinterpretation of students' performance, as for example conditions that can have prevented the students from displaying their knowledge.

1.3.4 Oral competence

Since the context of this study is oral competence in English, it is relevant to define the term competence, which can be described as "the sum of knowledge, skills and characteristics that allow a person to perform actions." (Council of Europe, 2001, p.9). I have chosen to use a definition provided by the Common European framework of reference for languages because it offers in-depth descriptions of oral competencies, which is more specific than the definitions in the Norwegian curriculum. This description display that the assessment of oral competence involves all aspects of the oral language of the learner. In the theoretical framework, I present how the English subject curriculum describes oral language competencies. Though the curriculum does not operate with the term competence in relation to the English subject, these characteristics display the content that students' performance should display.

1.4 Thesis structure

The thesis is divided into five chapters. At the end of the thesis, after chapter 5, there is a list of references followed by appendices. Below is a brief overview of the content of each chapter:

The second chapter presents the theoretical framework for this study. The framework emphasizes two theories on threats to validity, definitions of oral competence in English, in addition to aspects that influence assessment of oral competence.

The third chapter describes the methodology. This research project is an in-depth study of two research participants, with a case- study design. The collected data material derives from interviews, documents and diary notes collected over a period of 8 weeks.

The fourth chapter presents the analysis and findings. This chapter demonstrates the threats that emerged from the analysis of the data material and provides evidence to justify the relevance of the identified threats to the validity of assessment of oral competence in English.

The fifth chapter discusses the findings in light of the theoretical framework of the study. It highlights relevant features in both theory and findings and relate them to the research question with an emphasis on what the findings imply for assessment of oral competence in English. Additionally, it illustrates the relevance of the interplay of educational policies and teachers for classroom assessment. Lastly, I present the implications of the study and suggest further research on the topic.

2.Theoretical framework

This chapter presents the theoretical framework for this research project. First, I highlight the theories on threats related to construct validity. Secondly, I present the chain of validation and address threats associated with each step of the model. Lastly, I shed light on relevant theoretical perspectives on oral competence, as a contextualization of the field that the project addresses and to provide a description of the construct that is the target of the research.

2.1 Threats to validity

2.1.1 Construct validity

The term construct is used when referring to the aspects that a test intends to measure (Messick, 1992, p. 1488). In other words, this means the skill or competence that the test is designed to assess. Making sure that the correct construct is measured is "the central issue in all assessment" (Buck, 2011, p.1). Ensuring validity in constructs means to provide evidence for the quality of the assessment. In practice, this is what teachers do when they present evidence of student performance, to justify their interpretation of students' competence, in the area the assessment intends to test. The importance of justification through solid arguments as a baseline for interpretation and test use is the target when evaluating construct validity. In other words, trustworthiness in the interpretations of scores and consequences is the root of construct validity (Messick, 1996, p.744).

Therefore, as mentioned in the introduction, to ensure construct validity, it is essential to evaluate the coherence between the evidence of students' oral competence and the description of the construct provided by the curriculum, to evaluate the justification provided by the teacher. The following definition describes the characteristics of construct validity:

> the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables. In its simplest terms, construct validity is the evidential basis for score interpretation (Messick, 1995, p. 743).

The justification of the aspects covered by construct validity is in other words based on the interpretation of both performances and contextual variables that influence the assessment. In this way, "construct validity addresses both score meaning and social values in test interpretation and test use" (Messick, 1995, p. 741). In the context of assessment of oral competence in English, this means that it is not only the scoring of students' oral competence

that should be addressed when evaluating construct validity, factors such as the learning environment and the circumstances surrounding the assessment in general, are also influential for the validity of an assessment.

This illustrates the importance of understanding the connection between the factors that influence the validity of the construct. The characteristics of construct validity are further described through a unified validity framework that involves "content, criteria and consequences" (Messick, 1995, p. 741). The framework includes six aspects that represent different perspectives that influence construct validity: content, substantive, structural, generalizability, external, consequential (Messick, 1995, p. 744). Despite their relevance for construct validity I have chosen to focus on only two major aspects that threaten construct validity, because of their direct relevance for oral assessment in English, and therefore this project.

2.1.2 Major threats to construct validity

Construct underrepresentation and *construct-irrelevant variance* are highlighted as two major threats to construct validity (Messick, 1995, p. 742). Construct underrepresentation is related to situations where "the assessment is too narrow and fails to include important dimensions or facets of the construct" (Messick, 1995, p. 742). This is, for example, relevant if the construct oral competence is assessed solely through performance that involves listening, which is only one aspect of the construct. The assessment would thus leave out facets of the construct and would therefore not measure the entire oral competence of the student. This illustrate how construct underrepresentation can threaten the validity of assessment.

Construct-irrelevant variance is described as a threat if the assessment is "containing excess reliable variance associated with other distinct constructs as well as method variance such as response sets or guessing propensities that affect responses in a manner irrelevant to the interpreted construct" (Messick, 1995, p. 742). This means that construct-irrelevant variance becomes a threat if the assessment includes aspects or facets that are not considered a part of the construct by the curricula.

A practical example of construct-irrelevant variance in the context of oral assessment in English is, for example, a test that intends to assess listening competence, which is considered a part of students' oral competence, but that demands students to hand in written answers (Norwegian Directorate for Education and Training, 2013a). This means that the students are required to use their written competence to answer a test that is supposed to assess a part of their oral competence. This can result in scores that present a misleading image of students' listening competence, especially for students that struggle with writing and are thus not able to display their knowledge through writing, even though they know the answer and can express it orally. In this way, one can discuss the validity of the grade or the feedback provided and evaluate if such a test can reflect student's listening competence.

Furthermore, the definition above sheds light on the fact that guessing can threaten the validity of assessment, because it leads to questioning if the performance represents the learner's actual competence. Another example of construct-irrelevant variance is predictability in tests, which leads to construct-irrelevant variance because it then assesses: "rote learned responses rather than demonstration of skills" (Stobart, 2009, p. 168). The examples above reflect the role of construct irrelevant-variance as a threat that can occur both because of actions taken by the assessor or a result of unintentional actions by the student.

The theory divides construct-irrelevant variance into two underlying categories: construct-irrelevance easiness and construct irrelevant difficulty (Messick, 1995, p. 742). The latter refers to assessment that is more difficult for some students, as for example if the assessment is supposed to measure content knowledge but the learner struggles with understanding the questions and is thus not answering correctly. The content of this threat display that it is not relevant for assessment of oral competence in the subject of English, since both language aspects and content are represented by the construct oral competence.

However, construct-irrelevant easiness is more relevant for assessment of oral competence in English. It is described as: "construct-irrelevant easiness occurs when extraneous clues in item or task formats permit some individuals to respond correctly or appropriately in ways irrelevant to the construct being assessed" (Messick, 1995, p. 743). Even though construct-

irrelevant variance is divided into two categories, both threats can occur simultaneously (Messick, 1995, p. 742). In the context of assessment of oral competence, it is thus most relevant to address construct-irrelevant-variance as one concept, as I will do in this thesis.

2.1.3 The chain of validation

While the theories on construct validity emphasize the perspectives of the construct, the theoretical framework of the chain of validation addresses assessment as a process where attention is given to each stage of the assessment process. The framework underlines that "validation can only take place if the intended purposes of the assessment are well understood" (Crooks, Kane & Cohen, 1996, p. 267). This illustrates the importance of the correlation between understanding of the construct that is measured, and the decisions taken in the assessment process. If the assessor does not fully understand what he is measuring, he is not able to validate the assessment. This means that all decisions taken by the teacher when designing an assessment are based on the teacher's understanding of the construct the assessment intends to test. In other words, if the teacher has misinterpreted the curriculum's guidelines for oral competence in English, the assessment will reflect this misinterpretation.

An argument- based framework to validity containing eight facets that are important for validation of assessment is thus developed to capture all stages involved in the assessment process (Crooks, Kane & Cohen, 1996, p 266). These aspects are connected to each other through an illustration of an eight-linked chain, where each chain represent significant steps of the validation process (See Figure 1). The framework highlights that "The appropriateness of the assessment tasks and procedures to those purposes will be a critical issue in evaluating the strengths of each link in the assessment chain." (Crooks, Kane & Cohen, 1996, p. 267). This means that the first step in the validation process should be to evaluate the level of importance of the links in relation to the purpose of the assessment.

The purpose of the model is to provide the validator a "pathway for the validity argument", as a help to ensure and evaluate the validity of assessment (Crooks, Kane & Cohen, 1996, p 266). The content of each step is founded in validation criteria. Together they represent the entire process from planning the assessment and conducting the assessment to consequences of the assessment. The model is argued to present a complete image of the validation process,

but it is also underlined that it is meant as an illustrative approach. The importance of each aspect involved in a specific assessment differs depending on the characteristics of the assessment. For example, threats related to aggregation of scores are less relevant for tests that only consist of one task. Furthermore, it is highlighted that the last step in the validation process should be: "identification of the weakest links in the chains" (Crooks, Kane & Cohen, 1996, p. 282). This means to address major the threats that limit the overall validity of the assessment, with an aim of revealing factors that can limit the validity of the assessment.

Through close examination of each step in the chain of validation, and by investigating possible threats and weaknesses, it provides "a systematic approach to validation" (Crooks, Kane & Cohen, 1996, p. 284). The validation of assessment is in other words based on scrutiny of these steps. The theoretical framework of the chain of validation is illustrated below (Figure 1), with the eight chains represented as linked corners of an octagon. These steps represent the stages teachers undergo when assessing students' performance.

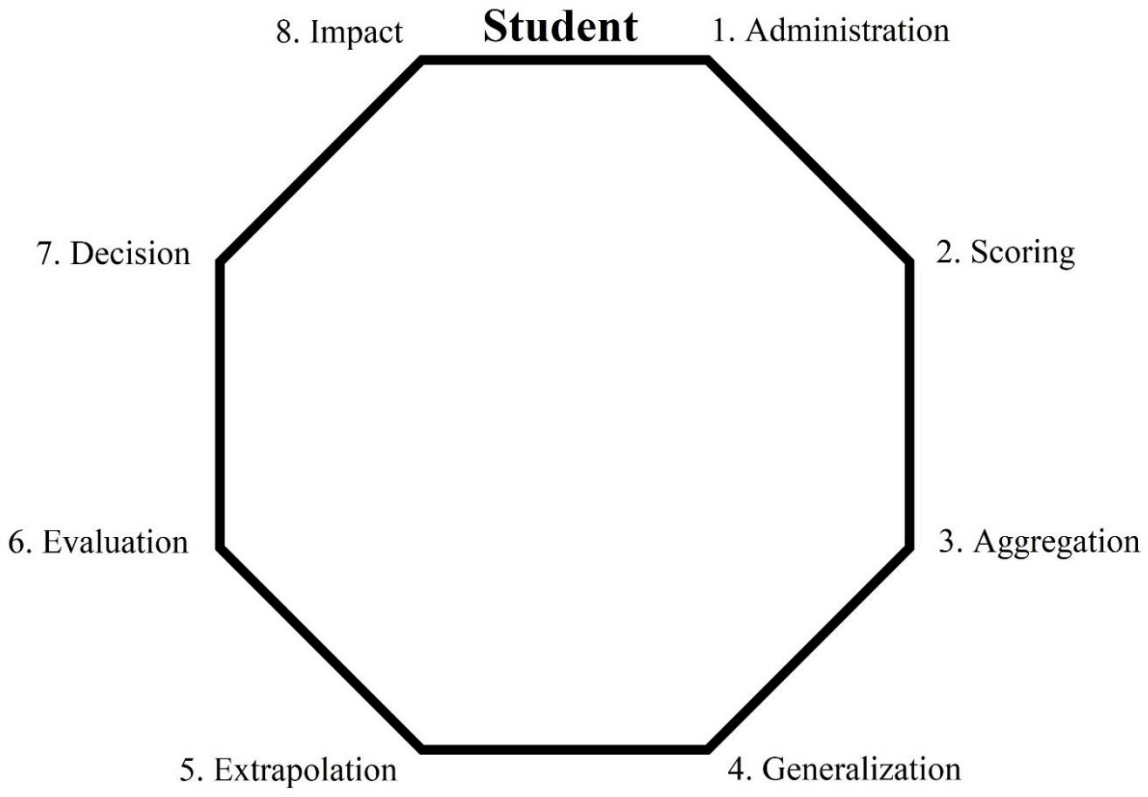

Figure 1: A model based on the chain of validation (Crooks, Kane & Cohen, 1996, p. 268)

The first step in the model addresses the practical *administration* of the assessment. Possible threats and decisions taken at this stage influence the task performances of students. The second stage of the assessment is *scoring* of student performance. The interpretations and decisions taken at this stage influences the task scores. The third stage is *aggregation* of task scores, which influences the combined scores of the performance. Stage four is *generalization* of students' knowledge in the specific competence that is tested. The interpretations and decisions taken by the teacher at this stage influences the assessed domain scores. The fifth stage is *extrapolation* of the competence that the assessment aims to assess. The decisions that are taken at this stage influence the target domain scores (for example the grade). The sixth stage of the assessment process is *evaluation* of the assessment process in relation to its intentions. This stage has an impact on the judgments given by the teacher on the students' performance. The seventh stage is when the teacher decides what to do with the information that derives from the assessment. The *decisions* taken at this stage influences the actions that are a result of the assessment. The last stage in the model is the *impact* of the assessment. It addresses how the entire assessment process influence the student (Crooks, Kane & Cohen, 1996, p.268).

The connections between each step and the direct consequences on different parts of the assessment are meant to illustrate the role of consequences in the validation process (Crooks, Kane & Cohen, 1996, p. 269). Hence, the model can be used both as a checklist when planning assessment and to identify threats to the validity of assessment. Furthermore, the links between the stages in the model symbolizes that the importance of each step in the validation process, and the connection between each step: "with weakness of any one link weakening the chain as a whole" (Crooks, Kane & Cohen, 1996, p. 266). In other words, securing validation of each step is crucial for the validity in the entire assessment. If the link between administration and scoring is weak, in addition to influencing the task performances, it influences the validity of the rest of the assessment.

2.1.4 Threats to validity

In the following part, I will elaborate on each step in the validation process and emphasize the threats associated with them.

1. Administration

Evaluating the validity of task administration involves an investigation of the context and the circumstances that surround the assessment. The aspects that are addressed in this stage of the assessment are motivational factors, anxiety, the assessment conditions and the communication of the task- and answers (Crooks, Kane & Cohen, 1996, p. 270-271). As well as pointing out what is implemented in this stage, the aspects above sheds light on facets that also functions as possible threats to the validity of administration. Student anxiety is, for example, an aspect that can influence the accuracy of students' performance in oral assessment because it can prevent the student from fully displaying his or her knowledge. Hence, preventing student anxiety is a way of increasing the validity in task administration. However, as listed above, other facets can still hinder the validity of this stage. This means that they also need to be taken into consideration by the teacher when planning the administration of oral assessment.

2. Scoring

Five threats are characterized as influential for the validity of the scoring of students' performance (Crooks, Kane & Cohen, 1996, p. 271). The first threat addresses situations where important facets of the performance are left out of the assessment, this can, for example, be specific competence inherent to the construct or caused by errors in "scoring rubrics or answer keys" (Crooks, Kane & 1996, p. 272). For example, in the context of oral assessment in English, this becomes a threat if the teacher intends to assess the total oral competence of students, but in practice only assesses students' pronunciation and leaves out other aspects of the students' oral competence. The result is then an assessment that does not give insight into the total oral competence of the students.

The second threat involves scoring that emphasizes a certain type of criteria or answers, which can be connected to construct underrepresentation. The third threat related to this step is inconsistency in the assessment. Such variance can, for example, be caused by differences between different teachers or can be a threat if the teacher is not consistent in his or her own assessment. The fourth threat that is identified in relation to this stage is that the scoring can be too analytic. In other words, this means that the assessment directs too much attention to grading each of the different parts of the task, leading to minimizing the chance of capturing a complete image of the entire performance. On the contrary, the fifth threat relates to assessment that is too holistic, if it does not provide the student with information about his or

her performance in the different parts of the assessment (Crooks, Kane & Cohen, 1996, p. 272).

3. Aggregation

The third step is described as: "Aggregation of the scores on individual tasks to produce one or more combined scores" (total score or subscale score) (Crooks, Kane & Cohen, 1996, p. 268). Two threats are associated with this step in the validation process. Firstly, assessment where the aggregated tasks are too diverse can lead to weakening the validity of the assessment. This threat is especially relevant for assessment that involves aggregation of many different topics. Such aggregation of scores can lead to decreasing the coherence of the aggregated score.

Secondly, inappropriate weights given to different aspects of performance are also identified as threats to aggregation (Crooks, Kane & Cohen, 1996, p. 273). This is what the theories of construct validity refers to as construct-irrelevant variance (Messick, 1995, p. 742). One should for example question the validity in an assessment of students' oral competence in English if it is only based on the content of a presentation that covers a topic that in itself is not specifically related to the definitions of oral competence in English (that is provided by the curriculum).

