English VG1 level oral examinations

How are they designed, conducted and assessed?

Lill Mari Yildiz

Mastergradsavhandling ved Institutt for Lærerutdanning
og Skoleutvikling, Engelsk fagdidaktikk

UNIVERSITETET I OSLO

Vår 2011

English VG1 level oral examinations

How are they designed, conducted and assessed?

Lill Mari Yildiz

Mastergradsavhandling ved Institutt for Lærerutdanning og Skoleutvikling, Engelsk fagdidaktikk

UNIVERSITETET I OSLO

2011

English VG1 level oral examinations – How are they designed, conducted and assessed?

Lill Mari Yildiz

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo

Abstract

The aim of this thesis has been to find out more about oral examinations in English at the VG1 level in Norwegian upper secondary school. More specifically I wanted to find out more about how the examinations are designed, conducted and assessed. In addition I wanted to find out more about the teachers' attitudes towards the oral examination, whether they think the oral examination is fair, and whether they would prefer a locally given or a centrally given oral examination. I also wanted, if possible, to try to find out more about the reliability and construct validity of the oral examinations.

The background for interest in this was first of all the fact that the Directorate of Education has left all responsibility for the oral examinations to the county administrations, while they are in full control of the written examination. Because of this we know very little about the practices of the oral examinations. However, both national and international studies indicate that there are great variations in the practice of oral examinations, especially with regards to assessment.

In this project I have used a qualitative research method and a phenomenological research design. I have conducted interviews and used open survey questions, and have collected information from 16 teachers, in 16 different schools, in 16 different counties. First I looked into theory concerning both assessment in general, and assessment of speaking in particular. I have especially been interested in theory concerning reliability and construct validity. I have also looked closely at the construct for the English VG1 oral examination, to see if the examination practices across the country actually test the construct at hand.

The findings of my project indicate that there is great variation in oral examinations with regards to both the format of the examinations and the assessment process. This involves what is found to be important elements to be assessed in an oral examination, and with regard to the use of assessment criteria. I argue that these findings may indicate issues with regards to reliability and construct validity.

However, this is a small-scale survey, and the findings can therefore not automatically be generalized to English oral examinations. It would therefore be very interesting to conduct a larger-scale qualitative or quantitative survey to find out more about the oral examinationsin English as well as in other subjects.

Sammendrag

Temaet for denne oppgaven har vært å finne ut mer om muntlig eksamen i engelsk på VG1 i den norske videregående skolen. Mer spesifikt ville jeg finne ut mer om hvordan eksamen er utformet, praktisert og vurdert. I tillegg ønsket jeg å finne ut mer om lærernes oppfatning av eksamen, om de synes eksamen er rettferdig, og om de foretrekker en lokalt eller sentralt gitt muntlig eksamen. Hvis mulig, ønsket jeg å finne mer informasjon om eksamen er reliabel og valid.

Bakgrunnen for min interesse for dette, var først og fremst at Utdanningsdirektoratet har gitt ansvaret for muntlig eksamen til fylkeskommunene, mens de selv har full kontroll over skriftlig eksamen. Dette har ført til at vi vet veldig lite om praksis rundt muntlig eksamen. Samtidig, finnes det både nasjonale og internasjonale studier som tyder på at det er store variasjoner i praksis av muntlig eksamen, spesielt i forhold til vurdering.

I dette prosjektet har jeg benyttet kvalitativ forskningsmetode og fenomenologisk forskningsdesign. Jeg har brukt intervjuer og åpne skriftlige spørsmål for å samle informasjon fra 16 lærere, 16 ulike skoler i 16 fylker. Men, først har jeg undersøkt teori som omhandler både vurdering generelt og vurdering av muntlig produksjon spesielt. Jeg har vært spesielt interessert i teori om reliabilitet og konstrukt validitet. Jeg har også sett nærmere på konstruktet for engelsk muntlig eksamen, for å finne ut om eksamenspraksisen faktisk tester konstruktet.

Funnene mine antyder at det er store variasjoner i muntlig eksamen med tanke på både eksamensformatet og vurderingsprosessen. Dette gjelder hvilke elementer som anses som viktig å vurdere på en muntlig eksamen, og om og i hvor stor grad det brukes vurderingskriterier. Basert på dette, har jeg argumentert for at dette tyder på at det kan være problemer knyttet til reliabiliteten og validiteten av muntlig eksamen.

Det er samtidig viktig å huske at dette er en liten undersøkelse med relativt få informanter, og at funnene ikke uten videre kan overføres til muntlig eksamen in Norge på generell basis. Likevel, tyder funnene på at det er visse mønster og tendenser, og dette styrkes med tanke på at informantene mine er fra forskjellige skoler i 16 fylker. Det ville være veldig interessant å utføre en store kvalitativ eller kvantitativ undersøkelse for å undersøke muntlig

eksamen nærmere. Dette kan hjelpe oss å finne svar på hvordan vi skal sikre at elevene får den individuelle og samme vurderingen som de har rett til.

Acknowledgements

First of all, I would like to express my thanks to all of the teachers who have been my informants in this project. Thank you for agreeing to be part of the project, and thank you for taking the time to answer all my questions. The information you have provided me with in invaluable. If it hadn't been for your help, there would not have been a finished thesis at this point.

I would also like to express my deepest gratitude to my thesis supervisor, Associate Professor at the University of Oslo, Glenn Ole Hellekjær. Thank you for pushing me to keep going, for your patience, support, guidance and all the helpful feedback you have given me throughout the project.

In addition I would like to thank lecturer Henrik Bøhn for sharing ideas and theoretical insight.

I would also like to thank my sister, Karine, for technical help, ideas on how to make the thesis better and for all her support.

Last, but not least I want to thank friends and family for having faith in me and never giving up on me. Special thanks to Mamma and Erik, and to my dear husband Kamil. Thank you for always taking the time to listen when I need someone to talk to. Thank you for encouraging me at times when I have wanted to give up, and thank you for always being there for me.

Oslo, May 2011

Lill Mari Yildiz

Table of contents

List of tables

1 Introduction

1.1 Background

According to the English subject curriculum in the Knowledge promotion reform, LK06, written examinations are to be prepared and graded centrally, while oral examinations are to be prepared and graded locally. This has led to extensive effort by the Norwegian Directorate for Education and Training to make sure we have good written examinations. They have also developed guidelines to help the external examiners correct the examinations, as well as courses for the examiners to ensure that they are well enough trained to correct exam papers in a consistent way. In addition it is also ensured that no exam paper is corrected by only one examiner, but at least two. All this is done to ensure reliable scoring of written examinations, reliability meaning that the scoring is consistent. It is very important to ensure reliable examinations, both written and oral, because examinations are what can be called high stakes for the students (Broadfoot 2007). This means that the results of the examination may have consequences for the students. For example it might have influence on what higher education the students are accepted to and what career they can choose later in life.

A key difference between the oral and written examination concerns who is responsible for the examination. The Directorate of Education and Training has left the responsibility for the oral examination to the county administrations, which is the reason why I became interested in this topic. I find it interesting how they do so much work to ensure that the written examinations are both valid and reliable, while they do little to ensure this for the oral examination.

Because the Directorate of Education and Training are responsible for the written examinations we know much about the exam papers as well as the assessment of exam papers, both the examination tasks and the guidelines are published online each year. At the same time, because the oral examination is to be prepared and graded locally, we know very little about the practices across the country when it comes to oral examinations.

The county administrations, that is responsible for the upper secondary education, receive few guidelines for the oral examination. The only thing that is stated is that the oral examination should be given and graded locally and that it should, as the written examination, «be based on the entire subject (140 teaching hours)» (Utdanningsdirektoratet 2006/2010). Together with this the Regulations to the Act of Education controls the oral examination. The Regulations § 3.30 specifies that the county administration is responsible for all locally givens examinations, while the local teachers must design suggestions for examination tasks. Further, it states that the students are to be made aware of what subject they are going to have their oral examination in 48 hours before the examination day. When it comes to structure the Regulations only say that the examination can have a preparation period before the examination up to 48 hours. There is not much in the Regulations about assessment, but they require that there have to be two examiners and that one of them should be external. If the examiners disagree on the result the external examiner will have the last word. There is nothing about assessment criteria in the regulations (Forskrift til Opplæringslova). As we see this is not much to go on, and because of this I find it interesting to try to find out more about how the oral examinations are developed and graded. I suspect that there are differences between counties, as well as between schools.

As mentioned above, it is very important that assessment of high-stakes examinations is fair and reliable. However, there are studies that show that assessment practices in Norway vary and may not be in consistence with the Act of Education. Bøhn (2011) says that an investigation by the County Governors in 2010 looking at examination practices in upper secondary education in Norway, shows that there is too much variation when it comes to examination formats. Furthermore, there are other studies by Galloway et al. (2011), Hægeland et al. (2005) and Throndsen et al. (2009), that show that many teachers assess performance in examinations in a norm-referenced manner, even if this is not consistent with the Regulations to the Act of Education. These specify that assessment in Norway should be criterion-referenced, which means that students should be assessed according to the competence aims of the subject curriculum, and not compared to each other, which would be a norm-referenced manner of assessment. One of the reasons why it is like this might be that until the new Regulations to the Act of Education came in 2001, the tradition in Norway has been dominated by a norm-referenced assessment approach (Stokke et al. 2008).

Furthermore, national research shows that it might be difficult for teachers to operationalize the curriculum goals, and because of this teachers develop and use different assessment criteria (Prøitz and Borgen 2010; Throndsen et al. 2009). To operationalize the curriculum goals means to break them down to less complex aims for learning, which the students can use to evaluate their learning progress. Teachers also report that they find it difficult to describe competence at different levels (Throndsen et al. 2009), which is necessary when one is going to assess in an examination.

A number of international studies support the findings of the Norwegian research. For example studies show that raters often give different scores to performances at the same levels, and also the same scores to performances at different levels (Lumley & McNamara 1995; Orr 2002; Ang-Aw & Goh 2011). This shows that there may be issues concerning inter-rater reliability, which in short means that the examiners do not agree on what is required of a performance to qualify for a certain level. Further, Lumley (1998) and McNamara (1996) report that teachers often are willing to award marks for effort at the same time as they less likely to award extreme marks. This again produces a narrow range of marks (Ang-Aw & Goh 2011), which means that the scores gather at middle of the rating scale. Research also shows that raters use criteria inconsistent, and that raters focus different on criteria (Wigglesworth 1994). That raters use criteria inconsistent may be a sign of low intra-rater reliability, which means that an examiner has to agree with himself on which performances qualify for which grade. That different raters focus different on criteria, may be a sign of low inter-rater reliability, which I described above.

Studies by Brown (2000), Douglas (1994), Pollitt and Murray (1996) and Orr (2002) also report findings that show that raters may base their decisions on a wide range of non-criterion information (Ang-Aw & Goh 2011). They also suggest that this shows a deficiency in the raters' understanding of the construct, meaning the description of what should be tested, the rating scales are build on (Brown 2000; Orr 2002; Wigglesworth 1994). Furthermore, it is reported that raters have different understanding of and interpret differently the construct to be measured (Ang-Aw & Goh 2011). Moreover, it has also been argued that these differences in raters would still exist after rater training (Douglas 1994; Lumley 1998).

Another important finding is that research shows that when assessing speaking, raters can also differ in approach. Pollitt and Murray (1996) reported two approaches used:

- A *synthetic* process when raters first form a holistic image of the candidate's performance, from first impression, based on their understanding of language learners, and then compared it with subsequent observed performances
- A more objective, less natural process where there raters evaluate intuitively each

utterance, and then add these up to the final score.

(Ang-Aw & Goh 2011).

Brown (2000) added a third approach that combines the two above:

- Raters found elements supporting their holistic image of the candidate, and at the same time rated sections of the test independently before weighing them.

(Ang-Aw & Goh 2011).

Moreover, there is also international research on the assessment of written performance that supports the above presented findings. Eckes (2008) mentions five ways in which raters differ:

- in the extent to which they follow a scoring rubric
- in the way in which they interpret assessment criteria
- in their understanding and use of rating scale categories
- in the degree to which they score severely or leniently
- in the extent to which their ratings are consistent across examinees, scoring criteria and performance tasks

(Brown 2000, p. 156).

To conclude, research has shown that there to a large extent is inconsistency in the way raters assess both written and oral performances. It has been shown that there are variances in several aspects of assessment. It seems that there are variances in the way raters assess oral performance, but also in what assessment criteria they use and the approach they use when assessing. It is therefore necessary to do more research on this, both nationally and internationally. The research presented above shows the need for more research and studies in this field, and is therefore both an interesting and necessary topic for a Master thesis.

1.2 Aim and research statement

Based on the research presented above and interest in the topic, the aim for this study is to look into English oral examinations at VG1 level in Norway. I want to find out more about what different formats of the oral examination that exist, who develops them, and if there is a system for how the oral examinations are assessed. I will also focus on the differences in rater practices. The oral examinations are high stakes for the students just as the written examination is, and because of this it is necessary to ensure that the assessment of oral examinations is as made as reliable as possible. If there is a lack of guidelines and a common examination format, this together with variances in assessment practice, might well cause reliability problems in oral examinations.

With this as a background the research statement can be outlined as follows: *How are English oral examinations at VG1 level in Norway designed, carried out and assessed across the country?* This statement focuses on both the process of making the oral examinations and the grading of the exams. It can be broken down into more than one research question:

1) How are oral examinations designed, how is the format?
2) How do teachers assess the oral examinations with regards to assessment criteria and rating scales?
3) Are the examinations and the results of the examinations valid and reliable?

I will try to give answers to these questions, based on the findings of my research in the results and analysis chapter, and present methods and approaches used in the method chapter. In the following two chapters I will try to give a theoretical overview of assessment.

2 Theoretical overview – Assessment in general

2.1 Introduction

Before going through the results of my findings and discussing them, it is necessary to take a look at theory concerning my topic. In this chapter I will start by looking at what assessment is in general and then oral assessment specifically. In the next chapter I will present test operationalization and look at important concept as reliability and validity.

2.2 A short history of language testing

In this section I will give a very short introduction to the history of language testing and the development we have seen in assessment of language. According to Brown (1996), language testing can be split up in historical periods or movements. Hinofotis (quoted in Brown 1996) called those the prescientific movement, the psychometric-structuralist movement and the integrative-sociolinguistic movement (Hinofotis as cited by Brown, 1996, p. 23). Brown also added the communicative movement, which he feels is the latest trend in assessment movement. In the following I will give a short presentation of each of the movements.

2.2.1 The prescientific movement

The prescientific movement is associated with grammar and translation approaches to language teaching and testing. Therefore the tests are commonly based on translation and free composition, and it is the classroom teacher that make the tests on his or her own. Brown says these tests are rather difficult to score objectively, and to a large extent are dependent on a subjective scoring. During this movement's period one did not take into consideration factors such as objectivity, validity and reliability. Today most teachers would not accept practices that are not concerned with making fair and consistent tests. However, Brown says that a movement such as the prescientific movement never ends in language testing, because even today these practices are used in parts of the world and school systems. (Brown, 1996, pp. 23-24)

2.2.2 The psychometric-structuralist movement

When the psychometric-structuralist movement came, it introduced concerns about objectivity, reliability and validity, and work to make tests more reliable and precise than earlier. The tests from this movement, according to Brown, measured “the discrete structure points (Carroll 1972) being taught in audio-lingual and related teaching methods of the time” (Brown, 1996, p. 24), and were influenced by behavioral psychology. The tests made were often standardized, multiple choice tests that were easier to administer and score. This type of tests are still used to a large extent across the world, but are often supplemented by integrative tests (Brown, 1996, p. 24).

2.2.3 The integrative-sociolinguistic movement

This movement came as a reaction to the previous because many started to believe that language tests needed to test more than the discrete parts of language. Language professionals argued that communicative competence was more than using correct grammar, and also entailed using appropriate language in different situations. The tests that were now given were cloze tests and dictations, and one was still concerned with the aspect objectivity, reliability and validity. This is maybe what makes this period so important, because they saw that language tests needed to test more than they had done earlier, but they still used the techniques provided by the previous movements in their test making. (Brown, 1996, pp. 24-25).

2.2.4 The communicative movement

In addition to the movements presented above, Brown (1996) added the communicative movement, which he thought was a movement we had seen the beginning of at the time writing. Brown said that this new movement might lead to the tests in the future focusing on more authentic language situations where the students are to show ability to communicate in a purposeful way. I am not sure whether we have seen tests as this yet, but that we might be moving in that direction. Today we are trying to give the students tests that test more aspects of language at the same time, and preferably in more authentic ways than earlier.

2.2.5 Summing up

In the previous paragraphs I have given a very short introduction to the history of language testing and the movements and trends we have seen. I have used categories provided by Brown (1996). In the following I will look at important distinctions when it comes to terms and concepts in assessment and testing.

2.3 Basic distinctions in assessment

In this section I will look at the basic distinctions and concepts in assessment. I will start by looking at the differences between formative and summative assessment.

2.3.1 Formative and summative assessment

According to the Ministry of Education and Research, the Norwegian assessment system is based on individual assessment, of which we use the two basic types, formative and summative assessment[1]. Formative assessment is the kind of assessment that is done during the teaching- and learning process, while summative assessment is the testing that is to test the students' competence at the end of the teaching- and learning process, such as written and oral examinations. Formative assessment is also supposed to provide guidance to the students during their training by showing their strengths and weaknesses in the subject and making it possible to improve their competences. It is also supposed to show the students what they have to work on and what they already know according to the competence aims from the subject curriculum. Formative assessment can be done with or without grades, and it can be given in writing or as oral given feedback. Methods that are commonly used for formative assessment are; followup conversations with students, self-assessment and portfolio assessment.

Summative assessment is assessment given at the end of a learning- and teaching process. There are two types of summative assessment; the final grade and the final examination. Summative assessment is also connected to the competence aims from the subject curriculum and is supposed to show the level of competence the students have at a

[1] NOU 18 Vurdering og dokumentasjon av læringsutbytte. Retrieved from: http://www.regjeringen.no/nb/dep/kd/dok/nouer/2003/nou-2003-16/19.html?id=370792

given time. The English oral examination at VG1 level in Norway is clearly a form of summative assessment.

In this section I have in short described the differences between formative and summative assessment and I will in the following give a brief description of the terms formal and informal assessment. These terms are closely related to the terms described above.

2.3.2 Formal and informal assessment

Simensen distinguishes between informal and formal assessment (Simensen 1998, p. 251). Informal assessment is the assessment the teacher does on a daily basis, such as dialogues with students, observations of students and question-answer sequences. This is connected to formative assessment, but it is not exactly the same, because formative tests that I described above are not a part of informal assessment. Consequently, informal assessment is a part of formative assessment, but formative assessment is more than just informal assessment. Formal assessment on the other hand, is the use of examinations and other tests. This means that all tests given during the school year are part of the formal assessment, and that formal assessment includes both formative and summative assessment. When we look at the Norwegian VG1 level English oral examination it is clear that this is a formal test that is also summative, since it is the final examination of the course. In this section I have looked briefly at the differences between informal and formal assessment and in the following I will present more in detail the distinction between norm-referenced and criterion-referenced assessment.

2.3.3 Norm-referenced and criterion-referenced assessment

Another important distinction is the one made between norm-referenced assessment and criterion-referenced assessment. Brown describes the terms and says that a norm-referenced test is designed to measure global language abilities, for example language proficiency, reading comprehension and so on. The score result of one student is interpreted relative to the score results of all the others who took the test. The results of the test is then spread out in a normal distribution curve where the students with low abilities a placed in one end, and those with high abilities are paced at the other end, and with most students ending up in the middle. On the other hand, a criterion-referenced test measures well-defined and specific objectives that are specific to a course or subject. When interpreting the scores, different students' scores are not compared to each others, which means that a student's score is meaningful on its own because it shows how much the student has learned of the objectives that are tested. The results of such a test does not have to be normal distributed, because if all students know 100% of the material they should all get full score (Brown, 1996, p. 4). Furthermore, Brown says that the norm-referenced and the criterion-referenced assessment contrast in six ways; 1) the ways that the scores are interpreted, 2) the kinds of things that they are used to measure, 3) the purpose of testing, 4) the ways that the scores are distributed, 5) the structure of the test and 6) the students' knowledge of test question content. Brown sums this up in the following table:

Table 1: Differences Between Norm-Referenced and Criterion-Referenced Tests

Characteristics	Norm-referenced	Criterion-referenced
Type of interpretation	Relative (A student's performancies compared to that of all other students in the percentile terms)	Absolute (A student's performance is compared only to the amount, or percentage, of material learned)
Type of measurement	To measure general language abilities or proficiencies	To measure specific objectives-based language points
Purpose of testing	Spread students out along a continuum of general abilities or proficiencies	Assess the amount of material known, or learned, by each student
Distribution of scores	Normal distribution of scores around a mean	Varies, usually non-normal (students who know all of the material should all score 100%)
Test structure	A few relatively long subtests with a variety of question contents	A series of short, well-defined subtests with similar question contents
Knowledge of questions	Students have little or no idea of what content to expect in questions	Students know exactly what content to expect in test questions

Copied from Brown, 1996 p. 5

This table shows the main differences between the two types of tests. Norm-referenced and criterion-referenced tests are used for different decision purposes. Brown used four types of decision purposes; proficiency, placement, achievement and diagnostics. According to Brown these types are also described by Alderson, Krahnke and Stansfield, (1987), as the four most commonly used test types. These four can be placed within the categories norm-referenced and criterion-referenced assessment. Proficiency and placement tests are norm-referenced tests, while achievement and diagnostic tests are criterion-referenced tests. Again I have chosen to copy Brown's table that shows the differences between these types:

Table 2 Differences between proficiency, placement, achievement and diagnostic tests

	Norm-referenced		**Criterion-referenced**	
Test Qualities	**Proficiency**	**Placement**	**Achievement**	**Diagnostic**
Detail of information	Very General	General	Specific	Very specific
Focus	Usually, general skills prerequisite to entry	Learning points all levels and skills of program	Terminal objectives of course or program	Terminal and enabling objective of courses
Purpose of decision	To compare individual overall with other groups/individuals	To find each students appropriate level	To determine the degree of learning for advancement or graduation	To inform students an teachers of objectives needing more work
Relationship to program	Comparison with other institutions	Comparison within program	Directly related to objectives of program	Directly related to objectives still needing work
When administered	Before entry and sometimes at exit	Beginning of program	End of courses	Beginning and/or middle of courses
Interpretation of scores	Spread of scores	Spread of scores	Number and amount of objectives learned	Number and amount of objestives learned

Copied from Brown, 1996, p. 11

Table 2 shows the important differences between the four types of decision tests that Brown mentions. To sum up briefly, it shows that when it comes to what the tests focus on a proficiency test focuses on general skill, a placement test focuses on learning points that show level and skills, an achievement test focuses on terminal objectives of a course, and a diagnostic test focuses on terminal and enabling objectives of a course. When it comes to the purpose behind testing a proficiency test wants to compare individuals' scores to each other, a placement test wants to find students’ correct level of skills, an achievement test aims at determining how much the students have learned, while a diagnostic test attempts to find out what objectives need to be worked more on. There are also differences in when the tests are conducted; a proficiency test is usually conducted before entry to a program, and sometimes at exit. At placement test is conducted at the beginning of a program, while an achievement test is conducted at the end of a course and a diagnostic test can be conducted in the beginning or the middle of a course (Brown, 1996, pp. 11-17).