4. Generalization

Generalizability is defined as "the accuracy of generalizing from student's aggregated score to his or her universal score in the corresponding assessed domain" (Crooks, Kane & Cohen, 1996, p. 274). In other words, how tasks are used as representatives for the knowledge in a certain domain. For example, when teachers design assessment, they need to evaluate if the tasks in an oral test can display students' competence in oral English. The first threat linked to this stage occurs when the circumstances influencing the assessment is too variable. This relates to for example time, format and approaches used. The second threat is caused by assessment that contains too few tasks for the student to display his or her competence. The third threat involves the criteria and scoring of tasks and especially targets assessment where there is inconsistency between these aspects (Crooks, Kane & Cohen, 1996, p. 274- 275).

5. Extrapolation

The fifth stage in the chain of validation involves an investigation of the trustworthiness of the extrapolation (Crooks, Kane & Cohen 1996, p. 275). When assessing students' competence, the assessment contains a subset of the competence it is supposed to represent (Crooks, Kane and Cohen, 1996, p. 275). In other words, the tasks that are used to assess competence in an area "are often a systematically biased sample from the target domain because some categories of tasks or conditions of assessment in the target domain have been excluded from the assessed domain." (Crooks, Kane & Cohen, 1996, p. 275). Therefore, it is important that the teacher evaluates if the construct is fully represented in the tasks of an assessment. Since assessment demands such extrapolation, it is relevant to investigate the trustworthiness of the sampling of the competence it is designed to measure. Two threats are especially highlighted as relevant to investigate: firstly, the validity of the extrapolation, does the extrapolated tasks represent the target domain? Secondly, are some features of the construct given too little weight, or left out? This threat relates to the theories on construct underrepresentation Messick (1995, p. 742).

6. Evaluation

The evaluation link is described as following: "Leading from the target domain scores to judgments about the merit (and perhaps the strengths and weaknesses of the student's performance "(Crooks, Kane & Cohen, 1996, p. 276). In other words, what do the scores mean? This stage involves validation of the evaluation of the assessment, including interpretation of the assessed performance and the interpretation of the construct. This is what teachers do when they evaluate students' performance on a test, and question what it implies about, for example, the students' oral competence in general. The three following threats are associated with this evaluation: weak understanding of the information that is retrieved from the assessment, consequences related to weak understanding of the construct (it can, for example, lead to misinterpretations) and lastly, biased interpretation of the performance (Crooks, Kane & Cohen, 1996, p. 276-277). An example of the latter is if a teacher gives a student a better grade than what his performance deserve because the teacher knows that the student has done his best.

7. Decision

This step refers to the decisions that are following the judgments of the assessment. This stage is where the assessor decides what to do with the information he or she gained from the assessment. In other words, "what actions to take as a result of the judgments" (Crooks, Kane & Cohen, 1996, p. 277). Does the student need more help in a specific area, or does the assessment reflect a topic that the entire class need to work on?

The framework highlight that good decisions will "result in generally beneficial consequences for students and other participants in the assessment processes" (Crooks, Kane & Cohen, 1996, p. 278). A decision, as a result of a judgment can, for example, be that the teacher adapts the teaching to what the students' need to work on to improve their oral competence. For example, through more practice on oral interaction with peers. The two threats that are linked to this step are threats related to inappropriate assessment standards that are created by the assessor and threats caused by poor pedagogical decisions on what actions that follows the assessment (Crooks, Kane & Cohen, 1996, p. 278).

8. Impact

The final link in the chain of validation involves "the impact of assessment on students and other participants in the assessment process" (Crooks, Kane & Cohen, 1996, p. 279). In other words, this step investigates the consequences of the assessment. Additionally, it also addresses the influence of experiencing the entire assessment process. For example, if the teacher decides to give formative feedback, it can lead to facilitating the students' development on the field. The two major threats related to this step involves how the assessment influence students. Firstly, a positive consequence such as feedback that leads to development and motivation, is a possible threat if it is not given to the student. Secondly, the other threat is related to serious negative consequences caused by the assessment. This can, for example, lead to decreasing motivation or increasing anxiety (Crooks, Kane & Cohen, 1996, p. 279).

2.1.5 Threats to validity of oral assessment

In this research project, I will use the framework above to shed light on threats related to oral assessment in English, based on an analysis of teachers' reasonings. The theory of construct validity and the chain of validation provides a framework that can help the validator identify

threats related to assessment. These theories are therefore useful when evaluating the validity of oral assessment in English. Furthermore, the framework illustrates the importance of understanding the construct, or purpose of the assessment to ensure validity, and how this understanding influence the steps in the assessment process. In practice, this means that the decisions taken by the teacher in the assessment process are based on the teachers' interpretation of what counts as knowledge. If the teacher has misunderstood the definition and the content of oral competence in English, it will have an impact on the emphasis of the assessment and all decisions and interpretations related to the assessment. Moreover, both theories also highlight the importance of the contextual and consequential aspects that influence the validity of assessment. The question is what role the social context plays for oral assessment?

The complexity of the chain of validation provides a clear model for systematic evaluation of the validity of assessment. Teachers probably do not have in-depth knowledge of validation theory in specific, and one can therefore not expect that they are able to articulate specific reasoning on the validity of oral assessment in English. However, by analyzing their concrete examples, one can discover what they know. In other words, one can assume that a considerable amount of the knowledge they withhold on the field is tacit. This means that they cannot explicitly describe their knowledge, because they cannot identify their knowledge themselves (Polanyi, 1966, p. 5).

2.2 Oral competence

This section addresses the construct that is discussed in the teachers' reasoning, namely oral competence. There are multiple definitions of oral competence, but I have chosen to highlight two definitions that are provided by the curriculum, because of its role as the policy document that defines the content and aims for Norwegian education. The curriculum does not specifically define oral competence, but it sheds light on central features of oral language. Together, these definitions form the foundation for the understanding of oral competence that is advocated in this thesis.

Firstly, I highlight a definition of oral communication, which is described as following: "The main subject area oral communication deals with understanding and using the English language by listening, speaking, conversing and applying suitable communication strategies" (Norwegian Directorate for Education and Training, 2013a). This definition points out three major aspects of oral communication: oral production (speaking), reception (listening) and spoken interaction (conversations) in addition to including competence in adapting language to the context ("applying suitable communication strategies"). Since oral language is based on communication, these characteristics are argued to describe important facets of oral competence.

Secondly, since oral skills are defined as one of the five basic skills in Norwegian education, they are connected to the subject description of all subjects. In the subject of English, oral skills are defined as:

> 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 (Norwegian Directorate for Education and Training, 2013b)

This means that in addition to being able to speak, listen, and interact, oral skills involve knowledge on how the context and agents involved in the interaction influence the language. Adapting the language according to these aspects require knowledge of social norms and customs. This displays the complexity of oral language, in addition to illustrating the close connection between oral language and the context where it is produced. This reflects the communicative language approach that sets the premises for today's language instruction. In addition to emphasizing linguistic aspects of language it underlines the importance of sociolinguistic and pragmatic competence as important parts of language learning (Council of Europe, 2001, p.13). Together, these definitions outline the knowledge that should be emphasized when assessing oral competence in English.

2.2.1 Assessing oral competence

In addition to emphasizing the facets mentioned above when assessing oral competence, the assessment criteria are subject to the characteristics of oral language. For example, if assessing linguistic aspects of oral language, one must acknowledge that one cannot expect

the same complexity as with written language, because oral language is less complex and has a shorter length than written language (Chafe, as cited in Luoma, 2011, p.12). The reason for this is because of the context of oral speech, and the aim of it, which is communicating a message to a receiver that needs to remember the message. Relating this to the chain of validation, it demonstrates how the construct influences what steps to include when evaluation educational assessment (Crooks, Kane & Cohen, 1996, p 266).

2.3 Contextual perspectives of oral assessment

Since classroom assessment takes place in a classroom context, it is essential to question how this context influences the assessment. For example, when assessing oral competence in English, students' experience of the context will influence the oral language of the student, and his or her understanding of the tasks (Brookhart, 2005, p. 6). Therefore, to fully evaluate the validity of assessment, it is also necessary to evaluate the impact of the classroom context upon the assessment.

2.3.1 Assessment of oral competence

Therefore, when evaluating the validity of oral assessment, one should include the role of the context on the assessment. The contextual and social aspects that influence spoken interaction are "purpose of talk, the speaking situation and speaker roles" (Luoma, 2011, p. 22). This means that if the purpose of an assessment is to gain insight in students' oral competence, the circumstances surrounding the assessment and the people involved will influence their performance. For example, if the learning environment is negative, or if a student is peered with a person he or she feels nervous around, this will influence the oral language of the student. Therefore, ensuring that the performance reflects the students' oral competence implies an understanding of the fact that every conversation is unique, because: "speakers react to each other and construct discussions together", and because we adapt our language to the situation we find ourselves in (Luoma, 2011, p. 27). Thus, ensuring validity of oral assessment involves a consideration of how the context and the people involved influence the language of the speaker.

Furthermore, one should also take the level of formality of the situation into consideration when assessing oral language. Is it planned or unplanned? The degree of preparation that is

put into the communication influences its quality (Buck, 2011, p. 10). In planned speech, as for example presentations, the language is usually more formal than in unplanned speech such as classroom talk. Situations that are formal demands a "more written-like language" (Luoma, 2011, p. 13). This means that the level of formality influences the complexity of both grammar and vocabulary. The fact that unplanned speech is spontaneous makes it informal because it is produced in real time (Ochs, as cited in Buck, 2011, p. 9). Therefore, when assessing oral competence, one also need to adapt the assessment criteria to the level of formality of the assessment situation.

2.3.2 The role of the teacher

The relevance of the context reveals that the classroom environment itself is subject to the validity of the assessment. In addition to the relevance of these aspects for assessment, the theory also directs attention towards the role of the teacher, and how teachers influence the classroom environment, and thus the assessment context. "Teacher beliefs, teacher instructional practices, and teacher understanding of both the subject matter and students are relevant validity concerns" (McMillan, as cited in Brookhart, 2005, p.10). This highlights the reason for the relevance of the teacher as a contextual facet that influences assessment, also because of teachers' impact on teaching, understanding of the construct and assessment (Gipps, 2012, p. 55). One can therefore, for example, question how important the confidence of the teacher is for the validity of the assessment?

Additionally, when evaluating the validity of assessment, we address the connection between the assessment and the construct it assesses, and because of the role of teachers as interpreter of the curriculum, one can argue that the goals of the assessment or the construct, is influenced by the "teacher beliefs about the subject matter and about what constitutes appropriate instruction" (McMillan, as cited in Brookhart, 2005, p.10). This is also an argument that displays the importance of addressing the impact of the teacher when evaluating the validity of oral assessment. Firstly, how important is teachers' understanding of for example oral competence? Secondly, since assessment is connected to teachers' interpretation of the curriculum, how significant are the guidelines and definitions in the curriculum for the validity of oral assessment in English?

2.3.3 Validity in assessment of oral competence in English

The major influence of contextual factors highlights the importance of understanding how "social and situational needs", in other words, the learning conditions, can influence the oral language of the student (Luoma, 2011, p. 28). When evaluating the validity of oral assessment in English, these are, in addition to the other features addressed in the theoretical framework of the chain of validation and construct validity, aspects that need to be taken into consideration, to fully evaluate the justification of the teacher and the evidence of students' performances. This display the complexity of validity in assessment, and the challenge teachers meet when planning and evaluating validity. The relevance of these aspects for assessment of oral competence in English will be discussed in chapter 5.

3. Methodology

Within the framework of social constructivism, this thesis investigates how teachers' reasoning reflects threats to validity of oral assessment in English. The study is based on data material from two group interviews, two individual interviews and 16 diary entries collected from two English teachers working in upper secondary school. Additionally, I collected documents related to their practice of oral assessment (plans, criteria, assignments). In this chapter, I describe the design of the study and the methods used in terms of the data collection and the analysis.

Firstly, to provide transparency to the research, I position myself as a researcher. Secondly, I describe the methods used in this study and sheds light on the context of the research and the research participants. Thirdly, I describe the data collection and the methods used to collect data material. Fourthly, I provide an insight into the process of data analysis and the methods used to analyze the data material. Lastly, I discuss aspects related to validity, trustworthiness, limitations and ethical considerations.

3.1 Research context

This research project is connected to the larger SKUV project (school-based competence development in assessment), which is a school development project initiated by NTNU and Trøndelag county with a purpose of direct attention to development of assessment practices in upper secondary school. It is a four-year-long project where the participating schools agree to work collectively internally in the school with support from researchers to improve and develop own assessment competence. The aim of the project is to "develop models for practice and research on school development that contribute to strengthening assessment competence and assessment practices in upper secondary school" (NTNU, n.d). As a part of this project, teachers at participating schools choose self-chosen topics that they specifically want to work on as a part of the project to develop own assessment competence.

3.1.1 Methodological framework

As mentioned above, social constructivist theories are the theoretical framework that underpins the interpretations, decisions, and reflections in this thesis. The baseline for social

constructivism is the belief that reality is "constructed and reconstructed through actions and interaction between human beings" (Berger & Luckmann, as cited in Ringdal, 2013, p. 43). In other words, the tradition highlights human beings as agents that construct their own reality, where individuals add meaning to the world that surrounds them. Through interaction, beliefs and perceptions are challenged and altered. In this way, one can argue that from a social constructivist perspective, reality can be changed and improved by people (Ringdal, 2013, s. 43). Utilizing this framework as a baseline for this research project means that by gaining insight in the participants' perceptions of their own reality, through investigation of their reasonings and discussions, one can improve or develop current practices in the field of study.

3.2 Researcher bias and positioning

Before describing the choices made regarding design and methods in this thesis, I will briefly position myself as a researcher. In qualitative studies, the researcher is the research instrument that collects, analyze and interpret the data. This means that the researcher influences all aspects of the research. Thus, reflecting upon own positioning as a researcher is an important factor to give transparency to the research project (Nilssen, 2012, p. 21).

The reason for choosing to investigate threats to validity of oral assessment in English as a topic for my master's thesis is that I find it challenging myself to assess oral skills in the English subject. The curriculum offers few concrete guidelines for how to assess oral skills, and choices made by teachers are therefore often based on individual assumptions of what is considered good and bad assessment practice (Knowledge Promotion Reform, 2006). Discussing this with my colleagues, I found that we had different opinions on how assessment of oral competence should be practiced and how to ensure validity in such assessment. They also expressed that they found oral assessment challenging. This made me curious to find out more about how teachers reason when discussing assessment of oral competence, and how their reasoning reflects threats to validity of oral assessment.

My expectations for the research project was that the participants also found oral assessment challenging. In addition, I anticipated that they had different perceptions of how to assess oral competence, and what evidence to emphasize as justification for students' competence. I

especially expected that some participants could have strong opinions on how oral competence in English should be assessed, based on their own perceptions of what a "good" method is. Thus, I expected to find many features that could threaten the validity of oral assessment in English.

When conducting the research these expectations are aspects that could have influenced the investigation in this project. My expectations could, for example, have guided the research, and lead to an emphasis on evidence that supported and confirmed my expectations. In other words, lead to selectivity making me less open to findings that contradicted my expectations. This means that a consequence of my expectations could be that I left out important data. In specific, my biases could, for example, have influenced the way I asked questions when interviewing the participants, in addition to influencing what questions I chose to ask. Furthermore, it could also have influenced my decisions and interpretations when analyzing the data material. Being aware of own expectations can decrease their impact on the findings, but completely removing their influence is challenging, because they can affect interpretations and decisions both consciously and unconsciously throughout the entire research project. This means that even though I tried to limit their influence, they can still have had an impact on the research.

3.3 Method and research design

Since the goal of this study is to obtain in-depth knowledge of teachers' experiences, reflections and practice through an investigation of their reasoning, I have used qualitative methods to answer my research question. The thesis has a case study design which is described as research that is based on investigating few cases to collect in-depth knowledge (Ringdal, 2014, p. 170). In other words, the case study method can be characterized as descriptive research (Yin, 2003, p. 3-4). Using multiple sources of data is thus important to achieve a broad and complex image of what is being investigated (Baxter & Jack, 2008, p. 544). The fact that the data material is collected from few participants means that this study is not meant to generate generalized knowledge. The qualitative case study approach can be defined as "an approach to research that facilitates exploration of a phenomenon within its context using a variety of data sources" (Baxter & Jack, 2008, p. 544). This underlines the importance of understanding that the context influences the phenomenon and that this

connection needs to be elucidated to create a complete understanding of the phenomenon. Therefore, I have also chosen to investigate how educational policies influence teachers' validation of oral assessment.

3.4 Selection of participants

The selection of participants in this thesis is based on what is referred to as "purposeful sampling" (Patton, 2015, p.264). This is a method where the selection of participants is based on a process where the researcher first decides what target group that should represent the data that is relevant for the research, and in the next step chooses participants from this specified target group. I had four criteria for the selection of participants:

Firstly, the target group of my interest was English teachers working in upper secondary school. Secondly, although I did not set any limitations regarding the year and level they were teaching, I considered it interesting if I found participants that taught vocational studies and participants that taught general studies. The reason for this interest is that the first year of general studies and year one and two of vocational studies shares curriculum. Thirdly, I wanted 2-3 participants for my study. Fourthly, my research project aimed at choosing teachers that worked at schools participating in the SKUV project. Based on these criteria, I found schools that suited the profile and contacted English teachers that matched the target group I was interested in.