I will now try to place the Norwegian VG1 level English oral examination in the categories I described above. In doing so, I first want to see if it is norm-referenced or criterion-referenced. If we go to the Regulations of the Education Act, introduced in 2001, we see that these specify that assessment should be criterion-referenced in Norway. This means that the students are to be assessed according to the competence aims of the English subject curriculum, and that they should not be compared to each other as in a norm-referenced assessment. Moving on to the four following categories, the Norwegian VG1 level English oral examination seems to be most like the description of an achievement test. The examination is conducted at the end of the course, and the focus is the competence aims from the English subject curriculum, which are final objectives of the course. It is supposed to test the amount of objectives learned and the purpose is to determine the degree of learning for graduation. This makes me say that the VG1 level English oral examination in Norway is a criterion-referenced, achievement test.

In this section I have looked at the differences between norm-referenced and criterion-referenced assessment and the under categories proficiency, placement, achievement and diagnostic tests. I have looked at what characterizes each of the categories and then I have tried to place the VG1 level English oral exam in the right category. In the following I will look at another distinction in assessment; discrete point and integrative assessment.

2.3.4 Discrete-point vs integrative testing

An important distinction in assessment and testing is the one between discrete-point and integrative tests. In the following I will give a brief description of the differences between the two.

In a discrete-point test the small parts of language are measured separately. Usually this is done by using multiple-choice questions, for instance to test specific grammar issues. When using this method one often argues that if the multiple-choice questions test different aspects of grammar, the sum of them will test the students' overall grammar proficiency. It will also be argued that reading, writing, speaking and listening can be tested separately, and that the different elements of these individual skills can also be tested separately and isolated. On the other hand you have the integrative tests that test several skills at the same time. A good example of this is a dictation, where both listening skills and writing are tested at the same time. Other types of tests that are integrative are cloze tests and free writing tests. (Brown, 1996, pp. 29-30). Those who argue for using integrative tests, say that integrative tests should be used because they test language abilities in a complex way, in the same way that language is complex.

If we look at the English Vg1 level oral examination again, it is again obvious that this is an integrative test that tests multiple skills and competences at the same time, such as grammar, pragmatics, fluency, vocabulary and so on. It seems to me that it would not even be possible to administer an oral examination that would only test one skill and it is not the aim of the oral examination either, since the Regulations to The Education Act say that the oral examination should test as much of the students' competence as possible. (Forskrift til Opplæringslova § 3-30).

I have in the previous sections presented important distinctions in assessment in general, and I will in the following look at oral assessment specifically

2.4 Oral assessment

In this section I will look at oral assessment more specifically. I will look at what it special about assessing oral production, and at what is important to pay attention to when assessing speaking. Afterwards I will look at what oral proficiency is, and then at what the English subject curriculum of the Knowledge Promotion (LK06) says about oral proficiency.

2.4.1 Oral assessment in general

The assessment of oral production is as all assessment is, challenging, but is made even more so due to the nature of speaking in itself (Fulcher 1997; Luoma 2004). Luoma says that it is especially challenging to assess speaking because there are so many different factors that influence the way we evaluate someone's oral proficiency (Luoma 2004, p.1). Furthermore, it is also challenging to assess speaking tests because they are multimodal and test various skills, such as listening, speaking and often also reading comprehension (Brown 1996, p. 30). This means that how well the students do on the examination will, among other things, depend on how well they have understood the tasks given. In the following paragraphs I will look at the nature of speaking, and see why Luoma says that this in itself is a reason for assessment of speaking being so challenging.

2.4.2 The nature of speaking

As mentioned above there are several aspects of speaking that influence how we evaluate a person's oral proficiency. Elements are typically considered important are accent, grammar, vocabulary, mistakes and errors and the ability to use language appropriate to the purpose of speaking (Luoma 2004). Accent, or the sound of speech is difficult to assess because it is not easy to define what is really correct speech. Often one says that correct speech is how native speakers of a language speak. However, when it comes to a language that is as widely used as English, the way native speakers talk will vary greatly. This makes it very difficult to choose a standard in which students should be assessed according to (Luoma 2004). Research also shows that even though language learners are able to adapt a functional and understandable speech, it is very difficult to reach to the level of a native speaker. This means that most learners would fail if they were to be assessed according to a standard of native speakers (Luoma 2004). If sound of speech is going to be assessed, there are two elements that can be in focus; the accuracy of pronunciation or expressiveness of the speaker's use of voice (Luoma 2004, p. 11). It is often attractive to focus on pronunciation because that can be measured against a standard, even if it can be difficult to choose standard, as I mentioned above (Luoma 2004). In the Norwegian schools the most commonly used standards are standard British and American English. A focus on ability to create meaning in discourse would include the use of the students' ability to use stress and intonation to highlight important sentences or phrases and the choice of words to convey a message. No matter if you choose one focus or the other, or maybe both, it is important that the examiner is conscious of what he wants to focus on.

Furthermore, Luoma says that grammar is an element that influences the way we evaluate spoken production. This is an efficient way to measure proficiency because it is easy to detect how well the students master this. However, it is important to take in to account that the grammar of speech is different than that of writing. First of all, people do not usually speak in complete sentences, but often in idea units (Luoma 2004, p. 12). These are phrases and clauses that are connected with small words or with small pauses between them. This makes the grammar in speech quite different than written grammar. Second, planned speech should contain more of the written grammar structures than unplanned speech. This would be the situation with a prepared presentation in the oral examination. This is also connected to the level of formality in speech. There has to be expected a higher level of grammar in a formal situation than in an informal situation because of the nature of the situations (Luoma 1996, pp.12-13). This would be the case with an examination, which is a rather formal situation. However, you have to be conscious of the unplanned speech that will occur when the students are asked follow up questions or unprepared questions. It is important that an assessor is conscious about the differences in spoken grammar.

The third element Luoma mentions is vocabulary, or the choice of words in speech. She says that often the description of the high levels of vocabulary use, mentions the ability to “express oneself precisely and providing evidence f the richness of one's lexicon” (Luoma 2004, p. 16). It is however, important to remember that in spoken language, everyday and high frequency words are very common, and that if you manage to use these in a natural way in speaking, that is also a sign of advanced speaking skills (Read 2000). The use of generic words is also important in spoken language in contrast to written language where the use of specific words is much more common. Generic words are words such as *this, that one, that thing, fine* and *good*, while specific words are words that for example replace *this* and *that* with the word that says what *this* or *that* is, when it cannot be seen. Spoken language is much easier and faster and can refer to people and things that can be seen by saying *this* or *that*. However, the use of generic word does not always feel natural to EFL-learners because they rarely speak English outside the classroom, where the use of generic words is much more frequent. This makes it necessary to include the use of generic words in assessment criteria to show learners and assessors that the use of these words is important in natural use of spoken language (Luoma 2004). Furthermore, the use of fillers or hesitation markers is important together with the use of fixed phrases, these are important for the speaker to create time to speak and to speak naturally and fluent. There are studies that support Luomas theory that the uses of these elements makes the learners seem more fluent in their spoken production. Towell et al. (1996), Hasselgren (1998) and Nikula (1996) all support what Luoma says (references in Luoma 2004, pp. 18-19).

The fourth element Luoma discusses is slips and errors that occur in speech. It is natural that there are errors in spoken language, such as mispronunciations, the use of the wrong words or mixed sounds. The interesting thing is that if a listeners notices that native speakers have such errors in their language, they usually excuse the speaker because they believe that they know how it is supposed to be. However, when EFL-learners speak with such errors this is often considered lack of competence (Luoma 2004). Luoma says that assessors should maybe be trained in not counting all errors they hear, since this is a natural part of spoken language even for native speakers.

The last element Luoma mentions is that the speaker is using language appropriate to the purpose of speaking. As we have seen, spoken language is different from written language in various ways. However, spoken language also differs with the purpose of speaking. It is important that a speaker manages to use language appropriate to the situation he or she is in. The ability to use appropriate language in a situation is important to be a fluent speaker and it depends on many factors such as, the participants in a conversation, goal of speaking, what is going to be said etc. (Luoma 2004, pp.20-25).

In the previous paragraphs I have looked at elements that typically influence the way we evaluate a person's ability to speak, and that Luoma (2004) says that makes assessment of speaking especially challenging. I looked specifically at accent, grammar, vocabulary, slips and errors and using language appropriate to the purpose of speaking. These are elements that are important to be a fluent speaker and to score well on a speaking test.

2.5 Chapter summary

In this chapter I have looked at theory about assessment. I have looked at assessment in general and oral assessment specifically. In the next chapter I will give an introduction to the operationalization of the English oral examination. I will first look at what the construct of the examination is build on and then try to present the construct that is to be measured on the VG1 level English oral examination. I will also look at test making and development of assessment criteria. Furthermore, I will look close at the important concepts reliability and validity.

3 Theoretical overview – Construct and operationalizing

In the previous chapter I looked at assessment in general, and I gave an introduction to the history of assessment. Further I presented important distinctions between important concepts of assessment. I also looked at the nature of speaking and challenges connected to the assessment of speaking. This gave an introduction to what oral proficiency is. I will start this chapter by looking closer at oral proficiency. First I will look at what the Common European Framework of Reference for Languages says about this because this describes oral proficiency in general, and is also what the Norwegian English subject curriculum is build on. Following this, I will look closer at the Norwegian English subject curriculum, which is the construct[2] that is to be tested in the oral examination. Afterwards I will look at the operationalizing of speaking tests by looking close at different kinds of speaking tests and how to make assessment criteria. I will also discuss the important concepts of construct validity[3] and reliability, and how one can ensure that speaking tests are both valid and reliable.

3.1 The construct to be tested in the English oral examination

In this section I will try to present the construct that the English VG1 level oral examination is to test. According to the Regulations to the Act of Education it is the curriculum that is the construct for the examination (Forskrift til Opplæringslova § 3-3, §3-17, § 3-25). One part of this is oral proficiency, and I will here look closer at what that is. A study by Ang-Aw & Goh (2011) shows that raters do not have the same understanding of what the construct of oral proficiency is (Ang-Aw & Go 2011). This shows that it becomes very important to have a clear definition of what oral proficiency is. I started to look at this in the previous chapter, but in the following I will look into this in more detail.

[2] A construct what, for example a document, that explains what is to be tested

[3] Construct validity in short means that a test tests what it is supposed to test

To find out more about oral proficiency I have looked at what the Common European Framework of Reference for Languages (CEFR), says about this. The reason why I want to look at what the CEFR says is because that is what the Norwegian English subject curriculum is based on (Simensen 2010), and it gives a general description of oral proficiency. After presenting what the CEFR says about oral proficiency, I will move on to the English subject curriculum that constitutes the construct to be tested in the oral examination.

3.1.1 Oral proficiency according to the CEFR

When looking at what the CEFR says about oral proficiency, we have to look in the section about communicative language competence. CEFR says that communicative competence in a narrower sense has these components:

- linguistic competence
- sociolinguistic competence
- pragmatic competence

(Council of Europe, p.108)

This is of course just a list that needs to be explained more in detail.

Linguistic competence is described as knowledge of how to use lexical, phonological and syntactic aspect of the language. If we look back to the elements I presented in chapter 2 section 2.4.2 we see that some of those elements are a part of linguistic competence for example grammar. Furthermore, the CEFR says that linguistic competence is complex and consists of these components:

- lexical competence
- grammatical competence
- semantic competence
- phonological competence
- orthographic competence
- orthoepic competence

(Council of Europe, p.109)

Here, I will not go into further detail in what each of these comprise. Additional detail can be found in the section about communicative competence in the CEFR (Council of Europe, pp. 109-118)

The second component of communicative competence, according to the CEFR, is sociolinguistic competence. This is described as knowing the sociocultural conventions of language use. CEFR especially mentions five components:

- linguistic markers of social relations
- politeness conventions
- expressions of folk-wisdom
- register differences
- dialect and accent

(Council of Europe, p.118)

If we again look back to what I presented in chapter 2 section 2.4.2, we see that some of the elements presented there are a part of sociolinguistic competence, for example accents and vocabulary.

The third component of communicative competence is pragmatic competence. Pragmatic competence is described as knowledge of the functional use of linguistic resources, language functions, mastery of discourse, cohesion and coherence. The CEFR mentions three components of pragmatic competence:

- discourse competence
- functional competence
- design competence

(Council of Europe, p. 123)

For more detailed descriptions of what this is, again see the CEFR (Council of Europe, pp. 123-). In relation to the elements I presented in chapter 2 section 2.4.2 it is possible to place the ability to use language appropriate to the purpose of speaking, in the pragmatic competence component. This may however, also be placed in sociolinguistic competence.

As we see above the components presented are general and need to be broken down to more manageable units to be useful for assessment. The CEFR emphasizes that “no complete, exhaustive description of any language as a formal system for the expression of meaning has ever been produced” (Council of Europe, p. 108). This shows that it is very difficult to operationalize the concept of what communicative competence and oral proficiency is. However, the CEFR explains in detail what is meant by the components of each competence mentioned above and gives detailed descriptions of what is included in the competences. I will not, however, go more in detail about this here. This can however be found in the CEFR in the section about communicative competence. The reason why I will not go more in detail here is that I in the following will look closer at the details of what the English subject curriculum in Norway says about oral proficiency.

3.2 The construct definition according to the English subject curriculum

Bachman & Palmer say that “the definition of the construct is based on a frame of reference such as a course syllabus(*sic*), a needs analysis, or current research and/or theory of language use” (Bachman & Palmer 2010, p. 211). This is the case in Norway where the Regulations to the Act of Education says that it is the subject curriculum that is the construct for the oral examination. Because of this it is necessary to look closer at what the curriculum says about oral proficiency. However, as you will see when looking at the English-subject curriculum, and as I will show below, it is not only the competence aims concerning oral proficiency that are relevant for the oral examination. There are also competence aims concerning subject knowledge that are relevant for the English oral examination. In fact, the competence aims are so diverse, that I argue that there can actually be defined two constructs to be tested in the examination (cf. Bachman & Palmer 2010, pp. 239-240). I will come back to this below when I present the competence aims that constitute the construct(s).

3.2.1 The first construct: The competence aims for oral proficiency

The first construct that can be defined concerns the oral proficiency skills. Because the curriculum defines oral skills as “being able to both listen and speak” (Utdanningsdirektoratet 2006/2010), competence aims concerning both speaking and listening has to be included when we look at oral skills. According to Simensen (2010) the English subject curriculum in Norway is build on the recommendations in the CEFR, and because of this you find many of the points about communicative competence from the CEFR in the competence aims of the English subject curriculum.

The competence aims that are relevant to define oral proficiency are found in the main area *Communication* in the English subject curriculum. The competence aims are:

> The student shall be able to
>
> - understand and use a wide general vocabulary and an academic vocabulary related to his/her own education programme
> - understand oral and written presentations about general and specialised themes related to his/her own education programme
> - express him/herself in writing and orally in a varied, differentiated and precise manner, with good progression and coherence
> - select and use appropriate reading and listening strategies to locate information in oral and written texts
> - select and use appropriate writing and speaking strategies that are adapted to a purpose, situation and genre
> - take the initiative to begin, end and keep a conversation going
>
> (Utdanningsdirektoratet 2006/2010)

As we can see the competence aims are to a large extent inspired by the components from the CEFR that was presented above. They are detailed at the same time as they are open for a interpretation. For example it will be influenced by subjective opinion what a *wide vocabulary* is. It can also be discussed what a *varied, differentiated and precise manner* is. Even if the competence aims are detailed, they need to be interpreted for it to be possible to assess according to them. This can be done by making rating scales, which I will come back to later in the chapter.

3.2.2 The second construct: Competence aims for content competence

According to the Regulations to the Act of Education it is the curriculum that is to be tested in the examination (Forskrift til Opplæringslova § 3-3, §3-17, § 3-25). Furthermore, the Regulations also say that as many of the competence aims as possible should be tested in the examination (Forskrift til Opplæringslove § 3-30). This means that it is not only the communicative competence aims that are to be tested in the oral examination. In addition the competence aims concerning content competence have to be tested. This is the reason why I argue that there can be defined two constructs. By content competence I mean competence that reflect knowledge about various theoretical topics, for example knowledge about literature, knowledge about how people live in different speaking countries, knowledge about the history of English speaking countries, knowledge about culture and society, etc.

The competence aims I feel define the second construct, can be found in the main areas *Culture, society and literature* and *Language learning.* The competence aims concerning content competence are:

The student shall be able to

- discuss social and cultural conditions and values from a number of English-speaking countries
- present and discuss international news topics and current events
- give an account of the use of English as a universal world language
- discuss and elaborate on English texts from a selection of different genres, poems, short stories, novels, films and theatre plays from different epochs and parts of the world
- discuss literature by and about indigenous peoples in the English-speaking world
- describe and evaluate the effects of different verbal forms of expression
- assess and comment on his/her progress in learning English

(Utdanningsdirektoratet 2006/2010)

As we see these competence aims represent something completely different than those a presented in the previous section. However, they need to be a part of the assessment of the oral examination because the Regulations to the Act of Education say so. The remaining competence aims need not to be included because they are not directly relevant. They may be indirectly relevant, but the Directorate of Education points out that it is the students' performance at the examination that should be assessed (Utdanningsdirektoratet 2010).

3.2.3 Summing up

In the previous sections I have taken a look at what oral proficiency is and what should be assessed in an oral examination. I started by looking at the what the CEFR defines as communicative competence and further I looked closer at English subject curriculum, which is build on the CEFR. According to the Regulations to the Act of Education it is the subject curriculum that decides what should be tested in the oral examination and I have looked at the competence aims that are relevant for the oral examination. These show that there are many different aspects that have to be included. In fact, I have argued, that because of the large differences, one can say that there can be defined two constructs that are to be assessed in the oral examination (cf. Bachman and Palmer 2010). In the following sections I will look at operationalizing of speaking tests, first I will look at speaking test tasks and how to develop them, and further I will look at rating scales and how they can be made.

3.3 Operationalization – Speaking test tasks

In the previous section I presented the construct(s) that is to be assessed in an oral examination. In this section I will start to look at the operationalization of speaking tests by looking at the design of speaking test task and I will give a brief presentation to different task types.

3.3.1 Task design

When one is to design speaking tasks, one has to start by looking at the purpose of the test and the practical situation the test will be given in (Luoma 2004). The purpose of a test is important because as Ingram (1968) said “A test that is made up without a clear idea of what it is for, is no good” (Ingram 1968 quoted in Fulcher 2010, p. 95). However, the most important factor in task design is what information the result of the test has to show about the test takers' speaking skills (Luoma 2004). This means what the construct says is going to be tested in the test. An approach like this can be called a construct-centered approach which by Messick is described in this way:

> “A construct-centered approach would begin by asking what complex of knowledge, skills or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics.”
>
> (Messick 1994 quoted in Fulcher & Davidson 2007, p. 64)

This quote shows that the role of task design is closely linked to what kinds of knowledge or abilities the task designer want to say something about (Fulcher & Davidson 2007).

Furthermore a key-decision when designing tasks is to find out what the test takers are going to be asked to do with language (Luoma 2004). In task design it is common to focus on informational talk rather than social talk (Luoma 2004). Brown and Yule (1983) described four different types of informational talk: description, instruction, storytelling and opinion expressing (Luoma 2004, p. 31). Bygate (1987) made even finer distinctions between different types of informational talk:

Factual oriented talk	Evaluative talk
- description	- explanation
- narration	- justification
- instruction	- prediction
- comparison	- decision

(Luoma 2004, p. 32).

Both Brown & Yule and Bygate say that speakers' use of language is different in each of these categories. Because of this it is useful to test the different types of talk separately, and let all test takers use the same type of talk rather than different types (Luoma 2004). Consequently it is important for a task designer to be aware of the different types of talk there are and make a conscious decision of what they want to test.

Another decision the test designer has to make, is how the test is to be organized. The most common way to arrange speaking tests is by assessing the test takers one by one (Luoma 2004). This is often done in an interview format which has been the most common format since the beginning of oral assessment in the 1950's (Luoma 2004). By using the interview format, one ensures a flexible test where the questions can be adapted to each test taker at the same time as the examiners have the chance to be much in control of what happens in the interaction. The fact that the examiner has so much power over the test taker in this situation is one of the aspects that have received much criticism (Luoma 2004, p. 35 referring to

Savignon 1985; Bachman 1988; van Lier 1989; Lazaraton 1992). This format lets the examiner control the examination by asking all questions while the test taker has to answer the questions asked. However, there are other methods that can be used in one-to-one test. For example, it is possible to talk about a prepared topic, discuss announced topics, arrange a role play or do a reversed interview where the test taker asks questions of the interviewer (Luoma 2004). It is also possible that the test taker gives a prepared presentation on a given topic, and follows this up with a conversation or a discussion between test taker and examiner.

Even if the one-by-one method is the most common method to use in speaking assessment, it is also possible to give tests in pair or groups (Luoma 2004). Swain (2001) gives three arguments for this. First of all it is possible to include more types of talk when using this method and this can broaden the evidence of the test taker's skills. Second, this method encourages more pair work in teaching, and if pair work is already much in use, this method repeats what is already happening in teaching. Third, this method is more economic because it reduces the amount of examiner time needed to arrange the test. (Luoma 2004, p. 36). However, there are of course challenges with pair and group tests as well. It is argued by Weir (1993) and Iwashita (1999) that the test takers’ talk will be influenced by the other participants' personalities, communication styles and language levels, and that all the test takers may not get the equal opportunity to show their skills (Luoma 2004, p. 37). This may however, be a challenge in an examiner led examination as well. A study by Brown showed that an examiner’s personality and communication style may influence the test taker's performance. The study showed that examiners got different results when they were tested with different examiners (Luoma 2004, p. 38). Another concern with pair and group tests is that too much responsibility is left for the test takers, and that this might result in test takers not being able to show their skills at the best in this situation (Luoma 2004). This challenge can be met by the examiner ensuring that the instructions are completely clear so that the test takers know what they have to do to get a good result (Luoma 2004). In the following I will look at different test types.

3.3.2 Test types

Choosing test type is an important part of test design, and I will here give a short presentation of different speaking test types based on Luoma (2004). She splits test types into two categories; open-ended speaking tasks and structured speaking tasks, based on the relative amount of structure that the tasks provide (Luoma 2004).

Open-ended speaking tasks aim at getting the test takers to do something with language to show their language skills. This kind of tasks guide the discussion at the same time as it allows room for different methods of reaching the task requirements (Luoma 2004). Potential task types are to give a presentation, making a request, describe something, giving recommendations for something, and role-play. It is also possible to combine a role-play task with the other tasks mentioned above, for example by asking a test taker to give a speech or a presentation. Then the test taker has to structure the talk according to conventions of the task type at the same time as using the social conventions required by the role-play situation (Luoma 2004). Another task type is to make the test taker react in situations. The test takers read or hear the situation they are supposed to imagine being in and are then asked to say what they would say if they were there. The test takers then have to consider social conventions and what formulations to use (Luoma 2004). All tasks mentioned above are examples of open-ended speaking tasks.