3.5 Research participants

The research participants in this thesis work at an upper secondary school in a small municipality in the middle of Norway. This upper secondary school is a relatively small school, with 220 students (2018). The school offers both programs for vocational studies and programs for general studies. My initial intention was to recruit three English teachers from this school, which I did, but after the first group interview, the third teacher could not participate anymore. This means that there are three participants in the first interview, but only two in the rest of the data material.

The two teachers that participated in the entire research project have different backgrounds related to their teaching experience and education. Hannah has a master's degree in English

and has many years of experience as an English teacher, while Berit has less experience as a language teacher and does not have any formal pedagogical competence in English. The names used in this thesis are fictive. The research participants teach classes in both educational programs, and both teachers have another European language as their native language.

Since the school is participating in the SKUV project, the teachers were engaged in self-chosen projects related to assessment simultaneously as they participated in this project. The teachers at this school cooperated on SKUV-projects with colleagues that taught similar subjects. The selected research participants had chosen to focus on oral assessment but had not started to work on their SKUV project. However, their own project was directed towards the teaching of all foreign languages, not solely the English subject.

At the first meeting with the participants, they told me that they saw my research project as a way of achieving insight into the field they had chosen to focus on in their own project. This means that the participants also had their own personal interests in oral assessment, and perhaps a concrete motivation to participate in my project. This is an aspect I had in mind when collecting the data, since their intentions and motivation are factors that can influence the data material, and thus the research because the participants already were interested or at least engaged in the field of study. This points out the importance of understanding the relationship between the context and the participants, to gain in-depth knowledge on aspects that influences the participants.

3.6 Data collection

The project was approved by the Norwegian Centre for Research Data (Appendix A). After submitting the project application and receiving their confirmation letter, I contacted the participants and set a date for the first interview. In the information letter that was sent out in the first e-mail to the participants, I informed them about the topic of my master's thesis (assessment of oral skills in the English subject) (Appendix B). I purposely did not at any stage of the data collection process share with them the research question and the precise topic of the study. The reason for this decision was that I did not want to direct or influence the data

they provided. I told them that I was interested in all experiences and reflections related to assessment of oral competence in English.

Before providing an in-depth description of each stage of the data collection, I will present a brief overview of the entire process. The data collection period lasted for approximately 8 weeks, from November 2017 to February 2018. The first stage of the data collection was a group interview with three of the participants, which took place at the participants' workplace during the last week of November 2017. During eight of the following weeks after the group interview, from December 2017 through February 2018 (except the Christmas vacation), I collected diary notes written by the participants, one time per week. The diary notes were written reports containing teachers' reasoning on assessment of oral competence. In the middle of these eight weeks, in January 2018, I interviewed the participants individually. The last stage of the data collection was a final group interview, it took place at the same location as the first group interview in mid-February 2018. Furthermore, I collected relevant documents at the beginning of the period and in the middle of the eight weeks of writing (after the Christmas vacation).

3.7 Interviews

In total, I conducted four qualitative interviews: two group interviews and two individual interviews. They were recorded digitally, transferred to my computer, and transcribed by me. The interviews were conducted in Norwegian, despite that both research participants have another native language than Norwegian. This was a request from the participants themselves because they felt more safe and competent to discuss the topic in Norwegian than in English. Therefore, the excerpts that are displayed in the analysis are translated to English by me. In the translation process, I focused on maintaining the original messages rather than directly translating the words.

3.7.1 Group interviews

I based the questions in the interview guides on theories on threats to validation of assessment. In addition, they were also influenced by my own interpretations, understanding of the topic, and my position as a researcher. The interview guide for the first group interview contained ten open-ended questions (Appendix C). To ensure consistency in the data collection, I decided to start and end the data collection with a group interview. The intentions

of choosing the group interview form were to establish a common ground for the participants, where they through discussions with colleagues could share perspectives, ideas and reflect upon their experiences. Such discussions can also highlight differences between the participants' practice (Morgan, 1997, p.20-21). Thus, in group interviews, interactional features such as jointly constructed meanings or disagreement also communicate meaning, which can give further insight in the field.

Group interviews can also reduce the interference of the researcher upon the participants by reducing the researcher's control of the interview situation. In this way, the premises of the interview provided a starting point where the participants could reflect together in a conversation (Morgan, 1997, p. 22-23). Furthermore, the intentions of the last group interview were that the participants could share new ideas, experiences, and reflections that had derived from the diary writing, to see if their reflections and the focus on oral competence had changed their perspectives. The interview guide consisted of nine questions (Appendix E). The group interviews were structured interviews with open-ended questions, and they were held at the upper secondary school where the teachers work.

3.7.2 Individual interviews

I conducted the individual interviews in the middle of the diary writing- period. I chose to interview them individually to create a setting where the participants could share their reflections without interference from their colleagues. I asked questions related to their diary notes, together with five questions that were similar for all the participants to have some common ground for comparison (Appendix D). These interviews were conducted via telephone because of the distance to the participants' workplace. The individual interviews were semi-structured because it creates a feeling of having a conversation where each part is equal participants (Yin, 2003, p. 89).

Basing some of the questions on their diary notes lead to an interaction between the notes and the interviews. This method is a way of creating consistency between the different parts of the data collection, where one feature creates questions that directs the focus of another part of the data collection (Postholm, 2011, p. 72). Additionally, I did not want to direct the participants when sharing thoughts and reflections in relation to their diary notes, because I did not want to interfere and guide their thought process. This was the reason for the

distinction in structure between the group interviews and the individual interviews, where the individual interviews were constructed to be more informal. However, in terms of the validity of the interviews, it is relevant to underline that the data material that derives from the interviews only should be interpreted as verbal reports. This means that the answers are likely to be influenced by the participants' "bias, poor recall or poor or inaccurate articulation" (Yin, 2003, p.92). This highlights the importance of using multiple sources of data material, to ensure validity in the findings.

3.8 Diary method

Diaries are described as "self-reporting instruments used repeatedly to examine ongoing experiences" (Bolger, Davis & Rafaeli, 2003, p. 580). This description elucidates the metareflective qualities of diary notes, where the writer reflects upon own experiences. The diary method is a data collection method where writing of diary notes is used as a means to gain insight in participants' reflections of everyday situations (Bolger, Davis & Rafaeli, 2003, p.580) The method is based on similar principles as normal diary writing, but it especially emphasizes its reflective and metareflective qualities.

Moreover, diary notes also direct attention to the importance of the relationship between the context and the processes that unfold (Wheeler & Reis in Bolger, Davis & Rafaeli, 2003, p. 580). In other words, the writer reflects upon experiences in their natural setting, which is also considered an important aspect of case studies where the research participants and their surrounding context is investigated in a mutual relationship. Hence, diary notes can give insight into differences between the research participants in addition to elucidate the reasons for these differences (Bolger, Davis & Rafaeli, 2003, p. 587).

Furthermore, an interactive relationship between writing and thinking occurs when writing diaries, the relationship is described as following: "Written languages are not only an instrument of thinking representation but also a factor of thinking development" (Sá, 2002, p. 152). The idea underpinning this description of the outcomes of utilizing written text to process experiences is that there is an interactive connection between the diaries and the writer. The writer can develop his ideas through his or her diary entries and at the same time, the written representation of thoughts and ideas can help the writer develop his or her

reasoning through meta-reflection and awareness. This characteristic displays its relevance as both a method to gain insight into teachers' reasoning and a method to develop their reasoning.

However, the commitment of the participants will influence the reliability and validity of the data collected from the diary notes. A possible issue can occur if the writer develops a "[…] habitual response style when making diary entries" (Bolger, Gleason & Shrout, as cited in Bolger, Davis & Rafaeli, 2003, p. 592). This tendency can also influence the participants' perception of the guiding questions and lead to focusing their answers on what they perceive as relevant for the topic and omit other questions. Moreover, the process of diary writing can also lead to influencing the participants' entries so that they try to adjust the content according to the target domain or what they believe the researcher is looking for (Bolger, Davis & Rafaeli, 2003, p. 592). The importance of these factors reveals that these aspects must be taken into consideration when analyzing data in diary research.

The diary notes that were collected in this thesis were collected through e-mail and has a time-based design, which means that the collection of diary notes follows a specified interval or schedule (Bolger, Davis & Rafaeli, 2003, p. 588). In this project, I have as mentioned earlier, chosen to collect the participants' diary notes once a week for eight weeks. The diary notes were based on some guidelines that I gave the teachers in the beginning of the process, to help them in their reflection process, but the participants were also encouraged to write whatever they felt relevant for assessment of oral skills in the English subject (Appendix F). The guidelines are presented below in Table 1:

When you are writing the reflection notes, I want you to think about how you have worked with oral competence in English this week. You can also write about how you specifically worked when planning and conducting activities and assessment of oral English, and what you emphasize when assessing students' oral competence. What do you look for, and how does it cover the competence aims in the curriculum? What do you do to make sure that you actually know what students' oral competence is? Your notes should be a coherent- and narrative text. Ask if there are any questions!

Table 1:Instructions for the diary notes

Since the emphasis in diary notes is the content, and because I did not want the teachers to focus on to their writing skills when writing their diary notes, I decided to keep the language informal in the instruction, to set an example. Furthermore, I also decided to use the term reflection notes in the instructions because this is a genre the teachers were familiar with.

3.9 Documents

The third source of data was the collection of documents related to the teaching of oral competence in the English subject. These documents included: planning documents as for example projects plans, monthly plans and plans for specific parts of their teaching related to oral competence in addition to insight into criteria related to these plans. I gathered most of the documents at the beginning of the data collection period, but I also encouraged the teachers to send me documents that were related to oral competence during the data collection period. Despite that the documents gave some insight into how the participants worked, it did not provide relevant information about their reasoning on assessment of oral competence, or about threats to validity. Therefore, I decided to leave them out of the analysis and focus on the interviews and the diary notes.

3.10 Data analysis

3.10.1 Coding

In my analysis, I emerged the data material with an abductive approach, where the coding was founded on both theory and empirical evidence (Aliseda, 2006). I started by transcribing the interviews and diary notes. After that, I printed out the transcriptions and began to analyze the material from an inductive perspective. I coded them by searching for patterns and wrote codes on the right side of the paper, and my comments and reflections on the left side. I mostly used descriptive codes, which are codes that summarize the content of the excerpt, though some codes are taken directly from the utterance of the participant (Saldaña, 2016, p. 4) After this process some reoccurring codes emerged, such repetitive topics are referred to as patterns. One of the goals of the coding process was to find patterns, because they can give insight in important topics in the data material (Saldaña, 2016, p. 6).

As a guidance for the coding process, I utilized four different perspectives as lenses to direct my coding. I looked for similarities, differences, frequency and causation (Saldaña, 2016, p.

7). Below, in Table 2, are some examples of how I coded, the first two displays descriptive codes and the second example displays codes that are taken directly from the excerpts:

	Excerpt	Code
Descriptive code	"They work independently, but it is still hard to motivate the students to ask questions in English."	Challenge
	"We focus on strengths and aspects they can work on now, and to develop, and perform better next time."	Formative assessment
Directly code	"In this period, I do not want to put even more *pressure* on the students. They should have the possibility to express themselves without being under the impression of being assessed every second throughout the lecture."	Pressure
	"Oral competence is about how the student manages to express him or herself in the target language. In this case: English."	Oral competence

Table 2: Examples of coding

I wrote down a list of the codes that belonged to each of the documents. This gave me an overview of the codes and the content of the data material. I continued by looking at the material two or three times more, to see if I agreed with the codes I had written. When I was satisfied with the coding, I investigated the data material from a deductive approach. I made a document where I linked the codes to the eight steps in the chain of validation (presented in the theory chapter), to get a better understanding of the meaning of the codes by relating them to threats to validity. This process gave me an impression of what to focus on in the data material. At the same time, I constantly throughout each step of the analysis had the research question in front of me. To guide the focus of the analysis. I also made a document where I

wrote down all my reflections and comments to the data material. These thoughts helped me when defining the categories.

3.10.2 Translation and summaries

The next step was to write summaries of each of the transcribed documents. I went back to the original transcriptions without codes on them. I used the research questions, the document where I had gathered the codes and their connection to the chain of validation when choosing what was important to include in the summaries. At this stage, I translated the statements I used in the summaries, to English. When the summaries were written, I coded them again. This time as well, I wrote the codes in the left margin and my comments in the right margin. When I had coded everything, I compared it to the codes I had written in the previous stages of the analysis. In addition, I made a visual representation (a mind map) of the codes to get an overview of the material and to better see connections between the codes.

3.10.3 Categorization

When I had an overview of the codes and the connections between them, I organized them into categories. I made a new document where I wrote the names of the categories and sorted the relevant data from the summaries under each category. In the following examples, I present two examples of how I worked to identify two of the categories, which I refer to as threats. The first threat I will exemplify is *teacher insecurity*. I discovered this category when analyzing the teachers' answers on similar topics, collected at different stages in the data collection period. A comparison of their answers displayed that they were highly contradictory. In other words, the threat teacher insecurity relates to the teachers' shift in opinion. In chapter 4. I present concrete examples of this threat.

The second example of how I organized the categories is related to the threat *contextual dilemmas*. I defined this category after analyzing many of the teachers' responses related to contextual issues that the teachers' themselves brought up when discussing their own assessment practice. Furthermore, I also analyzed responses where the teachers' mentioned contextual dilemmas when being asked about something else related to their assessment of oral competence. Specific examples that illustrate this threat will be presented in chapter 4. These examples illustrate how two of the categories emerged from the data material. Additionally, I also used the theoretical framework as a guide to when identifying these

threats. This demonstrates the abductive approach that was used in the analysis of the data material.

3.11 Credibility of a qualitative study

3.11.1 Validity

The validity aspect of research addresses if the research presents a realistic image of what it intends to investigate (Gibbs, 2007, p.91). Validity criteria are based on documentation of evidence, logical analysis, and justification of interpretations in relation to the field of study. The researcher's subjectivity in all stages of the research, specifically when analyzing the data material is an example of an aspect that can influence the accuracy of the research. In other words, these aspects threaten the validity of qualitative research (Yin, 2003, p. 35). The influence of this threat is discussed in section 3.2 of this chapter. Another example of a threat to validity is the degree to which the categories function as representatives of the features they represent. This illustrates the need to addressing the validity aspect of research, to evaluate validity and ensure quality.

Multiple validity strategies can be used as methods to display the accuracy of the research. In other words, to demonstration the validity of the research (Creswell, 2009, p. 191). In the following part I present the validity strategies used in this project. Firstly, an intention of this study was to use triangulation of data material as a strategy to ensure validity in the research, by collecting three different sources of evidence that elucidated the field of study from different perspectives. Triangulation of data is a strategy that adds validity to the data material by investigating the coherence between the different sources of data involving the same topic. In this way, one can examine the accuracy of the collected findings, and investigate if they underpin each other (Creswell, 2009, p. 191). Moreover, triangulation of the data material leads to thick descriptions of the data material, which together provides in-depth knowledge on the topic, which is a key concept in case studies. However, since the documents did not contribute to the findings, one can question if triangulation can be used as a validity argument in this study.

Secondly, I have acknowledged my role as a researcher, by reflecting upon my biases and positioning, to provide transparency (see 3.2). Thirdly, I have described some validity concerns in relation to the methods used to collect data material (see 3.7.2. and 3.8). Fourthly, in chapter 4, I present contradictory evidence, which means that I have discussed my findings from different perspectives, to elucidate how different standpoints influence interpretation. By recognizing the role of different angles, one presents a broader image of the findings (Creswell, 2009, p. 192).

3.11.2 Trustworthiness

Reliability, or trustworthiness which is the term that is often used in qualitative studies, relates to the transparency of the research. In other words, the documentation of the different steps of the research process. This includes a documentation of all elements of the data material, methods, result and all decisions taken by the researcher during the project. To achieve such transparency, a well-written case description is required in addition to highlighting the context of the study (Yin, 2003, p. 38). The purpose is to display that the research presents a correct image of the participants' contribution. To create trustworthiness, I have for example in this chapter thus presented the methods and strategies used in this research project.

Furthermore, language use is one of the factors that can influence both the validity and the trustworthiness of the research. If the participants and the researcher do not have the same understandings of the language and terms, it can lead to misunderstandings (Postholm, 2011, p. 170; Yin, 2003, p.92). In the context of this study, this is, for example, relevant if the teachers used the wrong terms when describing their reasoning or if I chose the wrong words when transcribing their utterances. Such mistakes can influence the validity of the interpretations of their utterances and lead to creating a misleading image of the participants. Additionally, the participants can withhold information that they believe can put them in a bad light, or they can try to adjust their answers according to what they think the researcher is looking for in the study (Postholm, 2011, p. 170). These examples demonstrate how various factors can threaten the trustworthiness of the research.

3.12 Limitations of the study

The limited number of participants participating in this study means that one can question the generalizability of this study. However, since generalizability is not the intention of this study, it is irrelevant to interpret the findings as general representations. This study is only meant to give an in-depth description of the reasoning of two English teachers at an upper secondary school. Nevertheless, the limited number of participants will result in less data material than in studies with more participants. This is also a factor that has an impact on the findings in this study. One can for example also question how the findings would have differed if the teachers taught at a different level or if I interviewed more teachers.

As mentioned earlier, in qualitative studies, the research is based on the researchers' interpretation of evidence, arguments, and justification of them (see 3.2). Subjectivity is thus argued to be unavoidable. This is also a limitation of the study, though I have tried to maintain transparency in each stage of the study, to display my position by giving insight in biases, interpretations, and methods.

The role of the method, as a factor that influences the emphasis of the analysis and investigation in this research project, can also function as a limiting factor, because it directs the focus of the research. This means that a different method could have provided different findings.

3.13 Ethical considerations

The research in this project is based on informed consent and voluntary participation. I have worked according to the ethical guidelines provided by the Norwegian Data Protection Official for Research (the Norwegian center for research data). Before conducting the research, I applied to the Norwegian Centre for Research Data where I presented my research project and methods for containing data material (Appendix A). After receiving my approval, I contacted the participants and sent them the information letter containing specific information about their role in the project, and their rights regarding participation, anonymity, and handling of data (Appendix B).

Furthermore, I have taken relevant precautions regarding data storage to ensure complete confidentiality of the participants. Additionally, I have also utilized fictive names for the research participants.

4. Analysis and findings

In this study, I have identified the following five threats to the validity of oral assessment: teacher insecurity, unclear guidelines for teacher classroom assessment, variance in teacher assessment literacy, poorly defined constructs and contextual dilemmas. In the following chapter, I present each of the threats, accompanied by examples from the data material as justification for my arguments. Furthermore, while teacher insecurity is a threat to validity in itself, it is also exacerbated by the four other threats identified in this analysis. Hence, I have highlighted its connection to the other threats when discussing their influence on the validity of assessment of oral competence in English. Lastly, I present a model illustrating the five threats, and a description of the connection between them.

4.1 Teacher insecurity

The first threat I will describe and exemplify is *teacher insecurity*. This is a category that emerged when analyzing the data material because of the shifting opinions in the teachers' answers, on similar questions, at different stages of the data collection period. Their ambivalence communicates an insecurity, which I argue to be present in the teachers' discussion. Below, I have shed light on examples from the data material to display the role of insecurity, by comparing some of the statements from both participants. According to their own statements, their understanding on what to focus on is influenced by the contextual dilemmas they encounter in their classrooms, which also sheds light on the impact of multiple factors on teachers' assessment. Even though the specific reason for their insecurity is unclear, the following examples are meant to shed light on the presence of insecurity their reflections.

The first example displays the ambivalence in Hannah's reflections on what to assess in students' spontaneous interaction, through comparison of three excerpts from the data material. In the first two excerpts she advocates for one aspect, and in the last example, she advocates for the opposite. During the first group interview, the teachers discussed what they looked for when assessing oral competence. One of the teachers, Berit, said that participation in English, in class discussions was what she finds the most important when assessing oral competence. Hannah disagreed, and explained that:

> It is not assessed every day. It is […] noticeable if they move towards Norwegian, but it is not like I sit every day, after each lesson, and make a list of who answered correctly in English.

Hannah points out that participation is not something she assesses every lesson. However, her utterance can imply that it is assessed sometimes, but there is no evidence that concludes this. Furthermore, in the next example from the individual interview, Hannah's reflections give a deeper insight in her assessment philosophy, providing more detailed reflections on the topic. She argues that her assessment competence is something she has developed throughout her carrier. Hannah reflects upon her development: from mainly focusing on students' level of participation, to emphasizing content and language competence when assessing students' performance:

> I think that I have become aware of that answering a question is not enough to prove what you know, orally. […] I think it is a help to think: okay, a little bit more determined: What are they talking about? Have they understood? And can they present own reflections in English about what we are talking about? Can they participate with questions? Earlier, […] I counted how many times they spoke rather than what they said. So, maybe it is more important to think about the content and how they manage to come up with things in relation to what we have discussed.

These reflections imply that Hannah argues to have developed an understanding of what competence to emphasize when assessing oral competence. Her reflections imply that less experience could have been one of the reasons for her choosing to count how many times students participated rather than focusing on what they said.

On the contrary, during the last interview, when the teachers were asked what they believed would make oral assessment easier. Hannah suggested a system where they could score students' level of participation: "Maybe a system, it might sound mainstream: Every week they get some sort of scoring. This week you have participated, this week you did not participate… I do not know…?" In this suggestion, Hannah implies a focus on participation when assessing students. However, she does not seem confident about her idea. Her doubt is present in the last part: "I do not know…?" These examples illustrate mismatches in Hannah's reflections and can be interpreted as a sign of insecurity.

However, the ambivalence can also be a direct consequence of the contextual dilemmas the teachers describe in their classrooms, a class environment characterized by little oral participation. Hence, her suggestion can thus also be a way of exploring various methods and strategies to encourage students to talk. Nevertheless, I also argue that insecurity related to such dilemmas also goes under the category teacher insecurity.

On the other hand, one can also discuss if the contextual factors surrounding the interviews might have influenced her answers. For example, in the last interview, when she discussed the topic together with her colleague, could loyalty towards her colleague have been a factor that influenced her answer. Furthermore, the role of the researcher is also an aspect that should be taken into consideration: can her answers have been influenced by me? For example, by the way I asked the question, or by the situation itself? Did she shape her answers according to what she believed she should answer? Nonetheless, these examples shed light on the presence of teacher insecurity in the data material. The example displays how her thoughts vary during the data collection period.

The second example of teacher insecurity is illustrated in Berit's reflections on the same topic at different stages of the data collection period. Also, in her reflections, and when comparing her reflections, I discovered the presence of insecurity. At this point, it is also relevant to mention that Berit herself explained that she does not have a lot of experience as an English teacher and that she lacks formal teacher education. Hence, multiple times throughout the data collection period, Berit specifically explained that she finds oral assessment challenging and that she feels insecure on how to assess oral competence in English. This insecurity is visible when comparing the content of her answers and is demonstrated below. The examples also address the ambivalence in her reflections. The first excerpt is from one of her diary notes:

> I rarely feel secure in relation to oral assessment of students. I see more after quality than quantity in their contribution. At the same time, I wish to remove the pressure from the students, that sometimes can lead to silence. I am satisfied if everyone participates. Still, I look for differences in variation, vocabulary, and pronunciation.

Berit used the term quantity when addressing the role of participation, and quality when addressing the content and language. In addition to illustrating her insecurity on how to assess oral competence, her reflections direct attention to a contextual dimension that seems to influence her assessment: pressure. She finds content and language important, but on the other hand, participation itself seems to make her satisfied when taking contextual challenges as fare of causing pressure into consideration. Though these are significant issues to discuss, in this category, I will mainly focus on exemplifying teacher insecurity. The role and consequences of pressure and class environment are discussed under the threat contextual dilemmas. Her statement implies that she assesses multiple aspects, both participation, language, and content.

In the second group interview Berit seems to have determined what aspects she finds the most important to emphasize when assessing oral competence:

> I mean that assessment of oral competence goes in two directions: the quality of the contributions, I think, and also maybe the quantity of the contributions. How much it gives to the teaching, even though the quality of the answers might not be that good. Oral competence, assessment of oral competence, for me is that it is not always that they speak as correctly as possible, but that they talk. That is what I emphasize the most. If they work together, that they also answer correctly and right, it is even better. But first and formally that they open their mouth and answer in English. So that is what I assess. When I am assessing presentations, I in fact, stick to the competence aims I find online. And then I concretely have assessment criteria as a baseline for my assessment.

Despite that Berit in her diary notes wrote that she believed that both participation and language reflected students' oral language competence, she underlines that participation is what she finds the most significant when assessing oral competence. However, there seems to be a degree of uncertainty in the opening sentence of her explanation: "I think, and also maybe". Furthermore, her explanation seems to distinguish assessment of presentations from the assessment described above, where she seems to address classroom situations: "that they open their mouth and answer in English". This is also communicated through her last two sentences where she underlines that she follows the competence aims when assessing presentations, in other words: planned assessment. This raises the question: does she not assess according to the competence aims when assessing spontaneous interaction? Moreover, the last sentence seems to indicate that she finds planned oral assessment less challenging,

because she then has “[concrete] assessment criteria as a baseline for my assessment”, which points out the need for more concrete guidelines to facilitate teachers’ assessment of spontaneous interaction.

In these examples as well, it is important to mention that the contextual factors might have influenced Berit’s answers. In the first example she did an individual reflection and in the latter example, she reflected together with her colleague, where both seemed to have changed their answers towards a focus on participation. This can imply that the discussion influenced their reflections, but there is no evidence that proves this. Additionally, as mentioned earlier, it is significant to highlight that she herself addressed the role of her lack of experience as an English teacher, which imply that it is a factor that influences her insecurity. Rather than being a critique of her competence as a novice teacher, this question the engagement of the school to support novice teachers, and the possibilities they provide to employees regarding, for example, in-service training courses. Furthermore, as mentioned above, my role as a researcher can also have influenced her answers.

The examples of the ambivalence in Hannah’s and Berit’s reflections on how to assess oral competence sheds light on the role of insecurity in the data material. Although there might be multiple reasons for the inconsistency in their answers, I have chosen to discuss these examples under the category teacher insecurity. Firstly, the examples display the urgency of addressing teacher insecurity, because of its impact on the teachers’ assessment practice and decisions. Secondly, it illustrates the need for clearer guidelines for oral assessment in English, both to facilitate teachers in their assessment and to ensure a more unified assessment, practice to create a better foundation for validity. Thirdly, it is important to highlight that the teachers’ answers in isolation display an internal discussion that provides insight into not only insecurity, but also their ability to reflect on multiple aspects that are involved in assessment. The role of insecurity in their reflections and its connection to other aspects of the teachers’ role imply that teacher insecurity is a highly significant and influential threat to the validity of assessment.

4.2 Poorly defined constructs

The second threat that I have defined is *poorly defined constructs*. This category emerged after reading the teachers' reflections related to their perception of oral competence and on how this influenced their assessment practice. The discrepancy between their understanding of the construct and the theoretical definitions of oral competence directed attention towards their interpretation of curriculum and the guidelines given by the curriculum. Highlighting the role of the curriculum implies that their understanding of oral competence is perceived as closely related to their interpretation of the guidelines provided by the curriculum. In the following part, I present three examples from the data material that addresses their understanding of the construct. Additionally, I also present an example that illustrates what features the teachers interpret as evidence of students' oral competence, which also directs attention to their understanding of the construct.

The first example from the data material is from a discussion in the first group interview, where the teachers discussed their understanding of oral competence. The example below describes Berit's perception of the construct, where she points out what evidence she emphasizes when assessing students' oral competence:

> Yes, in fact, I do separate between the different subjects, or the classes I have. When I am teaching vocational students, I put an emphasis on the ability to have a conversation. I do not direct a lot of attention to structure, […]it is more about: is it possible for them to ask, can them manage daily situations?

Berit seems to argue that she does not perceive oral competence, in other words, the construct, as the same for all students: "I do separate between different subjects". Her statement implies that even though the competence aims are the same, she does not assess the same competence. It seems like her perception of the construct is different when she is teaching vocational students. As underlined in the example, Berit describes her individual interpretation of the construct: "I do, I put, I am". This can especially seem to be problematic if the teacher believes that the construct, is up for discussion. The example also illustrates the close relationship between the understanding of the construct and the assessment of the construct. Hence, this example illustrates how the teachers' subjective interpretation of the curriculum's description of the construct can threaten the validity in the assessment of oral competence.

However, the analysis revealed that the teachers had different perceptions of the construct. Hannah displayed a more in-depth understanding of oral competence when she is asked to describe her understanding of it:

> Oral competence is about how the student manages to express him/ herself in the target language, in this case, English. Not only express himself, maybe also to listen, to understand... Everything that is connected to oral conversations.

Hannah seems to have an understanding of the construct that reflects the definition of oral competence in the curriculum. However, the discrepancy between their construct understanding can be a consequence of the lack of concrete definitions in the curriculum, which means that the perception of the teacher heavily relies on the teacher's competence in the field. This implies that teachers with less theoretical insight and less practical competence on the field may have a disadvantage because it can be more challenging for them to interpret the curriculum.

The second example of poorly defined constructs is from the second group interview where Berit presents a concrete example that illustrates the challenges she meets when assessing spontaneous interaction. The example is especially related to what evidence she looks for when assessing students' competence. In her example, she discusses her insecurity related to assessment of two different kinds of students:

> So, that (oral assessment) is challenging. Should I look for the one that speaks English all the time, but not so good, or in general education for example, look for another that maybe says something twice a week and the answers are incredibly well formulated? What is important? To compare this and say you get this grade and you get this grade. It is not that easy, in my opinion. What is actually, what is actually important for me?

As Berit explains, she finds it challenging to decide what competence that is more important when assessing spoken language, especially when it comes to giving grades. This insecurity also implies an inadequacy in the guidelines provided by the curriculum, because it gives room for subjectivity related to construct understanding. Furthermore, the example also displays the lack of emphasis on feedback that facilitates progress. The challenges Berit describes imply that she is not certain of what evidence to look for in students' performance and that she does not know what is more significant, high participation or high competence?

As seen in her utterance, the subjectivity is visible in this example as well: “what is actually, actually important for me?”. This implies that the teacher believes that the construct is defined by the teacher, which is especially problematic for the validity of assessment.

The challenges they describe on grading students are illustrated in Figure 2. The hexagons represent the two kinds of students the teachers describe as challenging to assess: one with low competence and high participation, and one with high competence and low participation. The arrows illustrate the direction of desired development towards an idealized student that is less challenging to assess. Furthermore, the arrows also represent the aspect that is not identified by the teacher, namely how to facilitate students development:

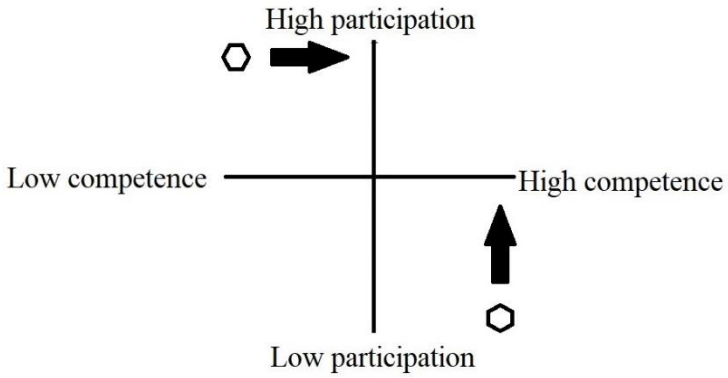

Figure 2: An illustration of the challenges the teachers discuss regarding participation, competence, and development

This means that the student with low competence and high participation needs feedback on how to improve language skills, while the student with high competence needs feedback on how to improve participation. Since they identify this as a challenge, it reflects their weak understanding of the construct. As Berit points out above, she does not know if she should give the student with high competence a good grade, because she mainly looks at participation, and he does not participate that often in class. These examples display the relationship between understanding of the construct, what evidence they consider as representative for their competence and their assessment practice.

Adapting the assessment to students’ preferences were only identified when the teachers discussed students with anxiety. This means the influence of the assessment method is mainly a challenge they seem to relate to students with anxiety, where Hannah suggests giving

options like using audio files as a method for such students to show their competence. This opinion is not something they describe as a possibility for students with high competence and low participation. This reflects that adapting the assessment is something they seem to relate to anxious students, one can thus argue that students with high competence and low participation are more at risk of not being assessed on their competence because they seem to get less attention. Hence, these examples also direct attention to what evidence they consider as representative for students' competence.

The third example that illustrate the role of poorly defined constructs are the teachers' challenges related to capturing a realistic evidence of students' competence. This is also an aspect the teachers themselves problematizes, especially related to the level of subjectivity that influences the interpretation of students' competence. Below is an example from the last interview, where Berit sheds light on the fact that she finds it nearly impossible to ensure that her interpretation of students' competence matches the students' actual level:

> It is not always easy to see […] if the assessment reflects the competence… No, I also think that subject conversations, we often begin with subject conversations… […] Still, it does not really say much about their competence. Because what I mean is that the students.. Since everyone is individuals, they are all individuals that are different, one is extrovert and the other is introverted. It is not always that easy.. And how I ensure it through conversations and explicitness of the assessment criteria and… after that, I reach my own limits. […] No, I do not think it is possible.. At least not in languages, in language subjects.