On the other hand you have structured speaking tasks that are the speaking equivalent of multiple-choice tasks (Luoma 2004, p. 50). They quite precisely express what the test takers should say, and gives a list of answers that are acceptable. The weakness of such a test is that it cannot assess the unexpected and creative sides of speaking. However, the main strength of these tests is comparability, since they are the same for all test takers (Luoma 2004). Possible task types are reading aloud, sentence repetition, sentence completion, factual short-answer questions and reacting to phrases (Luoma 2004, pp. 50-51). These test types, however, often test one aspect of speaking at a time and might not so easily give a complete picture of a test taker's speaking skills. In the next section I will look at a different part of test operationalizing, the process of making assessment criteria, or rating scales.

3.4 Operationalization – Rating scales

The results, or scores, of a speaking test shows how well the test taker speaks the language being tested. They are usually given as numerical grades that correspond with a description of the level the grade is on. These descriptions are the rating scale (Luoma 2004). One of the first attempts of such a rating scale was made by the Foreign Service Institute (FSI) and gave descriptions of proficiency on five levels, reaching from elementary proficiency to bilingual or native proficiency (Fulcher & Davidson 2007). To read the scale in detail, see Fulcher & Davidson (2007, pp. 95-96). This scale provided by FSI has been used as a template for many rating scales produced later. These descriptors are however, very general and it might be difficult to match performance with the descriptions. Moreover, it is considered quite difficult to make rating scales for speaking because of the complexity in speaking proficiency. North (1996) quoted in Luoma (2004, p. 59) describes it like this: "trying to describe complex phenomena in a small number of words on the basis of incomplete theory". Nevertheless, rating scales describe what test developers consider is a strong or weak performance and therefore is a part of the definition of the construct tested (Luoma 2004).

3.4.1 Types of rating scales

Rating scales can be holistic, primary trait scoring or multiple-trait scoring (Fulcher & Davidson 2007). Holistic rating scales are most often very general because they are usually made by a committee of experts trying to make level descriptions of performance to a construct. With primary trait scoring, there are scales for each task, which means that they are much less general than holistic rating scales. A primary trait scoring guide, according to Weigle (2002, p. 110), consists of these components:

- a description of the task
- a statement of the primary trait (construct) to be measured
- a rating scale with level descriptors
- samples of performance to illustrate each level
- explanations of why each sample was graded in the way it was

This is of course much more time consuming and complex to make than a holistic rating scale, because there are required rating scales for each task (Fulcher & Davidson 2007). These rating scales can, however, only be used for the task in question since it is so specific. Furthermore, multiple-trait scoring can be either general, as holistic scoring, or specific, as primary trait scoring. Fulcher & Davidson say that "instead of awarding a single score to a performance, multiple scores are awarded. Each score represents a separate claim about the relationship between the evidence and the multiple underlying constructs" (Fulcher & Davidson 2007, p. 97). Multiple-trait scoring systems can be developed either by expert committees or through analysis of samples.

3.4.2 Methods for developing rating scales

Making rating scales is difficult and challenging and because of that it might seem like a good idea to adapt an existing rating scale when developing a speaking test. However, a rating scale has to be related to the purpose of the test, and the definition of the construct being tested (Luoma 2004). It is of course possible to use already developed rating scales as a basis, and modify them according to the test at hand (Luoma 2004). Luoma (2004) describes three methods of developing rating scales for speaking tests that I will present here; intuitive methods, qualitative methods and quantitative methods.

Intuitive methods of rating scale development are based on interpretation of experience rather than on data collection. It can either be one person or a group of people developing the scales. The developers are usually experienced and trained persons with teaching experience at the relevant levels. It is common to consult already developed rating scales, curriculum and teaching material when making the scales. Frequently, the scales are also revised once or several times, especially when there is a group or committee of developers that are to agree upon the scales developed. (Luoma 2004).

In qualitative methods of rating scale development, a group of experts are asked to analyze data related to the scale. This can either be samples of performances on different levels, or scale level descriptors. If level descriptors are used, they can be given to the group of experts without saying which level the descriptions are on. The group is then asked to rank the descriptors according to difficulty and group them in a number of levels corresponding to the scale. If samples of performance are used, they can be analyzed by the group according to the rating scale, if there exists one. Another possibility is that the group members bring with them performances that they feel represents the different levels well. Yet another possibility is to analyze performances and grade them numerically. Afterwards, the raters are asked to explain why they gave the grade they did. These explanations are then used to make level descriptors. (Luoma 2004).

If one is to use quantitative methods of rating scale development it requires statistical expertise. An example is how Fulcher used a quantitative method when developing a rating scale for fluency. He conducted a discourse analysis of a set of performances and counted the occurrence of a range of fluency features in them. Then he used multiple regression analysis to determine which features were significant in determining the test taker scores. The features identified were then used to develop level descriptors. (Fulcher's method is described in Luoma 2004). Furthermore, Luoma (2004) says that another advanced quantitative method for developing rating scales is item response theory (IRT), which is a development of probability theory. There are several IRT model to choose from, but the most simple and robust is the George Rasch model.

Above I have presented three methods of developing rating scales provided by Luoma (2004). These are the intuitive method, the qualitative method and the quantitative method. They are different in several aspects and which one to use depends on a number of elements, such as purpose of testing, data availability, time at disposal etc. In the following section I will look at one of the most important concepts in assessment theory, reliability.

3.5 Reliability

Reliability is one of the most important concepts in assessment theory. In the following I will look at what reliability is in general and further move on to looking at reliability in speaking tests in particular.

3.5.1 Reliability of tests and scores in general

Reliability is closely connected to validity, and the reason for that is that performance scores can generally not be considered valid if they are not reliable (Alderson et al. 1995; Bachman 1990). The classic definition of reliability was provided by Lado (1961):

> “Does a test yield the same scores one day and the next if there has been no instruction intervening? That is, does the test yield dependable scores in the sense that they will not fluctuate very much so that we may know that the score obtained by a student is pretty close to the score he would obtain if we gave the test again? If it does, the test is reliable”
>
> (Lado 196, as cited in Fulcher 2010)

Said in another way, reliability has to do with the extent to which scores are consistent (Brown & Hudson 2002; Henning 1987; Luoma 2004). Reliability is important because if results are dependable, we can rely on them in decision taking (Luoma 2004). If test scores are not reliable they can have big consequences for test takers, for example wrong placements, unjustified promotions, or undeserved low grades (Luoma 2004).

There are of course factors that can threaten the reliability. Some of them are variation in administration, quality of the test, differences in test forms, changes in test takers over time, differences in scoring and differences in raters (Fulcher 2010). Lado (quoted in Fulcher 2010) claimed that no test is a perfect measure. The reason for this, Lado claimed, is for example the problem of how to choose what to be tested, since everything cannot be tested in one test. Furthermore, he noted that if a test's items test very different things, it would reduce its reliability. Finally, he highlights that unreliability can be caused by the scoring (Fulcher 2010). These elements that can threaten or influence the reliability are often referred to as measurement errors, and are sources of inconsistency in test scores that will affect the consistency of test scores (Bachman 2004).

To test the reliability of tests, Fulcher & Davidson (2007) present a few methods:

- Test-retest
- Parallel forms
- Split halves

Test-retest means to administer the same test twice and calculate a correlation between the scores of the tests. For this to be possible, there must not have been any learning taking place between the two tests and there can be no practice effect from taking the test a second time (Fulcher & Davidson 2007). Parallel forms is to make two forms of the same test in the way that they test the same construct and have similar means and variances. If there is a correlation between the scores of the two tests, it is taken as a measure of reliability (Fulcher & Davidson 2007). The method of split halves is to split the answers to a test in two halves. One half is taken to represent one form of the test and then correlated with the other half (Fulcher & Davidson 2007, p. 105). The correlation coefficient is taken as a measure of reliability.

Moreover, reliability can be calculated. Common ways to calculate reliability is by using Kuder-Richardson formula 21, Kuder-Richardson formula 20 and Cronbach's alpha. I will not look at these in detail, but for a presentation of the use of these, see Fulcher & Davidson (2007, pp. 106-108).

In this section I have looked at the issue of reliability in general. In the following I will look at the reliability of speaking tests in particular.

3.5.2 Reliability of speaking tests

According to Luoma (2004), the reliability of speaking tests “builds on high-quality instruments and procedures” (Luoma 2004, p. 176). She also says that the methods for ensuring reliability are different from formal tests given by examination boards and for classroom tests and assessment.

Luoma (2004) presents three types of reliability that are especially relevant to the assessment of speaking. These are:

- intra-rater reliability
- inter-rater reliability
- parallel form reliability

Intra-rater reliability means that raters agree with themselves, over a period of time, about the ratings they give. Inter-rater reliability means that different raters rate performances similarly. However, they do not have to agree completely even if the ratings should not be very different. This is helped with well-defined assessment criteria. Parallel form reliability is if there is more than one test format that is supposed to be interchangeable. Examinees are to take more than one of these test formats, and then the scores are analyzed for consistency. If the scores are not consistent, one might have to change some of the tasks in the test formats to make them more consistent. (Luoma 2004).

For formal tests given by examination boards the most common way to ensure reliability is rater training (Luoma 2004). Usually a rater training course starts with an introduction to a test and the assessment criteria. Then the different levels are presented through taped performances of students during a test, a *benchmark tape* (Luoma 2004, p. 178). After having this illustrated, the course participants are to assess other taped performances, report their assessment aloud and discuss the reasons for the scores given. Often the course participants are to discuss the scores they have given in groups and reach a consensus score. At the end of a rater training course there is often a qualifying test where the participants are given taped performances to assess on their own, to see if the score they have given is in consistence with scores given by already qualified raters. (Luoma 2004).

Rater training has been criticized because one says that the training changes the way the participants understand rating to ensure reliability, but does not provide any proof for the assessment criteria are valid (Fulcher 1997 quoted in Luoma 2004). However, Luoma (2004), says that this just means that the developers have to show that the assessment criteria are "related to the construct definition, the tasks and the kinds of skills that the speakers need outside the test" (Luoma 2004, p. 177). The fact that examination boards give rater training, shows that they know it is almost impossible to give comparable test scores without training, and that they are doing something to ensure this because it is so important.

Another method examination boards often use to ensure reliability is standard setting or the setting of cut scores (Luoma 2004). To set cut scores you have to let "known masters and known non-masters (of a certain score band) to take the test in order to define the cut point between that band and the one below it, or getting experts to describe how well examinees at certain levels, e.g. 1, 2 and 3, would do on the test tasks" (Luoma 2004, p. 178). Moreover, it is also possible to ensure reliability by ensuring consistency of rating procedures (Luoma 2004). A common way of doing this is to use rating forms.

When it comes to ensuring reliability for classroom assessment a common issue is subjectivity. To reduce subjectivity it is often suggested to assess test performances task by task and to try to assess performances anonymously (Brown & Hudson 2002; Luoma 2004). However, it is very difficult to use these methods for speaking tests. It is impractical to assess task by task and it is close to impossible to ensure anonymity since the test takers have recognizable voices that a teacher would recognize (Luoma 2004). Luoma (2004) says that practically the only way to ensure reliability in classroom assessment is for the teacher to be aware of the issue and reflect on his or her own assessment practice. It is necessary for the teacher to focus on being just and using assessment criteria correct.