As Berit discusses, there are many surrounding factors that can influence oral assessment, and she points out that these need to be taken into consideration when discussing if the assessment matches the competence. In other words, when discussing the validity of assessment. She mentions that personality plays a role in the assessment as well. As seen in the example that was illustrated in the diagram above, this is a factor the teachers seem to think a lot about when assessing students. Some students are extrovert and participate a lot, and some students are introvert and do not participate a lot orally. Nevertheless, as mentioned above, in the data material, most attention is directed towards students with low motivation or anxiety. Giving less attention to students with high competence and little participation. The influence of these contextual dilemmas on their understanding of the construct is visible in many of their

discussions, especially when they compare their assessment practice in the vocational classes and their general study classes.

However, what these examples also display, is that the teachers are aware of their limitations. The fact that they share these examples demonstrate that they themselves want to discuss their perceptions of the construct and their assessment practice. This implies that the teachers are open for a discussion on how to improve their practice, and on receiving guidance on how to assess oral competence in English.

Furthermore, their answers give insight into the complexity of their job, where multiple factors need to be taken into consideration to understand the full picture of what influence their decisions and perceptions of the construct. By raising the question on how to assess challenging students the teachers show that they are aware of how students' personality can influence their perception of their competence. Hence, the highlighted examples clearly illustrate the need for more concrete definitions of the construct, to ensure unity in assessment, but also because general guidelines put too much responsibility on each individual teacher.

4.3 Contextual dilemmas

The third threat I describe is *contextual dilemmas*. At many stages of the data collection period, the teachers brought up contextual dilemmas as justification for their decisions related to assessment of oral competence. Below I have highlighted examples from the data material that sheds light on the role of such dilemmas for teachers' assessment of oral competence in English. The main challenges that the teachers themselves identified as problematic were the differences in motivation, language skills and learning environment between the two study programs, vocational studies and general studies.

The first example describes Hannah's distinction between the classes she teaches. In one of her utterances, she highlights that English is the only language used in the general studies classes she teaches. Her emphasis on general study implies that she does not only use English in vocational classes. This illustrates that variance between student groups influence her

classroom practice. Her colleague continues by describing that anxious students in the vocational classes are the reason for the distinction between the language use in the classes:

> We are working on safety because when we started with this class, almost no one was willing to speak English. […] So, we figured, okay, let's be on their side. So, we lowered the level of ambitions and told them that we would go step by step. We started with speaking. It was only a couple of the students who wanted to say something in English.

On the one hand, these examples illustrate how contextual issues can result in prioritization that can influence the learning conditions for the students. In this case, the teachers' decided to emphasize social aspects over oral language use in class. The consequence of such a decision can be both positive and negative, depending on each individual student, because lowering the demands means to reduce the challenge to increasing students' competence, confidence, and motivation. As I have presented in other examples above, these contextual dilemmas seem to have influenced their assessment practice as well, leading to a strong focus on participation when assessing students' oral competence, because of fear of discouraging students.

On the other hand, the teachers describe a situation where they feel forced to reduce the language level and demands because of the students' lack of motivation and little participation in English. This aspect is also something that needs to be taken into consideration to understand how the dilemmas they meet influence both teaching and assessment. The teachers decide the level of importance based on what they perceive as more important for the students' language development. In the example above, it seems like the teachers have evaluated that their biggest challenge, that prevents language development, is that the students do not use the target language because they do not feel safe. Their priority is thus to create a learning environment where the students feel safe, and by this form positive learning conditions as a foundation for their development.

Nevertheless, Berit herself points out that she finds such decisions and prioritizations challenging. She communicates that she is insecure on what decision that is more beneficial for the students' development: "to find the right balance I think, between when it is okay to switch to English, or to Norwegian. Sometimes it is necessary […] So, that is challenging. I think oral assessment in general [is challenging]."

This implies that Berit is aware of how prioritizations can influence teaching and that deciding what is most important is not always easy. This display how the responsibility put on teachers in relation to such significant decisions also influence their confidence and can lead to teacher insecurity.

The examples above highlight threats to validity that are caused by contextual challenges surrounding the teaching situation and display how multiple aspects influence the decision of the teacher. In other words, looking at one action or decision in isolation will in not present the whole picture of the aspects that influence teachers' decisions when assessing students' oral competence in English. The impact of contextual dilemmas on teachers' oral assessment practice display its role as a threat to the validity of oral assessment.

4.4 Variance in teacher assessment literacy

The fourth threat to validity of oral competence in English that is identified in this analysis is *variance in teacher assessment literacy*. The data material illustrates a distinction between the teachers' reflections on assessment practices related to both planned and unplanned speech in the context of assessment of oral competence. Below, I present three examples that display the role of this threat in the data material.

The first example involves reflections on the challenges related to the content of the feedback provided to students. Hannah describes that keeping up students' motivation is one of the most important considerations for her as a teacher:

> The challenge is maybe that we have to be as positive as possible because this is a very touchy area. And if you give them the least hint that […] this was not really that good… So, lifting up the positive aspects, even though we ourselves mean that this was a little bit pathetic, but it is possible to get better.

Hannah's reflection indicates that students' low self-esteem in oral English leads to her giving less concrete feedback, she writes: "we have to be as positive as possible". This can be interpreted as excessive focus on positive feedback, even though the performance was not good. If this is the case, positive feedback that is not concrete can lead to little formative

feedback that can help the student develop. On the contrary, it can also mean that Hannah emphasizes positive reinforcement, which is closely related to formative feedback. Either way, the example describes a challenge Hannah herself has pointed out as problematic when discussing her assessment practices.

This example also demonstrates a relevant motivational challenge that is present in multiple discussions in the data material and how it influences the teachers' assessment practices. As the example above display, some of their decisions seems to be influenced by the fact that the teachers do not want to decrease the students' motivation if they are trying to use their oral language. Berit for example explains that she has a student that participates a lot in class, but his language is full of errors. Nevertheless, she explains that: "Still, I consider it very positive that he at least tries, and he works on his own progress, own improvement.". The latter example illustrate that the teachers focuses on motivating the students. However, the first example displays how this kind of focus can lead to feedback that is not addressing the aspects the students need to work on to develop his or her competence. This means that this kind of decision can lead to decisions that decrease student development.

Furthermore, the teachers explain that the feedback they give is usually written feedback, both when it comes to feedback on unplanned- and planned oral language. The teachers explain that they usually post feedback on *itslearning*, so that the students can read it individually. They do not mention what kind of feedback that is given orally, or if feedback is given orally in class. The teachers themselves were directly asked to discuss what kind of feedback they emphasized for spontaneous interaction. Hannah explained that they often use itslearning to give feedback on classroom talk, the example displays her emphasis on positive reinforcement:

> Sometimes I have… and you have also [looking at Berit] posted it on itslearning, if we feel that they have had excellent participation in class. It is possible to post a comment: really nice oral participation today. And it is only the student that sees it, and it is a GOOD mark.[…] We can of course not do it too often, but we have done it.

This display that Hannah does not mention what kind of spontaneous individual feedback she provides in classroom situations, but that does not mean that she does not give feedback in the

classroom. What her explanation seems to reflect is that the feedback is mainly general and positive. Both teachers seem to agree that positive feedback encourages and motivates the students. They continue the discussion by elucidating that they experience positive effects of giving general positive feedback to the students, if they have done a good job, outside of class as well. These examples illustrate that the aim of motivating students to speak English is considered as more significant for the teachers than to focus on concrete aspects to develop their oral competence. This focus can be a result of the dilemmas that the teachers meet in the context of the subject, and the choices they take based on their perception of what is more important. One can also direct attention to the fact that this can be a result of their assessment competence, unclear guidelines in the curricula or their own teacher education.

However, when discussing planned assessment, the teachers seem to highlight other aspects in their feedback. The teachers agree that providing formative feedback with concrete examples is important. Hannah explained that:

> we focus on strengths and aspects they can work on now, and to develop and perform better next time […] I would say that in formative assessment, we use it to get them to understand what they should work on, and hopefully get better.

This display how their assessment practice differs when the assessment is planned. Moreover, this example illustrates the connection between the educational principles and their practice, where the assessment practice Hannah describe is according to what is advocated by educational policies. The difference between planned and unplanned assessment can imply that the teachers are insecure on how to assess spontaneous interaction.

Furthermore, Hannah continues by describing how she provides feedback to the students:

> I at least, write thorough feedback on itslearing on what was good and on what was not that good, and how they can get better until next time. Then, they receive messages, and when I take them out, and talk to them, we go through it and I ask questions. Only in Norwegian, then, if they have understood what it says, and then I also hear if they did not get it. And if they [unclear] they have a possibility to say something.

Here as well, *itslearning* is mentioned as an important source for communication of feedback to students. Such communication is usually one- way communication. The positive aspect of

posting feedback individually on itslearning is that the students can read it multiple times and go back later if they forget the content of the feedback. Nevertheless, the downside is that the students do not get the chance to discuss their feedback right away, if they do not understand or if they need clarifications.

Hannah herself gives the impression of having reflected upon these limitations herself since she mentions that she arranges a conversation with students after they have received the message. Furthermore, she seems more confident in her answers when discussing planned assessment, and her concrete criteria for the content of the feedback is an example of the distinction between her confidence in planned assessment and unplanned assessment. This distinction also displays how assessment literacy influence decisions and the impact of the assessment, which also addresses the role of variance in teacher assessment literacy as a threat to validity of oral assessment.

4.5 Unclear guidelines for teacher classroom assessment

The fifth threat that is identified in this analysis is *unclear guidelines for teacher classroom assessment*. I identified this threat after discovering a connection between the teachers' insecurity and their interpretation of the competence aims and the curriculum. The outcome of the unclear guidelines is for example visible in their discussion of the relationship between competence and grades.

In the following two examples, unclear guidelines for teacher classroom assessment seem to be a factor that influences the teachers' insecurity of the role of oral competence in the students' final grade. As the teachers themselves highlight, they find the final grades especially problematic in vocational studies and the first year of general studies, where oral and written competence are represented in one grade. The lack of guidelines for how much weight that should be given to each of the competencies means that the decision relies on subjectivity and individual interpretations of the curriculum. Berit's utterance in the group interview is an example of the threat that unclear guidelines for teacher classroom assessment can cause validity of oral assessment:

> I think it is hard because I need to decide how much influence the oral part should have. […] if the student hand in incredible good texts, and I know that they are good in English, but almost never say anything voluntarily, or answer really short, it is hard to say that.. if I would say 50-50, then it would pull him down. Oral: two, written: six, that turns into a four. So, I feel that that would not be fair either.

Firstly, the role of Berit's subjectivity is highly present in her reflection above, already in the first sentence: "I need to decide how much influence the oral part should have". This implies that she believes that it is the teachers' responsibility to decide the role of oral competence in the final grade. If this is the case, missing guidelines will lead to multiple meanings of such grades. Secondly, she describes a situation where a student has low participation, and she problematizes this because she knows that this student has high competence in English. On the one hand this points out the major weight she gives to participation as an evidence of oral competence, and on the other hand what I discussed earlier, little focus on adapting assessment situations to students with high competence and low participation. She seems to argue that the student's competence is not visible through his participation, thus she cannot justify a grade based on his participation because of her knowledge of his skills. Despite that she draws these conclusions, that participation is not an evidence of his competence, she does not suggest the use of other methods to document his competence, nor does she reflect upon the problems with using participation as an indicator of knowledge in general.

The challenging aspect Berit seems to advocate is that she feels it is not fair that oral competence should count as much as written competence, her justification for this conclusion is that she knows that this student is "good". This is a concrete example of a situation that is caused by unclear guidelines for teacher classroom assessment. If the guidelines were clear, there would not be any discussion on how much the oral part should count.

On the other hand, Hannah interprets the situation Berit describes differently, she suggests taking a formative approach to it, and to focus on how this student can improve his oral competence. She suggests that if the student has low competence in one of the aspects, one could use it as a motivational factor for him to improve. The examples illustrate two different interpretations of the same situation, and how this leads to two very different decision, which

also influences the consequences of the assessment upon the student. These examples also demonstrate of how much responsibility each teacher is given when guidelines are unclear.

Furthermore, Berit seems to disagree with Hannah's suggestion, and describes that she despite her suggestion, consider the final grade problematic:

> Formative assessment, yes, but with a grade. They only get one number. […] We do it through subject conversations [fagsamtaler], the formative assessment, but the result is then, that if they do not follow it, I still need to decide…

The question Berit seems to ask is how she can give a grade that matches their competence in both aspects. Additionally, this is also an example of how her assessment competence influences her choices. Berit's reflection seems to express that formative assessment is something she connects to subject conversations: "we do it through subject conversations". This raises questions about her understanding of formative assessment; does she argue that she is only giving formative feedback through these conversations? Moreover, the examples above imply that the competence that is reflected in grades are defined by the individual teacher. Unclear guidelines for teacher classroom assessment is addressed as one of the reasons for this perception, which demonstrates the relevance of identifying it as a threat to the validity of oral competence.

4.6 Connections between the threats and teacher insecurity

The findings above illustrate a connection between some of these threats, which also sheds light on the complexity of the aspects that influence teachers' assessment of oral competence in English. Teacher insecurity is especially emphasized as a central threat and it is argued to be both a result of the other threats and a threat in itself. The connection between the threats and their relation to teacher insecurity is displayed in the model below:

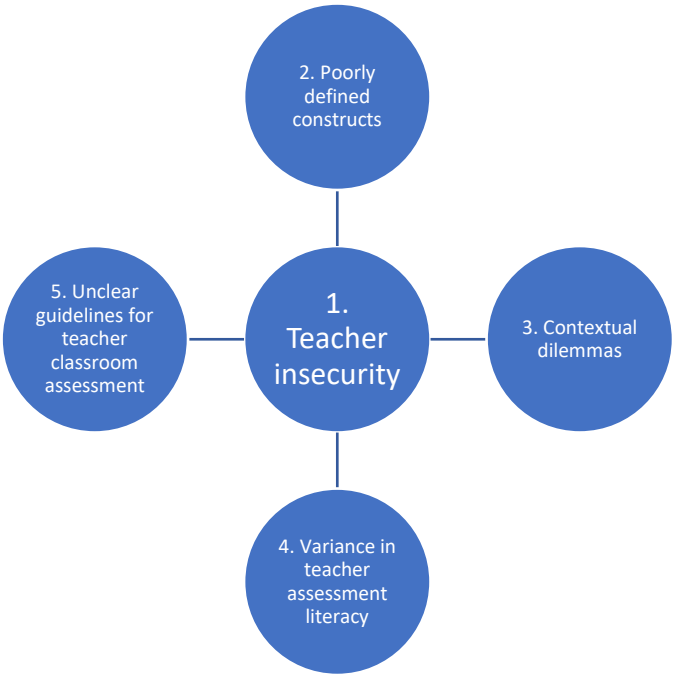

Figure 3: The connection between the five threats

Teacher insecurity is illustrated as a central threat to the validity of oral assessment. The model display how the other threats are linked to teacher insecurity. However, the model does not explicitly illustrate the connection between the other threats, but this is explained in the sections above. In addition to its connection to the other threats, in itself, teacher insecurity is present in multiple reasonings in the data material, both explicitly and implicitly. Below I have specified its relevance for each of the four other threats, where the numbering is linked to the figure above:

1. The findings display that poorly defined constructs and teacher insecurity is connected because of the lack of in-depth definitions of oral competence in the curriculum and specific descriptions on how to assess oral competence. This means that interpretation heavily leans on the individual teacher's perception of the curriculum. This can lead to insecurity because it puts a considerable amount of responsibility on the teacher since the teacher's interpretation of the construct influence both teaching and assessment. How can teachers know if their interpretation is correct if the curriculum does not contain in-depth descriptions of the construct?

2. As the findings illustrate, contextual dilemmas influence teachers' decisions regarding assessment practice and emphasize when assessing oral competence. This leads to insecurity because the teacher needs to prioritize what is more important, which means

that different teachers probably will prioritize differently. The result is multiple interpretation and practices of the same competence aims and constructs.

3. Variance in teacher assessment literacy leads to teacher insecurity because assessment competence influences the decisions taken by the teacher in lessons and assessment and his or her perception of the construct. Such differences lead to nuances between teachers, which can lead to insecurity on what the "correct" way to assess, for example, oral competence in English is.

4. Unclear guidelines for teacher classroom assessment can lead to teacher insecurity because it demands each individual teacher to interpret the guidelines before using them in own teaching. As mentioned above, since teachers are different, this will lead to multiple interpretations of the same construct, which can lead to insecurity. As poorly defined constructs, unclear guidelines also put a lot of responsibility on the teacher, because the teaching will be based on his or her perception of the guidelines.

4.7 Chapter summary

In this chapter, I have presented and exemplified five threats to the validity of oral assessment. These threats were identified when analyzing the data material, and are presented in the list below:

1. teacher insecurity
2. poorly defined constructs
3. contextual dilemmas
4. variance in teacher assessment literacy
5. unclear guidelines for teacher classroom assessment

In the next chapter, I will go deeper into what the findings mean for the validity of oral assessment in English. Furthermore, I will also discuss what they imply about the relationship between educational principles and teachers' practice of oral assessment in English.

5. Discussion

The findings identified five threats that influence the validity of oral assessment in English. This chapter discusses how we can use this knowledge to gain insight into the field of study, and more specifically discuss how these threats have an impact on assessment of oral competence in English and how the interplay of educational principles and teachers' practice influences these threats.