In this section I have looked at theory concerning reliability which together with validity are two of the most important issues in assessment theory. I have presented theory of reliability in general and theory that is especially relevant to speaking tests. In the following section I will look closer at validity.

3.6 Construct validity

Validity is one of the most important concepts when talking about assessment. It is at the center of all discussions of assessment (Fulcher & Davidson 2007; Fulcher 2010). In this section I will present what construct validity is, give a historical overview of how validity thinking in assessment has developed and look at validity in speaking tests in particular.

3.6.1 Defining construct validity

Until 1989 there had been used the same definition of construct validity for decades:

> “By [construct]validity is meant the degree to which a test or examination measures what it purports to measure. Validity might also be expressed more simply as the 'worthwhileness' of an examination. For an examination to possess validity it is necessary that the material actually included be of prime importance, that the questions sample widely among the essential over which complete mastery can reasonably be expected on the part of the pupils, and that proof can be brought forward that the test elements (questions) can be defended by arguments based on more than mere personal opinion.”
>
> (Ruch 1924 quoted in Fulcher 2010)

With this definition one can say that the important validation question was: does my test measure what I think it does? (Fulcher 2010). After Messick’s work from 1989 our understanding of validity has changed. Now it is seen as “a single concept, with a number of different facets, or aspects” (Fulcher 2010, p. 20). After this validity is traditionally understood to mean to find out if a test “measures accurately what it is intended to measure” (Hughes 1989 quoted in Fulcher & Davidson 2007, p. 4), or uncovering the “appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure” (Henning 1987 quoted in Fulcher & Davidson 2007, p. 4). This in short terms means that validity is a question of whether a test measures what it is supposed to. For a more detailed introduction to the history of validity thinking, see appendix 3.

3.6.2 Validity in speaking tests

According to Luoma (2004) the question of validity is so fundamental in all assessment theory that it is very difficult to highlight aspects of validity that it more relevant to speaking tests than others. However, she has points of view on how to ensure a valid speaking test. She says that “the validity of speaking scores is grounded in the purpose that the scores are intended to serve “ (Luoma 2004, p. 185). Consequently, the first thing a speaking test developer needs to do, is to take some time to specify what is the purpose of the test. Further, the test developer should try to define what kind of speaking the test is supposed to assess, and what the construct is. When the purpose of the test is specified and the construct defined, the next step to ensure validity is to show evidence from the test development process that the test actually implements the construct (Luoma 2004).

To produce such evidence one has to use the task specifications to characterize the tasks, then explain the relevance of the tasks to the purpose and the degree to which they are representative of possible tasks this purpose can include (Luoma 2004, p. 186). The next step in ensuring valid speaking tests, is to evaluate the assessment criteria (Luoma 2004). These have to be coherent with the purpose of the test, the construct definition and the tasks. They also have to be concretely defined so that they are easy to use. Finally, a step to ensure valid speaking tests is to look at the planning and monitoring the test developers do related to score use (Luoma 2004). One can look at administrative records, but it can also be relevant to look at the examinees experience with the test and look at the washback effect the test has (Luoma 2004).

3.6.3 Summing up

In the section about construct validity I have defined construct validity and looked at the development in validity theory. I have also looked closer at some theories concerning validity. In the last paragraphs I looked at validity in speaking tests in particular and presented steps to ensure valid speaking tests. In the following section I will give a chapter summary and briefly repeat what I have presented in this chapter.

3.7 Chapter summary

In this chapter I have looked at several theoretical aspects of assessment. I started the chapter by introducing what the Common European Framework of Reference for Languages (CEFR) defines as communicative competence. The reason why I did this is that the English subject curriculum in Norway is build upon the recommendations of the CEFR. Because the curriculum is the construct to be tested in the oral examination, I needed to start by looking at what the curriculum is build on. Furthermore, I looked at the English subject curriculum. In doing so, I found that it is possible to argue that the competence aims that are relevant for the oral examination contain are so diverse aspects that it is possible to argue that there can be defined two constructs. The first construct that can be defined concerns the oral proficiency skills. These competence aims are found in the main area *Communication* in the English subject curriculum. The second construct that can be defined concerns content competence and these competence aims can be found in the main areas *Culture, society and literature* and *Language learning*.

After having presented the construct(s) to be tested in the oral examination, I moved on to looking at operationalizing of speaking tests, first by looking at task design, and next by looking at the development of rating scales. Finally, I presented the important concepts of reliability and construct validity. Reliability is in short terms a question of whether the scores of a test are consistent or not. In speaking tests there are three types of reliability that are especially relevant; intra-rater reliability, inter-rater reliability and parallel form reliability. Furthermore, the most common methods to ensure reliability in speaking tests, is the use of assessment criteria, rater training, the setting of cut scores and of course to make sure that the raters are aware of what might lower the reliability so that they focus on being consistent in their assessment. Validity is a question of whether or not a test tests what it is intended to test. To ensure the validity of a speaking test, there are several steps that can be taken; a speaking test developer needs to take time to consider what is the purpose of the test, define what type of speaking the test is to assess and also define what the construct is, then evidence from the test development process that the test actually implements the construct has to be shown.

In this last section I have tried to sum up what I have gone through in this chapter. In the following chapter I will present the methods I have used for my project before I present the results and findings of my project in chapter 5.

4 Methods

4.1 Introduction

In this chapter I will go through the procedures that have been used during the research process and the choices I have made while working on this project. First of all I will briefly go through an overview of what a research process consists of and then next I will go on to explaining my research process. I will go through the parts; preparations, collecting data, analyzing data, presentation of the data, and in the end I will discuss the validity of the results of my project.

4.2 The research process

Before saying anything specific about my research process I will here give a general introduction to what a research process is. The process is normally divided into four phases (Johannesen, Tufte, & Kristoffersen, 2006, p. 37). These phases are:

1) Preparations
2) Collecting data
3) Analyzing data
4) Presenting results/ report

4.2.1 Preparations

Starting off with the preparations, this is a crucial part in all research. First of all you have to have an idea and an area of interest that you want to have a closer look at. As a researcher you have to be curious and have questions you want to find answers to. The next step is to narrow these questions down to a research question that can be used as a starting point for your project. In the preparations you also have to decide what kind of research you want to do, and what research design you want to use. There are several research designs you can choose between; ethnographic, phenomenology, case study, grounded theory, experiments and questionnaire (Johannesen, et al., 2006, p. 37).Both qualitative and quantitative research designs are represented in these categories and I will give a description of the differences further below.

4.2.2 Collecting data

The second phase is collecting data. When conducting research one has to collect documentation. No matter if you decide to do a qualitative or a quantitative research and regardless of what research design you have chosen, you need to decide who will participate in the survey or which textbooks or documents you will study. As a researcher you also need to decide how many informants you want, and how you are going to find and choose these informants. At this point the choice of how to collect data is dependent on if it is a qualitative or a quantitative study. In qualitative studies common ways to collect data are by interviews, observation or focus groups. In quantitative studies one usually uses questionnaires to collect data (Johannesen, et al., 2006, p. 38).

4.2.3 Analyzing data

After collecting the data it is necessary to analyze it. In any case it is necessary to reduce the amount of data to make it possible to handle for the researcher. Methods of analysis differ from qualitative to quantitative studies. Analysis of qualitative data has to do with editing and analyzing, while analysis of quantitative data is done by using statistical techniques (Johannesen, et al., 2006, p. 38).

I will in the following connect these three phases to my own study and more in detail explain them as I explain how I have gone about the different phases in my study.

4.3 Preparations

At this step there were a number of decisions that had to be made. I needed to find a research statement, find out if I wanted to do a qualitative or a quantitative study, and, finally, the research design I wanted to use.

4.3.1 Research statement

It is not as easy as it might seem to find a good research question. It has to be focused and narrowed down from the idea you initially had, and you have to find a question that hopefully can be answered within the limits of a master thesis. First I carried out a pilot study where I examined the oral examinations at the VG1 level in three different schools across the country. My focus was on the structure of the oral examinations and to see how big the differences were. I also wanted to see if the examinations were fair for the students, if they were reliable. This felt satisfactory for the pilot study, but for the master thesis I wanted to do a study showing more and had to change the research statement. After changing it back and forth I in the end decided to have this research statement: *How are English oral examinations at VG1 level in Norway designed, carried out and assessed across the country?*

4.3.2 A qualitative or quantitative study?

As mentioned above there are important differences between quantitative and qualitative approaches to a study. A quantitative study counts how often phenomena occur to find out more about the phenomenon and the data can be counted and statistically processed. The qualitative approach on the other hand, aims at describing and characterizing, providing extensive information about the topic. This approach is especially applicable when examining phenomena we know little about. (Johannesen, et al., 2006, p. 36; Bakke, 2010, p. 43).

When deciding which approach to use you have to take into consideration what the purpose behind the research is. If you want to have a large number of respondents to make generalizations, and/or test already developed theories and hypotheses, a quantitative approach is the most suitable. However, if you want to describe a phenomenon the way it is understood and evaluated by people, and perhaps also develop new theories, then a qualitative approach is often the best. (Ragin, 1994; Faye-Schøll, 2010). If you choose a quantitative approach you need a large number of respondents, while if you choose a qualitative approach you will just need a few respondents.

The methods used for collecting data also vary between these two approaches. For a qualitative study the most common methods are observations and in-depth interviews, while for a quantitative study methods such as these would not be appropriate because of the number of respondents you need in such a study. The most common method to use for a quantitative study is to distribute a questionnaire.

In my case it was from the start clear to me that I wanted to do a qualitative study. I wanted to find out more about how oral examinations were structured across the country. This was because I found it very interesting that the Directorate of Education had left the responsibility for these examinations to the county administrations, while at the same time it was so important for them to control the written examination. I also wanted to see if there were as big differences between the counties and schools as I expected, and to find out what the teachers' attitudes towards the system of the oral examinations were. I am sure that I could have chosen both qualitative or quantitative approaches to this, but I chose a qualitative approach because I found this the best approach for an exploratory survey. I also had heard from people doing their master thesis earlier that doing a quantitative study could be even more time consuming and that it often is very difficult to recruit enough respondents. Within the limits of this thesis I did not consider it possible to conduct a study around this subject with a quantitative approach. After deciding the approach, the next step was to find out what kind of research design I wanted to use. I will first present some of the research designs that exist in qualitative method and then explain how I made my choice.

4.3.3 Research design

Johannesen defines research design as “How a study is organized and conducted to make it possible to answer the research question “ (my translation Johannesen, et al., 2006, p. 347). There are a number of possible qualitative research designs to choose from. When deciding which one to use, the researcher has to consider what the purpose of the study is. I wanted to describe and try to understand the situation of the oral examinations in upper secondary school, and according to Johannesen, there are the four possible research designs for me to choose from; grounded theory, case design, ethnographic design and phenomenology (Johannesen, et al., 2006, p. 88). Grounded theory aims at developing new theories with the new data as a starting point (Johanesen, et al., 2006, p. 82). My aim was not to develop new theories, so I therefore ruled this out. Ethnography wants to describe and analyze a culture, a social group or a social system, the researcher tries to find patterns, typologies and categories within the group he is studying (Johannesen, et al., 2006, p. 83). Neither was that my goal for the study, so I had to rule out this as well. A case study studies one or two cases of a phenomenon (Johannesen, et al., 2006, p. 95) and that was not suitable for me since I wanted to look at the situation across the country. This means that I was left with phenomenology as the research design. Phenomenology aims at describing people and their experiences with and understanding of a phenomenon (Johannesen, et al., 2006, p. 80), and I wanted to describe the oral examinations in upper secondary and the teachers' experiences with this. The best method of collecting data for this purpose is through conduct interviews with 5-25 informants according to Johannesen (Johannesen, et al., 2006, p. 81). This brings me on the next step in the research process; collecting data.