Firstly, I discuss the findings in light of the theoretical framework of construct validity, where I provide examples from the data material of both construct underrepresentation and construct-irrelevant variance. Secondly, I direct attention to the chain of validation and discusses evidence from the data material that confirms the theories in the framework. Thirdly, I highlight aspects of the findings that are not represented in the current model of the chain of validation. Fourthly, I suggest an expansion of the model. Fifthly, I discuss possible consequences of reducing some of the threats. Sixthly, I provide a summary of the entire thesis. Finally, I present the implications of this study and suggest further research on the field.

5.1 Confirming the importance of construct validity

The theoretical framework of construct validity describes that construct validity refers to the trustworthiness of interpretations and arguments provided by the teacher. Hence, discussing construct validity involves evaluating the trustworthiness of the evidence provided by the teachers as justification for their interpretations (Messick, 1996, p.744). The findings display examples of both construct underrepresentation and construct-irrelevant variance. The presence of these threats to construct validity confirms the importance of acknowledging these aspects as relevant threats to assessment of oral competence in English. In addition to displaying the coherence between the theoretical framework on threats to validity of assessment and the findings, this chapter address how these threats can influence teachers' assessment of students' oral competence in English. Therefore, below, I discuss some specific examples of the findings that illustrate construct underrepresentation and construct-irrelevant variance.

5.1.1 Construct underrepresentation

As presented in the theoretical framework, construct underrepresentation relates to situations where the assessment leaves out important aspects of the construct (Messick, 1995, p. 742). To exemplify construct underrepresentation in the data material, I highlight the example where Berit described that she especially finds overall achievement grades challenging because both oral and written language competencies are represented in one grade in the first year of general studies and both years of vocational studies in upper secondary school. She illustrated this through an example of a student with high written competence, but that did not display the same level of knowledge when speaking orally. Her conclusion was that she found it unfair if the student's oral competence influenced the overall grade to the same degree as the student's written competence because it then would result in a lower grade than what she believed this student "deserved". This illustrates how a biased interpretation can influence the evaluation of a performance and can lead to construct underrepresentation (Crook, Kane & Cohen, 1996, p. 276- 277).

If Berit in this situation decided that the student's written competence would count more than the oral competence, this is an example of construct underrepresentation, because the grade is supposed to reflect the written and the oral competence combined. Thus, one can argue for the importance of both aspects having equal influence on the grade. Furthermore, even though Berit seems to believe that the assessment does not reflect his competence, she does not seem to question the validity aspect of the grade. Her main concern is only the role that oral competence should play in the final grade. This an example that also displays the relevance of addressing the influence of teachers' understanding of students as a relevant validity concern (McMillan in Brookhart, 2005, p.10).

However, since the share of how much influence the two language aspects should have is not specified in the curriculum, it means that this is a decision that relies on the teacher's interpretation. If the teacher does not fully understand what construct the combined grade is supposed to represent, it is challenging to ensure validation (Crooks, Kane & Cohen, 1996, p. 267). However, because the construct definition is unclear, it is also possible to justify a grade where written language, for example, counts 70 percent and the oral competence counts 30 percent. The question is then if it is possible to ensure validity in such combined grades when

the role of oral and written aspects is not specified? This is especially problematic since the grades are used to compare students' competence.

If grades do not represent the same competence, one can thus argue that they should not be used to compare students when applying for higher education. As the example above demonstrates, if the curriculum relies on teachers' interpretation in contexts like these, one can argue that grades only are evidence of teachers' interpretation of the construct (and the curriculum). Inconsistency between raters is also identified as a threat to validity related to scoring of student competence by the chain of validation (Crooks, Kane & Cohen, 1996, p. 272). This example sheds light on the impact of the teacher on the assessment, in addition to the need for more concrete guidelines and definitions of the construct, to ensure construct validity in assessment of oral competence in English.

5.1.2 Construct-irrelevant variance

The findings also displayed an example where Berit and Hannah discussed assessment of students' level of participation as evidence of students' oral competence. If the teachers decided to base assessment of oral competence on the level of participation, this is an example of construct-irrelevant variance, which means that the assessment emphasizes aspects that are not relevant for the construct it is supposed to assess (Messick, 1995, p. 742). The irrelevance of basing grades in oral competence in English on participation is also described in § 3-3 in the Regulations of the Education Act highlights, which states that the only subject where the level of participation can be used as evidence of student competence, is physical education (Regulations of the Education Act, 2006).

The teachers justified an emphasis on the level of participation when assessing oral competence because their students did not participate in class, and they, therefore, felt it necessary to provide positive feedback, to encourage them to talk. Firstly, choosing to work on the learning environment is supported by the theories on learning environments, because the classroom context influences the language of the students (Brookhart, 2005, p. 6). Secondly, this is in line with what current research on the field also display, teachers seem to prioritize the learning environment when assessing oral competence (Svenkerud, Klette & Hertzberg, 2012, p. 44). Nevertheless, this is also an example of the impact the threat

contextual dilemmas can have on construct validity, if teachers' prioritizations lead to emphasizing irrelevant aspects of the construct.

Furthermore, as Figure 2 display, another consequence of emphasizing participation is that it can decrease the attention to feedback on aspects that the student needs to improve to develop competence, because the student is given feedback or a grade that does not represent his oral competence. In this way, construct-irrelevant variance leads to assessment that can prevent progress in the actual construct. This is an example of a negative consequence that can follow an assessment if the student does not receive feedback (Crooks, Kane & Cohen, 1996, p. 279). As seen in the discussion on assessment of two different kinds of students, Berit did not mention student progress. However, Hannah interpreted the exemplified situation differently and was more focused on how to encourage progress and development. As mentioned earlier, this also displays the impact of variance in teacher assessment literacy, where students in one class are assessed differently than students in another class because of variance between teachers' competence.

Another problematic aspect regarding assessing the level of participation is that participation can be closely related to personality. An extrovert student can thus be given an advantage because he or she likes to talk. This kind of student can, therefore, be able to get a better grade than an introvert student, even though the competence of the extrovert student is lower. Hence, justifying students' oral competence through their level of participation will influence the validity of the assessment because it gives advantages to certain students. This illustrates how construct-irrelevant variance can lead to serious threats to construct validity.

However, participation in class or on a test can also be argued to be a necessary condition for assessment of oral competence. If the student does not display his or her competence, it is impossible to assess it. Hence, it is important that the teacher creates situations where the students are given a chance to demonstrate their competence. Adapting the assessment to students' needs is something the teachers themselves brought up in their discussion, but it was only mentioned as an option when discussing how to assess the oral competence of anxious students. Choosing to adapt the assessment according to contextual dilemmas such as anxiety instead of focusing on maintaining construct validity is in this way also justified by the theory

which underlines the importance of considering "social and situational needs" when assessing language skills (Luoma, 2011, p. 28).

Moreover, one can also argue that their discussion demonstrated a willingness to reward students that participate orally and punish students that do not participate orally. This can be a consequence of the lack of oral participation that they describe as present in their classes. It seems like the teachers believe that by giving a weak student with high participation a good grade, they can encourage students to participate orally. Therefore, it is arguably easier to give a weak student that participates in class a better grade than what his or her competence indicates, than giving a student with high competence and low participation a good grade. Rewarding students' participation, is in addition to functioning as positive reinforcement for the individual student, a factor that can be used to motivate and encourage all students in a class to participate orally. This is an example that demonstrates how contextual dilemmas in the classroom can influence teachers' prioritization, and lead to an emphasis on features that threatens the validity of the assessment (McMillan, 2013, p.102).

5.2 Following the steps of the chain of validation

The theories on the chain of validation underline the importance of understanding the purpose of the assessment to be able to ensure validity in the assessment (Crooks, Kane & Cohen, 1996, p. 267). As the examples above illustrate, insecurity regarding construct understanding can, for example, make it challenging or even impossible to ensure validity in all stages of the assessment, because the purpose of the assessment correlates with the decisions taken in the assessment process. Purpose understanding (or construct understanding) is thus the baseline for validity of assessment. Many of the threats addressed in this study underpin the threats that are presented in the chain of validation. This display the correspondence between theory and practice. Below are some specific examples of the findings that confirm the relevance of some of the threats that are pointed out in the chain of validation.

5.2.1 Correspondence between theory and practice

Both policy documents and theories on oral competence shed light on the role of the context as a factor that influences the language of the learner. In addition, it points out that contextual dilemmas can function as threats that influence both the language of the students and the assessment in itself (Brookhart, 2005, p. 6; Luoma, 2011, p. 22; Norwegian Directorate for

Education and Training, 2013). The role of the context upon the language of the learner is especially highlighted as a threat in the first stage of chain of validation, administration, because students adapt their language to the context that surrounds the assessment (Crooks, Kane & Cohen, 1996, p. 270-271; Luoma, 2011, p. 27).

The presence of contextual dilemmas as threats that emerged in the data material displays the connection between theory and practice and confirms the perspectives addressed in both the chain of validation and the theories on contextual influence on oral language and assessment. This highlights its influence on the purpose, the situation and the relationship between speakers (Luoma, 2011, p. 22). Moreover, the findings also acknowledged motivation and anxiety as threats to validity of oral assessment. Such threats can result in a performance that gives a misleading image of students' performance (Crooks, Kane & Cohen, 1996, p.270). This demonstrates the importance of reducing contextual threats to decrease the threats related to administration of the assessment.

In addition to focusing on participation when assessing oral competence, the teachers discussed how the contextual dilemmas influenced their teaching of oral competence. The findings displayed that one of the major consequences on the teachers' decisions was that they decided to lower the level of ambitions and use more Norwegian in the English lessons. One the one hand, maybe the impact on students that do not want to talk English will have positive consequences in terms of motivation because they will feel safer in class since they can talk in Norwegian. On the other hand, what about the students that already speak English? What is the impact on their language learning? In this way, one can also argue that this decision can lead to negative consequences because it can prevent student learning. This display that decisions can have positive consequences for the motivation of some students, while at the same time have negative consequences for the motivation of others, where both aspects display possible threats to validity (Crooks, Kane & Cohen, 1996). Either way, this reflects how teacher believes about appropriate instruction and definitions of the construct influence their teaching and goals of the assessment (McMillan in Brookhart, 2005, p.10).

Firstly, if students do not practice their oral language in class, it also means that they will receive less feedback that can help them develop their language. How can they, for example, develop their oral interactional skills if they do not practice them? This illustrates how this decision can make it more challenging for students in the long run, because they will not

become used to speaking orally, which can make them even more nervous when they finally have to speak, for example, in a test situation. This decision can thus be related to threats identified in relation to the impact stage of the chain of validation, because it reduced the possibility of "helpful feedback to improve learning", which the model connects to the threat "positive consequences not achieved (Crooks Kane & Cohen, 1996, p. 279).

Secondly, if the teachers decide to talk Norwegian in class, the students will also receive less input in English. This is also a condition that can make it more challenging for them to learn the language. Nevertheless, as mentioned above, for students that are especially nervous, it can be helpful to lower the pressure, so they can concentrate on increasing their confidence so that they become safe enough to use the target language.

However, it is questionable if the entire class benefits from such a decision. An alternative could, for example, be to divide the students into smaller groups when speaking orally, so that they can practice their language together with students they feel safe around. Nevertheless, these perspectives display how contextual dilemmas can lead to trade-offs that threaten the validity of the assessment, because it requires teachers to prioritize what is more important (McMillan, 2013, p.102). If such decisions lead to learning situations that hinder the progression of some learners, one should question if it is the right decision.

On the contrary, if the teachers decide to only focus on subject development and not work on the classroom environment it can lead to decreasing the development of students, leading to for example low competence or low motivation. This could therefore also threaten the validity of the assessment. Already in the first stage of the assessment process, administration, these aspects are identified as major threats (Crooks, Kane & Cohen, 1996, p. 270). In addition to influencing the task performances, these validity threats would influence the validity of all upcoming stages. Either way, this illustrates how the contextual dilemmas teachers' meet in the classroom have an impact on decisions and consequences of the assessment and leads to decisions where they often need to prioritize one aspect, and down prioritize another.

5.2.2 Feedback practices

The teachers describe that they feel that they have to give mostly positive feedback because oral English is "a touchy area". This highlights the influence of contextual dilemmas on

teachers' assessment. Discussing this in light of the theories on the chain of validation it is connected to the decision and the impact stage. The decision stage addresses when the teacher needs to decide what feedback to provide or "what actions to take as a result of the judgments" (Crooks, Kane & Cohen, 1996, p. 277). On one hand, positive feedback, despite that the performance was not good, does not provide any information about how the student can improve own performance. This means that it can decrease the development of the student.

On the other hand, deciding to emphasize positive feedback is a consequence of students' lack of motivation. Hence, as discussed earlier one can also justify providing positive feedback because of the importance of motivation for the ability to learn. However, if this is the feedback the teachers provide every time, one can question if they are doing the students a disservice, because of the consequences that follow the decision in terms of students' development, because it limits the concrete guidance they receive that can help them develop.

This example also illustrates how one decision can lower one threat, but at the same time raise the level of another, which demonstrate the complexity of assessment, and the consequences of both the contextual dilemmas the teachers meet and the impact of their decisions (McMillan, 2013; Pøitz & Borgen, 2010). This means that the teachers need to evaluate what prioritization that threatens development the most. Weak assessment literacy and poor understanding of the construct can lead to decisions that harm development more than it facilitates it. This illustrates the need for better guidance, to help teachers make good decisions in situations where they need to prioritize.

The responsibility put upon teachers when meeting such dilemmas demonstrate some of the challenges of assessment and reveals why teachers can feel insecure. Taking the wrong decision can have large consequences. For example, if an anxious student is forced to talk orally in front of the class, or if a student with high competence do not get to practice his oral language. The question is what tools or guidelines that are necessary to facilitate teachers' prioritization when meeting such influential contextual dilemmas?

5.3 Identification of new threats

In addition to providing evidence that underpins current theories on threats to validity, this study also contributes to strengthening our understanding of the chain of validation, through addressing threats that are not mentioned in the theoretical framework. By highlighting these aspects, this project contributes to expanding current knowledge on validation of assessment.

5.3.1 Teacher insecurity

Teacher insecurity is a threat that is not addressed by the theoretical framework of the chain of validation (see 2.1.3), but that emerged as a central threat in the data material of this thesis. However, the role of the teacher is visible in the stages of the chain of validation, but it does not explicitly address teacher insecurity as a threat to validity. In the current model, decisions of the teacher are described from a student-centered perspective, and it does not discuss the factors that influence the teacher' interpretations and decisions.

The findings in this project reveal the importance of also highlighting the teacher perspective, to acknowledge the role of the teacher as an interpreter of knowledge, assessor and validator, and how teachers influence the assessment of students. When teachers meet situations where they need to prioritize what to focus on, their decisions influence the entire assessment. In other words, identifying teacher insecurity as a threat can help expand our understanding of the chain of validation and our understanding of assessment literacy, because it sheds light on the teacher as an influential agent for the validity in oral assessment. Excluding this threat from a theoretical model of threats to validity can give a misleading illustration of classroom assessment. In the following part, I specifically discuss the relevance of this threat to validity of oral assessment.

The intention of pointing out the ambivalence in the teachers' statements was as mentioned earlier to justify the influence of teacher insecurity as a threat to the validity of oral assessment. The examples that were discussed displayed a discrepancy in the teachers' answers. Nevertheless, despite that the teachers did not provide concrete arguments when discussing oral competence, it does not necessarily mean that they lack knowledge on the field. The knowledge can be tacit, in other words: the knowledge can be unconscious, and articulating it may thus be challenging (Polanyi, 1996, p. 5). This means that teachers can possess more knowledge than what they managed to communicate. Yet, based on the findings

of this study, I advocate for the influence of teacher insecurity on all stages in the validation process because insecurity related to construct understanding and own assessment literacy influences the decisions taken in all stages.

Additionally, as the findings display, insecurity related to contextual dilemmas can also result in prioritization that decreases the validity of the assessment. This display that teacher insecurity is a key issue in the validation process. Hence, I argue that teacher insecurity needs to gain more attention in validation theories, to provide a clearer image of the assessment process. This means that neglecting the personal characteristics of the teacher, as both assessor and validator, is a threat in itself. However, as mentioned above, it is important to highlight that such aspects already are mentioned in the current model (for example poor pedagogical decisions), but their attention is too modest in relation to the relevance of its influence on the decisions taken by teachers (Crooks, Kane & Cohen, 1996, p. 278).

5.3.2 The role of the curriculum and teacher educational programs

The threats poorly defined constructs and unclear guidelines represent the values communicated by the curriculum. The content and coverage of such guidelines are argued to be a threat that influences the validity of assessment because they require interpretation. Despite that current research addresses the inconsistency in the interplay between educational principles and teachers' classroom assessment, such specific threats are not covered by the theoretical framework of the chain of validation (Brookhart, 2005, p. 11; Stiggins & Conclin, 1992, p. 20). This implies a need to highlight how these aspects also influence assessment, to gain a deeper understanding of the validity of oral assessment in English and a better understanding of the complexity of assessment.