4.4 Collecting data

In the following I will go through how I chose interview method, I will present my interview guide, how the informants were chosen, how the interviews were conducted and in the end I will give a short presentation of my informants.

4.4.1 Method of interview and interview guide

When conducting interviews there are 3 different types of interviews that can be used; unstructured interviews, semi-structured interviews and structured interviews. An unstructured interview is an informal interview with open questions. You have a topic but you have not constructed the questions before the interview. These are made during the interview and adjusted to each informant. Some of the advantages of this are that it is very informal and because of that it is comfortable for informants. It seems more like a conversation than an interview. The biggest disadvantage with this approach is that it makes it very difficult to compare the interviews and systematize the data after conducting the interviews. (Johannesen, et al., 2006, p. 138)

A semi-structured interview is more structured with regards to the questions used. In this case the researcher has decided on the topic and made an interview guide in beforehand. The questions are general and can be changed around when needed, and the researcher can ask follow up questions if that is needed (Johannesen, et al., 2006, p. 139). The advantages of this are that the data can more easily be compared and systematized, at the same time as you are not bound to ask only the questions that you have prepared.

In comparison, in a structured interview the researcher has questions decided before the interview and with no possibility of adding extra questions or changing the order of the questions. An interview like this makes it very easy to compare interviews, it will also be less time consuming to conduct and can be conducted by different interviewers. The disadvantage is of course that there is little or no flexibility to add questions that can be interesting. (Johannesen, et al., 2006, pp. 137-138).

I chose to use the semi-structured interview for my study. It seemed to be the most appropriate method since I wanted to be able to compare the answers and therefore did not want to run the risk of that being too difficult using an unstructured interview. At the same time as I wanted it to be possible for me to add questions I found of interest when interviewing the informants.

However, what I found when I was trying to recruit informants, was that it was much easier to convince them to say yes, if I offered them to answers the questions by email. I will come back to this in section 4.4.3. This led me to make the decision of using a combination of oral interviews and open questions answered by email.

The next step was to design my interview guide. First of all I had to find out what I actually wanted to find out. I wanted to know how the oral examination was structured, who was responsible for making the examination, what they aim at testing, how the examination is assessed, whether the teachers found the examinations fair or not, and whether they wanted a centrally or locally given examination. This gave me an interview guide with the following six questions:

1) How is the oral exam in your school structured? How is it conducted?
2) What is tested in the exam?
3) Who is responsible for making the exam? Is it locally given or has the county administration given any instructions for it?
4) How is the exam assessed? Who assesses? Do you have assessment criteria that have to be used by all examiners?
5) Your opinion: is it a fair exam?
6) Your opinion: Is it good that it is the county administration that it responsible for the exam, or do you wish for a centrally given oral exam as the written exam?

After having written my pilot paper using this interview guide when interviewing three teachers, I found that there was one more question I wanted to add:

8) If you are to make a prioritized list, what are the five most important criteria you look for when assessing in an oral exam?

These are the questions I have used when interviewing all my informants, when it has been necessary I have added questions. In the following I will go through the process of choosing my informants.

4.4.2 Choosing informants

At this point in the research process I had to decide how I wanted to find informants and how many informants I wanted to have. Since the idea for the thesis came from that it is the county administrations that are responsible for the oral exams, I found out that it would be appropriate with one informant from each county, namely 19. It is, of course, also important to find reliable sources that fit the study. My research question asked for the situation in upper secondary, and since it is only the first year of English that is compulsory, I had to find informants who had experience in teaching English at VG1 level and who also had experience in conducting oral examinations. Other than that there were no formal requirements for my informants. From earlier experience I had found that it does not work to send emails to schools asking if they might have teachers there who can be interested in being part of my study. Therefore, I understood that I had to find another way. My first three informants were provided to me by my thesis supervisor. Two informants were found by contacting relatives of mine that had contacts willing to be a part of my study. And for the rest, I simply entered the county administrations' web pages, found links to upper secondary schools which I called and explained the situation to. In some schools the administrators said that they would get back to me, but they rarely did. However, in many schools I got names of possible teachers that I could contact either by telephone or email. Of course not all I contacted answered me or wanted to be a part of my study, but I got in contact with many very helpful teachers who said they were willing to participate in this study. After calling many schools I finally managed to have one informant from almost each county, but as the time went I could not wait any longer to find informants from the last 2 counties. In these two counties it seemed almost impossible to find someone who had the time and wanted to be a part of my study, so because of the lack of time I ended up with informants from 17 counties. In one of the counties I actually ended up doing two interviews, this because the first interview did not provide enough information because of the teachers' lack of experience with oral examinations. I have chosen not to include that interview in my thesis. Unfortunately, the material from one of the interviews was lost because of technical problems and could not be restored. This means that I conducted 18 interviews, but the thesis is built on 16 of them, which was less than the initial goal for number of informants, but still provided more than enough information for me to build my thesis on.

Later in this chapter I will give a short presentation of my informants and I will now only give brief information about the spread in age, experience, gender etc. From my 16 informants, there were 12 women and 4 men. The number of men should maybe have been higher, but since there are more female teachers in the Norwegian schools, and because the information they provided is not dependent on their gender, I do not see this as a problem. In age they are spread from the age of 32 to the age of 63, which I also think represents a reasonable spread in age. When it comes to education some of the teachers are educated at a teachers colleges while others have done a major at the university in addition to their pedagogical course, PPU, some are what we in the Norwegian school system call *adjunkt* and others *lektor*. I asked the teachers for how many years experience they had from working in school and this varied from 5 to 35 years, and many had 20 years or more experience. To sum up, I think I got informants that were reasonably diverse with regards to age, education, gender and experience. In the following I will go through how I conducted the interviews.

4.4.3 Conducting the interviews and open surveys

As mentioned above I interviewed 18 teachers from 18 different schools from a 17 counties across the country, and the present study is based on 16 of these interviews. For practical reasons only one of the interviews were made in person. This would have been too money and time consuming, considering the informants are spread all across the country.

The three first interviews were done by telephone. I first had email contact with the teachers, and sent them the interview guide in beforehand. I considered whether or not to do so, but decided that I wanted them to be able to think a little about the questions before we had the interview. Of course there is a possibility that when the answers are not spontaneously given, they can be too well thought through and maybe not correct, but because I have asked for their opinion on the subjects and also asked them to make a prioritized list of important elements I wanted to give them time to think well through this before giving the answer. I also hoped to get more detailed answers when the teachers had this opportunity. I did not tape the interviews, at first this was my intention, but I was not able to get a tape recorder in time for the interviews. However, I made sure to make thorough notes during the interviews, and directly after the interviews I went through the notes again and wrote them over to make sure I would not loose valuable information.

When I was to conduct the interviews I did after finishing the pilot study, I first wanted to do all the interviews by telephone, but to make it as convenient as possible for the informants, I asked them whether they wanted to answer the questions by telephone or by email, or in person if that was better. I found that the questions could be answered by e-mail as long as I had the possibility to send follow up questions when that was necessary. Out of the teachers I interviewed in this process, one of them chose to do the interview in person and the rest wanted to answer by e-mail. I think the reason for this is that when the teachers can answer by e-mail, they can do it at a time that fits them the best. I also noticed that it was much easier to get teachers to answer the questions when I said that they could be answered by e-mail. In some cases there was no need to send follow up questions since the answers I got were sufficient, in other cases I sent another e-mail with clarifying questions when I felt that was needed. An advantage I found when the interviews where done by e-mail, was that I immediately got the answers in writing and that made the process of documenting and sorting the answers much easier.

When it comes to the choice of language, all of the interviews were done in Norwegian. Of course I could have given the teachers the choice of language, but none of them seemed uncomfortable with speaking or writing Norwegian when I first contacted them, so that was the reason why I thought it would be convenient to use Norwegian as the interview language.

Of course all names of teachers and schools have been changed to secure the informants' anonymity, which is important according to ethical guidelines (Johannesen, et al., 2006, p. 98). When I in the following will give a short presentation of the informants all names and locations have been changed.

4.4.4 Short presentation of informants

In the following I will give a very short presentation of my informants. As mentioned I did not know any of the informants and they were chosen as randomly as possible. All of my informants work in upper secondary school and teach English at the VG1 level. All of them have experience when it comes to structuring and conducting oral examinations. The teachers work in schools spread all across the country, all from different counties and all of the schools are public schools. I did not ask the teachers very much of their background, what I did ask them was about their age, their educational background and how many years of teaching experience they have. All the names in the following are fictitious. I have chosen to present the background of the teachers in a table.

Table 3 Teachers' bakground

Teacher	Age	Educational background	Years as teacher
Helga	39	7 years studying at university level. One year of studying law. Major in German in addition to English studies	13
Birgitte	53	*Adjunkt* educated in a university. Minor subject in English. Studied 1 year of business English in Norges Handelshøyskole. Has finished the 1st year of English as major subject.	26
Liv	36	Has bachelor degree specialized in English subjects, and in addition has studied social science/history and has attended courses in local history. Has also taken the PPU course.	5
Erik	55	*Lektor* with major in English	28
Ellen	51	3 years of teachers education in a teachers college. Has started English major, but did not finish it. Have also studied 2 years computer subjects	26
Kari	57	*Lektor*	6
Silje	35	*Adjunkt* with English and Geography as subjects	9
Tor	39	Cand mag with Norwegian, English and History of ideas. Has also taken the PPU course. Is qualified as *adjunkt* with additional education	14
Karine	53	*Lektor* with additional educaion. 3 years of teachers education. Has a minor degree in English and a major degree in Norwegian didactics. Has also attended courses in Music.	28
Tonje	39	*Lektor* with additional education. Has a major degree in American Literature and a minor degree in French.	16
Mari	50	*Adjunkt* with English philology and hunanistic sciences as her subjects.	24
Ingrid	63	Has a major degree in English. Also have experience and education as a translator in English.	25
Elisabeth	32	4 years of teachers education and in addition has a minor degree in History and a nasters degree in English literature. Qualifies as *lektor* with additional education.	7
Lars	48	Cand philol with a major degree in English.	17
Hanne	59	Has a major degree in English in addition to basic course in German. Has also attednded a basic course in Social anthropology. 2 years of education at a teachers college	35
Petter	56	Has economical studies as a background and is today department manager in his school. Has also been responsible for all exams in	29

		the school earlier.	

In the table above you see the answers the teachers gave me about their background and as mentioned above they are fairly diverse in all aspects some with interesting subject combinations and educational backgrounds. In the following I will go through how I have worked to analyze my data.

4.5 Analyzing data

After having collected all the data, the next step in the research process is to analyze the data. This is necessary to be able to answer your research question based on the data you have collected. In qualitative research this means that you have to work through the data, in my case interviews, compare the answers, find similarities and differences, patterns and other interesting elements (Johannesen, et al., 2006, p. 164). When doing this with my interviews I started by making a table with all the answers from all the teachers, later I found that this was not enough and understood I had to make a table for each question in order to get a complete overview. This made it easier to compare the questions and look for similarities and differences. The data was then made accessible and it was possible to use the answers to build a text and a presentation of the data. It was very interesting to see all the answers in that way, because then I could truly see the patterns or the chaos that appeared. In the following I will go through the next step of the research process; presenting data.

4.6 Presenting data

As I mentioned above research such as this is commonly presented in written reports, such as this thesis. It was now necessary to find out how to present the data in a best way possible to make it easy for the readers of my thesis to read and understand my findings. In my pilot study I chose to present my data as three case studies. However, this did not seem appropriate for this larger study since it would result in repeating much information several times. Consequently, I have decided to present my data thematically, going through the questions one by one and looking at my findings from my 16 respondents. This seems as the most effective and appropriate way to go about this. In addition I have made some tables to clarify my findings where that seemed appropriate.