Looking back at the definition of oral skills in the English subject curriculum (see 2.2), it demonstrates the lack of concrete guidelines and the general nature of the curriculum. Although the definition addresses key aspects of oral competence: production, reception, interaction and mediation, it does not provide information about the specific skills or competencies that underlie these language competencies (Norwegian Directorate for Education and Training, 2013). This leads to questioning why there are so few concrete guidelines for how to assess oral competence?

The absence of specific guidelines is an example of the challenges teachers meet when assessing oral competence in English, a challenge that is caused by educational policies. The definitions provided in the theoretical framework in this thesis display that it not only gives the opportunity to personal interpretation it requires personal interpretation. In other words, the foundation for the education that is offered to students is based on teachers' perception and understanding of the curriculum. This means that it is underpinned by the teachers' theoretical competence, experiences, and biases. In other words, the general guidelines are in themselves a threat the validity of assessment. This confirms the theory on the impact of teachers upon the validity of assessment, which addresses that "teacher beliefs, teacher instructional practices, and teacher understanding of both the subject matter and students are relevant validity concerns" (McMillan in Brookhart, 2005, p. 10).

However, it is also relevant to shed light on the role of teacher education when discussing the guidelines and definitions in the curriculum. Since they require teachers' interpretation, one should discuss what competence the curriculum expects that teachers have? If it requires a certain level of competence, is it the responsibility of teacher education programs to ensure that teachers have the knowledge they need to interpret the curriculum? If this is the case, it means that it requires that the teacher education emphasizes development of assessment literacy, to ensure teacher interpretation. How can we ensure that teachers have the required knowledge and competence that the curriculum demands?

The interplay between the three factors: teacher education, teachers and the curriculum demonstrate the importance of ensuring quality in all these aspects to create a solid baseline for the validity argument. For example, the importance of including teacher education programs is because of their impact on teachers' competence and therefore their interpretation of the curriculum. Seeing this from a larger perspective, this means that validity of oral assessment in English also is influenced by the quality of teacher education programs and the curriculum. As the findings in this study display, these aspects display themselves as influential to the validity of oral assessment in English, and I thus advocate for the relevance of expanding the current model of the chain of validation.

5.4 Expanding our understanding of the chain of validation

Since the current model of the chain of validation does not include an evaluation of the validity of the guidelines provided by the curriculum or the role of teacher education programs for teachers' classroom assessment, it implies that the framework assumes that these are valid. This means that the validity of the model itself is based on full validity in the education provided to teachers and validity in the guidelines and definitions provided by the curriculum. These are aspects that contradicts the findings of this research project and are also highlighted as problematic by other research projects on the field (Prøitz & Borgen, 2010, p. 112).

Therefore, to fully illustrate the validity aspect of assessment, I argue for the need to direct attention towards the curriculum, because of its influence on classroom assessment and because of its role as a guide that set the premises for education. If the curriculum does not provide applicable definitions and guidelines, I argue that it will influence all aspects of assessment of oral competence. Secondly, I argue to direct attention towards teacher education. If it does not emphasize the competence that is needed to interpret the curriculum and to take valid decisions when assessing oral competence in English, one needs to discuss what changes that are necessary to make to alter this impact on the validity of the assessment.

Therefore, I argue that the following aspects should be addressed in the framework in addition to the current features, to expand our understanding of validation:

1. The validity of the guidelines and definitions provided by the curriculum.
2. The indirect impact of teacher education programs on classroom assessment.
3. The impact of the classroom context on teachers' decisions,
4. The influence of teacher insecurity on assessment

These aspects combined are added to the model as step one in the assessment process. I have named this step *the curriculum*, because the curriculum is the baseline that sets the premises for assessment, and it defines what is considered as knowledge in oral competence in English. Additionally, the curriculum is the link between educational policies and teachers' classroom practice, and it hence directly influence the knowledge gap that is addressed by research on the field. Moreover, this step also highlights the importance of including aspects that address

the influence of teachers' interpretations and decisions upon assessment, as factors that also have an impact on what is considered as knowledge in oral competence in English. Furthermore, I suggest five threats that are associated with this step:

• Unclear guidelines for teacher classroom assessment

As described above, unclear guidelines are a threat to validity because it requires individual interpretation of the curriculum. This makes it challenging to ensure validity.

• Poorly defined constructs

The consequences of poorly defined constructs are also described above. It is identified as a threat because it can lead to multiple interpretations of the same construct in addition to teacher insecurity.

• Misinterpretation

If the guidelines provided to teachers give room for interpretation, misinterpretation is a relevant threat because their meaning relies on the teachers' interpretation, which is influenced by their competence and experience. Reducing this threat requires that the teacher is competent to interpret according to its intentions and to take the correct decision. Furthermore, this reflects the importance of teacher education programs, to give teachers the competence and tools they need to interpret the curriculum.

• Teacher insecurity

When guidelines are too general and unspecific it can lead to teacher insecurity because it requires that individual teachers interpret its guidelines, which as mentioned above can lead to multiple interpretations of the same definitions or guidelines. Since interpretation is based on the knowledge that the teacher has, it demands that teachers have both theoretical and practical knowledge to take right decisions based on their interpretation. This also addresses the role of teacher education programs, which foster the mind of novice teachers.

• Variance in teacher assessment literacy

Since teachers' interpretation is underpinned by their current competence in the field. This means that, for example, a teacher with little knowledge of assessment will interpret the curriculum different than for example a teacher with higher assessment competence. This threat addresses the urgency of emphasizing development of teacher assessment literacy in

teacher education programs to enable teachers to interpret guidelines according to their intention.

5.4.1 Expanding the chain of validation

Below, Figure 4 displays the knowledge this study has brought to the field of study. It specifically demonstrates how this research project expands our understanding of the chain of validation by adding *the curriculum to* the model as the first stage of the assessment process:

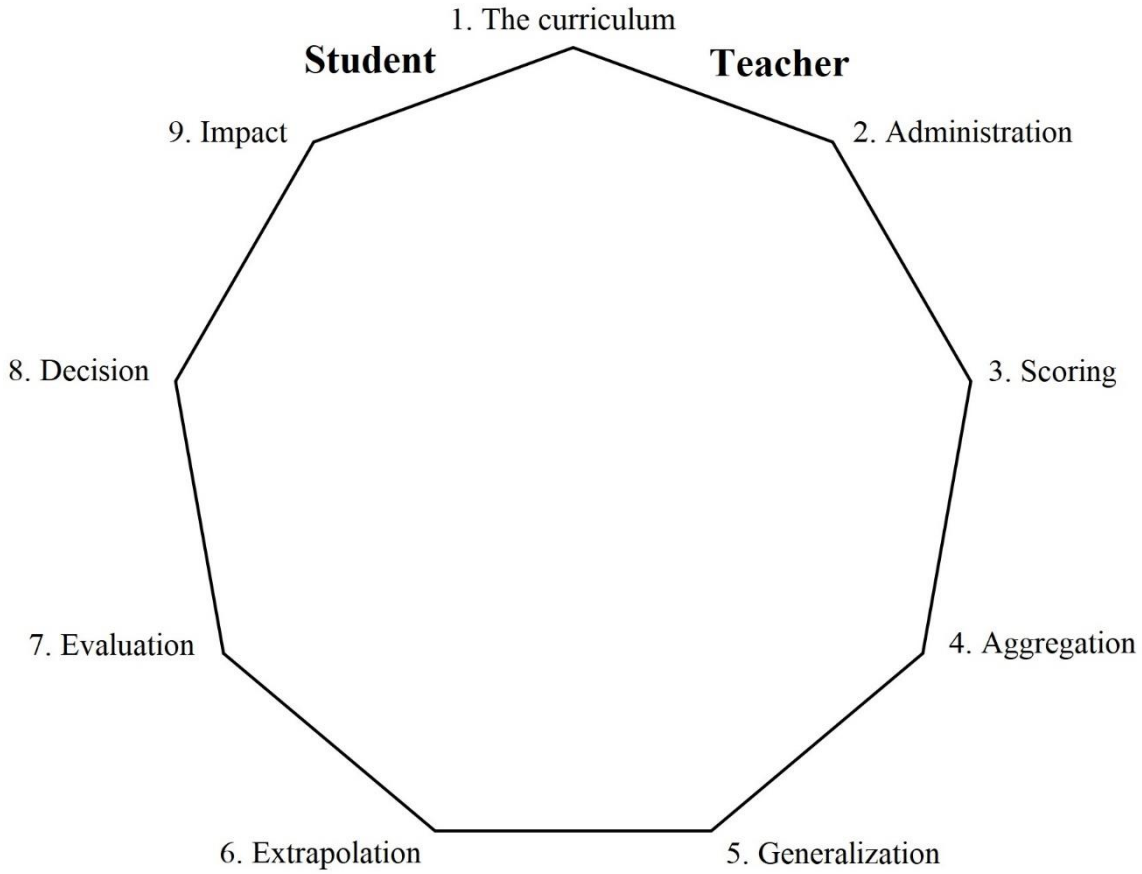

Figure 4: An expanded version of the chain of validation (Crooks, Kane & Cohen, 1996, p. 268)

In addition to adding a new stage to the framework, the role of both agents is highlighted in this expanded version of the chain of validation because of their importance for the assessment process. In the original version, the teacher is not emphasized in the model, but as argued earlier teachers interpret the curriculum, design the assessment, assess and validate the process. The student is directly influenced by all decisions taken by the teacher in the assessment process, in addition to the fact that the teacher's decisions are based on students'

personalities and needs. For example, as the findings display, if the students were demotivated or anxious, the teachers adapted the teaching and assessment in relation to the students' needs. The influence of teachers and students on the assessment process displays the relevance of implementing both as central agents in the chain of validation.

5.5 Possible consequences of reducing threats

The discussion above reveals the connection between the five threats and how they influence each other. In this part, I question the possible consequences of a threat-reduction related to threats caused by educational policies, and what it could mean for teachers' practice of oral assessment in English? For example, will the threats teacher insecurity, contextual dilemmas and variance in teacher assessment literacy decrease if the threats poorly defined constructs and unclear guidelines for teacher classroom assessment were reduced or removed?

If considering the power of the curriculum, as a guiding document that directs the education in Norway, one can argue that it is highly influential upon teachers' practice. I thus argue that if the curriculum provided clearer guidelines and definitions, it would at least reduce teacher insecurity and variance in teacher assessment literacy, because specific guidelines would give less room for personal interpretation. This would also mean that clearer guidelines and definitions would put less responsibility upon the individual teacher. However, a stronger emphasis on assessment competence in teacher education programs could also lead to less insecurity and variance between teachers' competence. This could also lead to a more unified assessment of oral competence in English because it might lead to less misinterpretation. Nonetheless, there would still be some variance in assessment practice, caused by the different personalities of teachers. Furthermore, giving teachers less room for subjectivity could also result in reluctance, and by that decrease motivation amongst teachers, because it would give each individual teacher less power to influence own teaching and assessment.

What is more questionable is if reducing threats related to educational policies would influence the threat contextual dilemmas. On the one hand, if the curriculum provided specific strategies and methods for handling contextual dilemmas, it could facilitate the teacher when meeting such challenges, but contextual dilemmas will always be present because classroom

assessment involves humans. This means that such challenges are influenced by multiple aspects. On the one hand, concrete tools and guidelines could therefore make prioritization when countering dilemmas easier. On the other hand, one can also argue that more concrete guidelines could make assessment more challenging because it could make it harder to adapt the education to the needs of each individual student. Nevertheless, if the guidelines for how to assess oral competence in English would have provided practical tools or specific suggestions for how to assess oral competence, it would facilitate teachers to some extent.

This discussion highlights that both the curriculum and the teacher education programs have a strong impact on teacher insecurity, contextual dilemmas and variance in teacher assessment literacy. Furthermore, it also display the power of the educational system, and the importance of also highlighting the interplay between educational policies and teachers' classroom practice to fully understand how these factors influence each other, and how they influence the validity of oral assessment in English.

5.6 Summary

Through a case study approach, this research project has investigated how teachers validate assessment of oral competence in English. The aim of the study was to answer the following research question: "What threats to assessment of oral competence in English are revealed through teachers' reasoning about own assessment literacy?"

Through diary notes and interviews I have gained insight into the reasoning of two English teachers working at an upper secondary school. Five threats to validity of oral assessment emerged from the analysis of the data material:

1. Teacher insecurity
2. Poorly defined constructs
3. Contextual dilemmas
4. Variance in teacher assessment literacy
5. Unclear guidelines for teacher classroom assessment.

The thesis displays the major consequences of contextual dilemmas and it sheds light on the influence of the curriculum upon teachers' practice and validity. Teacher insecurity was emphasized as a central threat and was also considered a consequence of the other threats. These findings were discussed in light of theories on construct validity and the chain of validation. The study confirmed the threats related to the theoretical framework of construct validity. Furthermore, the discussion illustrated the importance of the curriculum as a guide that influences teachers understanding and practice. Connecting the findings to the chain of validation displayed that the role of the curriculum and teachers' interpretation of the curriculum were missing in the current validation model. Hence, this study advocates for the importance of implementing "the curriculum" as the first stage in the assessment process, because of its major impact on teachers' confidence, understanding of oral competence and decisions taken in the assessment process.

5.7 Implications

Qualitative studies like this one cannot provide a general picture, however, it offers insight into the field of study, which can lead to developing our knowledge of threats related to validity of oral assessment in English. As mentioned in the introduction, this study aims to have a system approach to threats to validity of oral assessment. Therefore, the main purpose of analyzing the teachers' reasoning was not to criticize their practice, but to highlight teachers' role in the assessment process and investigate how the factors that influence teachers' assessment threaten the validity of oral assessment. As a final remark of this study, I direct attention to the following implications:

• The need for more specific guidelines for assessment of oral competence in English. Concrete guidelines would provide a better foundation for teachers to ensure uniformity in assessment, to prevent teacher insecurity and to facilitate teachers when assessing oral competence. More specific guidelines can also help reduce the threats caused by contextual dilemmas if the curriculum provides specific guidelines for what to emphasize in the assessment.

• The necessity for practical tools that can facilitate assessment of oral competence in English, especially related to unplanned assessment. Providing specific tools that help teachers in their assessment can contribute to increasing the validity in all parts of assessment. Furthermore,

practical tools can also facilitate teachers when handling contextual dilemmas, to prevent prioritization that can threaten the validity of the assessment.

• The need to enhance teachers' knowledge of assessment and oral competence in English, to reduce variance in teacher assessment literacy and to reduce teacher insecurity. Developing teachers' assessment literacy can also lead to offering a better education because it can prevent misinterpretations of the curriculum. Therefore, I argue that development of assessment literacy and construct understanding should get more attention in the teacher education programs.

5.8 Further research

As mentioned in the introduction, there are many knowledge gaps in the field of study. This means that it is necessary to investigate multiple aspects to get a better understanding of validity of oral assessment in English. The implications of this research project highlight aspects that should be developed to increase the validity of oral assessment in English. Below I provide some concrete suggestions for further research on the field:

Firstly, a suggestion for further research on the field is to investigate what concrete changes that are necessary to improve the guidelines and definitions in the curriculum, so that they become more applicable for teachers. Such research can expand our understanding of threats to validity, facilitate teachers' practice, and hopefully increase the validity in oral assessment in English.

Secondly, to provide better insight into the field one could conduct a similar investigation, but with a larger number of research participants. This could provide more detailed information of what threats that are relevant. This could also help address what specific changes to current practices or policies that are necessary to provide a better foundation for validity in oral assessment. Furthermore, it would also be interesting to continue this research project by investigating how these threats can be used to develop the educational system.

Thirdly, this research project also identified a connection between the threats. Thus, a suggestion for further research is to specifically investigate the relationship between the

threats and their causality. For example, would the threat teacher insecurity be less relevant if one of the other threats were removed? A quantitative study is suggested as a means to find answers to such questions.

Lastly, a suggestion for further research is also a project that aims to investigate the role of assessment literacy in current teacher education programs. This knowledge could be used to point out specific areas in the field of validity of oral assessment in English that should receive more attention by teacher education programs.

6. Reference list

Aliseda, A. (2006). *Abductive reasoning: logical investigations into discovery and explanation.* Netherlands: Springer Netherlands.

Baxter, P. & Jack, S. (2008). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. *The Qualitative Report,* 13(4), 544-559. Retrieved from: http://nsuworks.nova.edu/tqr/

Bolger, N., Davis, A. & Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annual Review of psychology,* 54, 579-616. Retrieved from: http://www.annualreviews.org/doi/10.1146/annurev.psych.54.101601.145030

Brookhart, S. M. (2005). Developing Measurement Theory for Classroom Assessment Purpose and Uses. *Educational Measurement: Issues and Practice*, 22(4), 5-12. Retrieved from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1745-3992.2003.tb00139.x

Buck, G. (2011). *Assessing listening.* New York: Cambridge University Press.

Council of Europe. (2001). *Common European framework of reference for languages: learning, teaching, assessment.* Cambridge: Cambridge University Press.

Creswell, J. W. (2009). *Research design: qualitative, quantitative, and mixed methods approaches.* Thousand Oaks: SAGE Publications

Crooks, T. J., Kane, M. T. & Cohen, A. S. (1996). Threats to the valid use of assessments, *Assessment in Education: Principles, Policy & Practice*, 3(3), 265-286.

Fjørtoft, H. (2016). Vurdering av muntlighet i klasserommet. In K. Kverndokken (Ed.), *101 måter å fremme muntlige ferdigheter på- om muntlig kompetanse og muntlighetsdidaktikk,* (119- 135). Bergen: Fagbokforlaget.

Gibbs, G. (2007). *Analyzing qualitative data.* London: SAGE.

Gipps, C. V. (2012). *Beyond testing, towards a theory of educational assessment.* London: Routledge.

Gipps, C. V. (2015). *Beyond testing: towards a theory of educational assessment.* New York: Routledge.

Kane, M. T. (2016) Explicating validity, *Assessment in Education: Principles, Policy & Practice*, 23(2), 198-211.

Lovdata. (2009). *Forskrift til opplæringsloven* [Regulations of the Education Act]. Retrieved 06.09.17 from https://lovdata.no/dokument/SF/forskrift/2006-06-23-724/KAPITTEL_4-2#§3-1

Luoma, S. (2004). *Assessing speaking.* Cambridge: Cambridge University Press.

McMillan, J. (2013). *SAGE Handbook on Research on Classroom Assessment.* California: Thousand Oaks: SAGE Publications.

Messick, S. (1988). Validity. In R. L. Linn (Ed.), *Educational measurement*, 3rd. ed. New York: Macmillan.

Messick, S. (1992). Validity in test interpretation and use. In M.C. Alkin (Ed.), *Encyclopedia of Educational Research*, 4(6) (1487-1495). New York: Macmillan Publishing Company.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments, *Educational Researcher*, 23(2), 13-23.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning, *American Psychologist*, 50(9), 741-749.

Morgan, D.L. (1997). *Focus Groups as Qualitative Research.* California: SAGE Publications.

Myles, F. (2010). Research timeline: the development of theories of second language acquisition, *Cambridge Journals* 43(3), 320-332.

Nilssen, V. (2012). *Analyse I kvalitative studier: Den skrivende forskeren.* Oslo: Universitetsforlaget.

Norwegian Directorate for Education and Training. (2013a). *English subject curriculum-Main subject areas*. Retrieved 10.29.17 from https://www.udir.no/kl06/ENG1-03/Hele/Hovedomraader?lplang=http://data.udir.no/kl06/eng

Norwegian Directorate for Education and Training. (2013b). *English subject curriculum: Basic skills.* Retrieved 03.03.18 from https://www.udir.no/kl06/ENG1-03/Hele/Grunnleggende_ferdigheter?lplang=http://data.udir.no/kl06/eng

Norwegian Directorate for Education and Training. (2006). *English subject curriculum (ENG1-03).* Retrieved 01.21.18 from https://www.udir.no/kl06/ENG1-03?lplang=eng

NTNU. (n.d). *Skolebasert kompetanseutvikling i vurdering-SKUV.* Retrieved 04.05.18 from https://www.ntnu.no/ilu/skuv

Patton, M. Q. (2015). *Qualitative Research & Evaluation Methods.* California: SAGE Publications.

Polanyi, M. (1966). The logic of tacit inference. *Philosophy,* 41(155), p. 1-18.

Postholm, M. B. (2010). *Kvalitativ metode: En innføring med fokus på fenomenologi, etnografi og kasusstudier.* Oslo: Universitetsforlaget.

Prøitz, T. S. & Borgen, J. S. (2010). *Rettferdig standpunktvurdering- det (u)muliges kunst?: Læreres setting av standpunktkarakter i fem fag i grunnopplæringen.* Oslo: NIFU STEP

Ringdal, K. (2014). *Enhet og mangfold: Samfunnsvitenskapelig forskning og kvantitativ metode.* Bergen: Fagbokforlaget.

Sá, J. (2002). Diary writing: An interpretative research method of teaching and learning. *Educational Research and Evaluation,* 8(2), 149-168. Retrieved from: http://www.tandfonline.com/doi/abs/10.1076/edre.8.2.149.3858

Saldaña, J. (2016). *The coding manual for qualitative researchers.* London: SAGE Publications.

Stobart, G. (2009). Determining validity in national curriculum assessments, *Educational Research*, 52(2), 161-179.

Stiggins, R. J., & Conclin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment.* Albany: SUNY Press.

Svenkerud, S., Klette, K. & Hertzberg, F. (2012). Opplæring i muntlige ferdigheter. *Nordic Studies in Education*, 32 (1), 35-49. Retrieved from: https://www.idunn.no/file/pdf/53214176/opplaering_i_muntlige_ferdigheter.pdf

Yin, R. K. (2003). *Case study research: Design and methods*. Thousand Oaks: SAGE Publications.

7. Appendices

Appendix A: Approval from the Norwegian Centre for Research Data

Henning Fjørtoft

7491 TRONDHEIM

Vår dato: 24.11.2017 Vår ref: 56798 / 3 / OOS Deres dato: Deres ref:

Vurdering fra NSD Personvernombudet for forskning § 31

Personvernombudet for forskning viser til meldeskjema mottatt 25.10.2017 for prosjektet:

56798	*Læreres arbeid med vurdering av muntlig språkkompetanse i engelsk*
Behandlingsansvarlig	*NTNU, ved institusjonens øverste leder*
Daglig ansvarlig	*Henning Fjørtoft*
Student	*Kine Brattland*

Vurdering

Etter gjennomgang av opplysningene i meldeskjemaet og øvrig dokumentasjon finner vi at prosjektet er meldepliktig og at personopplysningene som blir samlet inn i dette prosjektet er regulert av personopplysningsloven § 31. På den neste siden er vår vurdering av prosjektopplegget slik det er meldt til oss. Du kan nå gå i gang med å behandle personopplysninger.

Vilkår for vår anbefaling

Vår anbefaling forutsetter at du gjennomfører prosjektet i tråd med:

- opplysningene gitt i meldeskjemaet og øvrig dokumentasjon
- vår prosjektvurdering, se side 2
- eventuell korrespondanse med oss

Vi forutsetter at du ikke innhenter sensitive personopplysninger.

Meld fra hvis du gjør vesentlige endringer i prosjektet

Dersom prosjektet endrer seg, kan det være nødvendig å sende inn endringsmelding. På våre nettsider finner du svar på hvilke endringer du må melde, samt endringsskjema.

Opplysninger om prosjektet blir lagt ut på våre nettsider og i Meldingsarkivet

Vi har lagt ut opplysninger om prosjektet på nettsidene våre. Alle våre institusjoner har også tilgang til egne prosjekter i Meldingsarkivet.

Vi tar kontakt om status for behandling av personopplysninger ved prosjektslutt

Dokumentet er elektronisk produsert og godkjent ved NSDs rutiner for elektronisk godkjenning.

NSD – Norsk senter for forskningsdata AS Harald Hårfagres gate 29 Tel: +47-55 58 21 17 nsd@nsd.no Org.nr. 985 321 884
NSD – Norwegian Centre for Research Data NO-5007 Bergen, NORWAY Faks: +47-55 58 96 50 www.nsd.no

Ved prosjektslutt 25.05.2018 vil vi ta kontakt for å avklare status for behandlingen av personopplysninger.

Se våre nettsider eller ta kontakt dersom du har spørsmål. Vi ønsker lykke til med prosjektet!

Marianne Høgetveit Myhren

Øyvind Straume

Kontaktperson: Øyvind Straume tlf: 55 58 21 88 / Oyvind.Straume@nsd.no

Vedlegg: Prosjektvurdering
Kopi: Kine Brattland, kine.brattland@gmail.com

Personvernombudet for forskning

Prosjektvurdering - Kommentar

Prosjektnr: 56798

REKRUTTERING OG SAMTYKKE

Formålet med prosjektet er å få innsikt i lærere sin erfaring, og studien innebærer intervju med lærerne. Disse informeres skriftlig og muntlig om prosjektet og samtykker til deltakelse. Informasjonsskrivet er godt utformet, men personvernombudet gjør oppmerksom på en skrivefeil i prosjektslutt (2017, korrekt år er 2018).

OBSERVASJON

I følge prosjektmeldingen skal det gjennomføres observasjon på skolen for å få et inntrykk av lærerne som deltar. Mens skole er en obligatorisk arena for barn skal deltagelse i forskning være frivillig. Observasjoner der det ikke registreres hverken direkte eller indirekte personidentifiserbare opplysninger omfattes ikke av meldeplikten. Av etiske hensyn ber vi om at likevel det gis informasjon til elevene, slik at disse er klare over hva som skjer. Vi gjør oppmerksom på at forsker også på forhånd må avklare gjennomføring av prosjektet med ledelsen i skolen.

DATASIKKERHET

Personvernombudet legger til grunn at forsker etterfølger NTNU sine interne rutiner for datasikkerhet. Dersom personopplysninger skal sendes elektronisk eller lagres på mobile enheter, bør opplysningene krypteres tilstrekkelig.

PROSJEKTSLUTT

Forventet prosjektslutt er 25.05.2018. Ifølge prosjektmeldingen skal innsamlede opplysninger da anonymiseres. Anonymisering innebærer å bearbeide datamaterialet slik at ingen enkeltpersoner kan gjenkjennes. Personvernombudet gjør oppmerksom på at det gjøres ved å:

- slette direkte personopplysninger (som e-post)
- slette/omskrive indirekte personopplysninger (identifiserende sammenstilling av bakgrunnsopplysninger som f.eks. arbeidssted, yrke og kjønn)
- slette digitale lydopptak

Appendix B: Information letter

Forespørsel om deltakelse i forskningsprosjektet *«Læreres arbeid med vurdering av muntlig språkkompetanse i engelsk»*

Bakgrunn og formål
Jeg sender dere denne formelle forespørselen om å delta i mitt masterprosjekt på bakgrunn av interessen dere allerede har vist.

Som en del av min masterstudie ved NTNU skal jeg skrive en masteroppgave om hvordan lærere jobber med vurdering av muntlighet i engelskfaget. Jeg er interessert i å få innsikt i hvilke erfaringer dere har med slikt vurderingsarbeid og hvordan dere konkret jobber med vurdering av muntlig språkkompetanse i engelsk.

Hva innebærer deltakelse i studien?
Jeg vil samle inn data i form av intervju, dokumenter, observasjon i klasserommet og dagbokdata. Data fra intervju registreres i form av lydopptak og vil omhandle spørsmål knyttet til vurdering av muntlighet i engelskfaget. I observasjonene vil jeg fokusere på vurderingssituasjoner som foregår i klasserommet, relatert til muntlig språkkompetanse. Jeg vil også be dere skrive dagboknotater en gang hver uke, der dere forteller om undervisningen. Dette kan være en fin mulighet til å reflektere over egen praksis. Jeg håper derfor at dere ser dette som en mulighet til å lære mer om egen- og kollegers vurderingspraksis.

Hva skjer med informasjonen om deg?
Alle personopplysninger vil bli behandlet konfidensielt, og vil kun være tilgjengelig for meg og veileder. Lydopptakene transkriberes, og vil deretter slettes. Deltagerne anonymiseres og vil ikke kunne gjenkjennes i publikasjonen av oppgaven. Prosjektet skal etter planen avsluttes 25.05.18.

Frivillig deltakelse
Det er frivillig å delta i studien, og du kan når som helst trekke ditt samtykke uten å oppgi noen grunn. Dersom du trekker deg, vil alle opplysninger om deg bli anonymisert og slettet. Studien er meldt til Personvernombudet for forskning, NSD - Norsk senter for forskningsdata AS.

Dersom du har spørsmål, ta kontakt med:

Kine Brattland
kine.brattland@gmail.com,
Tlf.: 95899508

Veileder:
Henning Fjørtoft
Tlf: 48044018
henning.fjortoft@ntnu.no

Samtykke til deltakelse i studien

Jeg har mottatt informasjon om studien, og er villig til å delta:

(Signert av prosjektdeltaker, dato)

Appendix C: Interview guide, group interview 1

Intervjuguide: gruppeintervju 1

1.Hva er vurdering?

2.Hva er vurdering av muntlighet?

3.Hvordan kan man få innsikt i elevenes muntlige språkkompetanse?

4.Hvordan jobber dere med vurdering av muntlig kompetanse?

5.Hva bruker dere å legge vekt på i slik vurdering?

6.Hvilken innsikt i elevens kompetanse opplever dere at slik vurderingen gir?

7.Hva legger dere vekt på i tilbakemeldinger dere gir til elevene?

8.Når erfarer dere at tilbakemeldingene dere gir har mest innvirkning på elevenes læring?

9. Hvilke tilbakemeldinger erfarer dere at elevene deres forstår best?

English translation:

Interview guide: group interview 1

1. What is assessment?
2. What is oral assessment?
3. How can you get insight into students' oral language competence?
4. How do you work when assessing oral competence?
5. What do you emphasize in such assessment?
6. What insight into students' competence do you experience that this kind of assessment gives?
7. What aspects do you emphasize in the feedback you provide to students?
8. When do you experience that the feedback you give to students is the most influential on their learning?
9. What kind of feedback do you experience that students understand the best?

Appendix D: Interview guide, individual interview

Intervjuguide: individuelt intervju

1.Hvordan bruker du å jobbe når du skal planlegge en muntlig vurderingssituasjon?

2. Hvordan sikrer du at oppgavene du gir elevene gir innsikt i den kompetansen du ønsker at de skal belyse?

3.Hvilke tilbakemeldinger erfarer du at elever har mest nytte av?

4.Hvis du tenker på tidligere erfaringer med muntlig vurdering av engelsk, hvilke erfaringer har du tatt med deg videre i ditt vurderingsarbeid? Hvordan bruker du disse erfaringene?

5.Hva mener du avgjør om vurderingen er vellykket?

English translation:

Interview guide, individual interview

1. How do you usually work when planning oral assessment?

2. How do you ensure that tasks give insight into the intended competence of the student?

3.What feedback do you experience as the most useful for students?

4. If you think about previous experiences with assessment of oral competence in English, what experiences have shaped your assessment practice? How do you use these experiences?

5. In your opinion, what determine if an assessment is successful?

Appendix E: Interview guide, group interview 2

Intervjuguide: gruppeintervju 2

1. Hva forstår dere med «vurdering av muntlige ferdigheter»?

2. Hvordan jobber dere for å utvikle elevers muntlige ferdigheter?

3. Hvordan tenker dere når dere lager vurderingskriterier?

4. Det er vanlig å skille mellom *spontan muntlighet* (samtaler uten forberedelse) og *planlagt muntlighet* (forberedte presentasjoner og lignende) i skolen. Hvilke tilbakemeldinger bruker dere å gi til elever i spontane samtaler?

5. Hvordan bruker dere informasjonen dere får fra muntlig vurdering?

6. Hvilke utfordringer møter dere når dere tolker elevens kompetanse i muntlige ferdigheter?

7. Hvordan sikrer dere at deres tolkning av elevens kompetanse stemmer med elevens nivå?

8. Kan vurdering av muntlige ferdigheter gjøres enklere?

9. Hvordan har deltagelse i dette prosjektet påvirket deres forhold til muntlig vurdering?

English translation:

Interview guide: second group interview

1.What is your understanding of assessment of oral competence?

2.How do you work to develop students' oral competence?

3.How do you think when making assessment criteria?

4.It is normal to distinguish between oral activities (spontaneous conversations) and planned oral activities (presentations, tests) in education. What feedback do you usually give students in spontaneous conversations?

5.How do you use the information you obtain from oral assessment?

6.What challenges do you encounter when interpreting students' oral competence?

7.How do you ensure that your interpretation of the students' competence matches/represents the students' language level?

8. Can we conduct/ carry out assessment of oral competence in an easier way?

9. How has participation in this project influenced your relationship to assessment of oral competence?

Appendix F: Example of a diary note

Week 3

In many ways I thought this week went better than the previous ones when I think about spoken skills. The pupils in all of my classes seem to be getting more comfortable with the fact that English is spoken in the classroom. Some of them still prefer to speak Norwegian, but the majority are making a real effort.

My 3rd year students worked well on an analysis of a Yeat's poem and finished up the lesson with presenting their findings infront of the class in groups of three. I noticed that their speech had more free flow to it and they weren't so caught up in their notes. This was perhaps because they hadn't been given time to write so many notes. Maybe this is an idea that I could use for other classes – i.e. that students are given a certain short time to produce something (eg.15 minutes) and then asked to present it infront of the others?

It might prove difficult for the vocational students, however, since they are still struggling with presenting things infront of their classmates. It's important to remember that my vocational students are young and only in their first year of upper secondary education. I strongly feel that time and thus maturity are two important ingredients that students need to help them with their confidence in a foreign language. Making mistakes in pronunciation, grammar, vocabularty etc…are all things that need to be worked upon and DO often improve with time and accompanying confidence.