4.7 Are the results of the study trustworthy?

As a last part of the method chapter I am going to try to evaluate whether my findings can be trusted. When evaluating a qualitative study as this, there are according to Johannesen three elements that need to be discussed. These are: construct validity, transferability and confirmability. In the following I will go through these elements one by one.

4.7.1 Construct validity

Construct validity questions whether we have measured what we want to measure in the study, whether we got answers to the questions we wanted or not. Validity is concerned with whether we got answers to our questions that actually reflect the reality we wanted to look at, and made observations that were able to describe the phenomenon we wanted to study (Johannesen, et al., 2006, p. 199). There are methods that can be used to assess the validity, these are continuous observation and triangulation. Continuous observation means to observe a phenomenon over a longer period to get to know it so well that you can distinguish between relevant and irrelevant information. Triangulation means to use different methods to study the same phenomenon. (Johannesen, et al., 2006, p. 199).

Within the limits of my study I did not find it possible to use any of these methods. However, I have conducted 18 interviews of which 16 have been used for this thesis. I made sure that my informants came from different parts of the country and that they were chosen randomly. I feel that I have found some answers to my questions and in that way have measured what I wanted to measure. Of course I understand that a selection of 16 teachers out of hundreds in Norway cannot give an absolute answer to my questions, but it does show patterns that are interesting and that I feel show some of what I wanted to find out. So, the external validity might be limited, but I still feel that my respondents have given a useful picture of the situation of oral examinations at the VG1 level.

4.7.2 Transferability

Transferability is also called generality and is concerned with whether it is possible to transfer you results to other related phenomena (Johannesen, et al., 2006, p. 200). As I said above I understand that 16 teachers are a small group and cannot give a complete picture of the situation in Norway. At the same time I do think that they can give a picture that presents the main patterns of the larger picture.

Many of the teachers told of a system where there were certain guidelines from the county administration that need to be followed by all schools within the county, which should mean that the oral examinations are rather similar within these counties. What I do think my study established, is the fact that we have a situation in Norway where students in upper secondary are not assessed in the same way in oral examinations across the country, at the same time as they are supposed to compete for the same entrance to higher education with their results from this assessment. I find this phenomenon very interesting and I think that a larger study might show that the patterns I have found in my research are correct. However, I cannot say that I can generalize and claim that this is the situation in all schools at the same time. What I have discovered is that there are large differences between schools and counties.

4.7.3 Confirmability

Confirmability means that the results of a study reflect the results of the research itself and not the subjective assumptions and attitudes of the researcher. This can be ensured by you as a researcher describing all of the research process so that the reader can for himself assess the confirmability (Johannesen, et al., 2006, p. 201).

Hopefully, I have described the whole process step by step and the results of my findings, and I believe that the results reflect the reality I found and not my own opinions.

4.8 Summary

In this chapter I have described the methods used when conducting this study. I have given a general overview of the research process and connected it to my own study discussing preparations, collection of data, analysis of data and how to present data. I have also given and overview of how to ensure that a research is reliable, trying to evaluate my own study.

5 Results and analysis

5.1 Introduction

In this chapter I will present the findings from my project. I will do so thematically, using the main questions that I asked the teachers as themes. These are as follows:

1) Structure of the oral examination
2) What is tested in the exam?
3) Who is responsible for making the examination?
4) Assessment
5) The five most important criteria when assessing an oral exam
6) Teachers' opinion: Is it a fair exam?
7) Teachers' opinion: Centrally or locally given oral exam?

In going through these I will try to analyze and interpret the findings. As I presented in the method chapter I have interviewed 16 teachers in 16 different schools in 16 different counties. In this chapter I will give examples of what I have found, but for more details I have included a table with an overview of each informant's answers in the appendix.

5.2 Structure of the examination

All my informants were asked these following questions about the structure of the exam: How is the oral exam in your school structured? How is it conducted? By asking these questions I wanted to find out about the format of the oral examinations across the country. I expected that there would be quite big differences, and even though I was surprised that there were fewer differences than I first expected, the differences are worth noticing.

The main finding is that there is a pattern of two different models of the oral examinations, which I have called the project model and the question model. Not all of the informants call the format of the exam by these names, but the structure of the examination comprises what is included in these models even if they do not call it the project model or the question model. I will now describe these two models briefly.

The question model gives the students a 30-minute preparation period prior to a 30-minute examination. During the 30 minutes of preparation the students are given a subject to prepare as well as possible for the examination, and they are allowed to use all available sources. The subject is connected to topics that have been gone through during the school year and can either be given to the student, or it can be drawn from a selection of subjects.

During the examination the students are asked questions about the subject by the examiners. Usually it is the internal examiner who leads the questioning, while the external examiner can add questions and comments when they want to. The examination comprises a 30 minutes dialogue between the student and the examiners during which, hopefully, the student is the most active speaker.

Next you have the project model, which has become more and more common the last years. This model gives a 48-hour preparation period to the students. The students get information about which subject they are to have an examination in 48 hours before the examination, and at the same time they are given a subject or a theme to work with for the examination. The students are then to prepare a presentation on the given subject for the examination day. During the 48-hour preparation period the students are allowed to use all sources they wish. They can receive help from friends and family, and do whatever they feel necessary to prepare a best possible presentation for the exam. The topic can be given as a general topic so that the students have to make their own problem formulation, or it can be given as a concrete problem formulation that the students have to answer in their presentation. On the examination day the examination most commonly comprises two parts; first the students give their prepared presentation. Afterwards this is followed up by a conversation between the student and the examiners about the topic the student has given a presentation on. It may also include other topics that have been gone through during the school year. As mentioned, these are the models most commonly used with some variations when it comes to details. Both of the models are in some cases combined with a listening test.

My findings show that out of the 16 informants I have talked with, 12 told me that they use the project model where the students are given a 48-hour preparation period to prepare a presentation on a given subject. I will come back to variations within this model below. In comparison, only two informants told me that they use the question model with only a 30-minute preparation period, in which the students are to prepare for being questioned on a given topic. Furthermore, in six schools they use a listening test in combination with either the project model or the question model. I will also come back to how they do this later.

In the following I will present how the two models, the project model and the question model, are being conducted, since there are differences also within the schools that use the same models. I will start by looking at the schools that use the project model before I look at the schools that use the question model. In the end I will present how some schools use a listening test in combination with either of the two models.

5.2.1 The schools that use the project model

As I said above, 12 of my 16 informants practice an oral examination with a 48-hour preparation period. However, there are variations in details among these 12. One of the things that vary is how long the prepared presentation should be. My findings show there are four different approaches to this. The most common is approximately 15 minutes for the students' presentation. This is common in the schools where they have a two-part examination. In the first part the students give their presentation and in the second part they have a follow up conversation with the examiners. However, one informant, Lars told me of a second approach that gives the students only 10 minutes for their presentation. Lars also uses a listening test on the oral examination, and this results in the prepared presentation being shorter since they also need time to discuss and ask questions to the listening test. Furthermore, this informant also said that the examiner asks the students questions concerning a movie or a book they have seen or read during the school year. But, they do not have a conversation that follows up the presentation the students made. This is interesting since this makes it possible that the students' knowledge about the topic for the presentation can be limited to the information they present without the examiners noticing this. The third approach to how long the prepared presentations was presented by Lars, another informant, who told me that the prepared presentation in their school is shorter than in most other schools. In this school the students are given a listening test 30 minutes before the examination, and the examination then starts with approximately 10 minutes of questions to the listening test. After that the students give a 5-minute introduction to a subject they have been given 48 hours earlier. The remaining 15 minutes are devoted to a conversation between the student and the examiners based on the introduction the student has given. They can also discuss other aspects of the topic the introduction is based on. The fourth approach was a three-part examination with a listening test, a prepared presentation and a follow up conversation, where each part is given approximately 10 minutes each. All the other informants who use the project model told me that they use the first approach, in which they have a two-part examination where the first half is the students' presentation and the next half is a following up conversation.

Another aspect that varies among the 12 schools that use the project model, is how the students get the subject or topic for their preparation period. I found that this is practiced in four different ways. The first way is that the students are only given a topic based on one of the competency aims from the subject curriculum, and are expected to make their own problem formulation. For example Birgitte said: “My students choose their own problem formulation to the given topic, but the problem formulations have to be approved by me before the exam”. The second way to do this is to let the students draw a problem formulation from a selection of eight. The third way is to make it possible for the students to choose from a list of problem formulations or make one themselves. The last method is to give the students two possible tasks to choose between. The above presented ways show that there are a number of ways that the students are given a theme for their preparation period. I find this interesting because I think this measures different competences that the students have. I will come back to this in the discussion. It is also necessary to mention that all informants did not provide information about this, so there might be even more ways in which this is done.

Above I have presented a selection of the differences between the schools that have chosen the project model. As we see there are differences that at first glance might not seem so big, but that at a closer examination reveals differences in detail that may be very important as to what competences are actually tested in the examination. In the following I will present the schools that use the question model for the oral examination.

5.2.2 The schools that use the question model

As mentioned, only two of the 16 informants told me that they use the question model. Both of these informants combine the question model with a listening test. The first of the two, Liv, lets the students draw a topic from a selection of topics based on the curriculum. The students are then given 30 minutes to prepare for this topic. During the examination the students are first given a listening test that they have to answer questions to. After this they are asked questions about the topic they have drawn. This topic includes a task that contains elements from both literature and culture. In addition to this the students will also be asked about their topic of in depth research[4].

[4] This is what in Norwegian is called *fordypningsemne*, and is compulsory at the English VG1 level

The other informant who uses the question model, Ellen, does not do this by own choice, because the county administration gives guidelines where both models are presented, and has decided on the question model for language subjects. In this school the student are given a listening test to prepare for 30 minutes before the examination starts. During the examination they discuss and are asked questions about the listening test by the examiner. In addition the students are asked questions about a poem, novel or short story they have read during the school year. It is, however, not decided whether or not the students will be told during the 30 minutes preparation which literary work they will get questions about.

Above I have presented the two schools that use the question model for the oral examination in English. It is interesting that it is such a low percentage of the schools that use this model, since this was much more dominant earlier. In the following section I will take a closer look at the schools that use a listening test in the examination.

5.2.3 The use of listening tests

When it comes to the six schools that use a listening test I have presented some of them above. Five of the informants all use a listening test in the ways described above. However, one informant, Mari, told me about a completely different way of using a listening test than the other schools did. In this school the examination starts with a listening test in a group in addition to the prepared presentation and the follow up conversation. The students listen to a text in groups and after that they are to have a conversation together about what they have listened to. Mari said: “I include this [a listening test in groups] in order to test the students' ability to make informal conversation about an unfamiliar topic. This way we can see how well the students manage to communicate in groups not only in a dialogue or monologue”. With this kind of listening test it is possible to test how well they master turn-taking, discussion and other conversational skills. This test format includes and assesses elements that none of the other examination formats include, and according to Mari, this examination format might be more difficult than what other schools practice.

5.2.4 Schools where it is possible to choose between the two models

Above I have presented how the project model and the question model are conducted in practice and it is found that there are differences in detail. I addition to this, I will here present two cases that have a different solution to the choice of examination format than the other schools. One informant, Petter, told me that in his school the teachers have the possibility to choose between the two models, but the students are not given the possibility to choose. However, he thinks 90 % of the teachers currently choose the project model. Furthermore, he told me that the county administration specifies that if you want to have a listening test on the examination you have to use the question model, because they do not want a listening test combined with the project model. However, Petter also told me that he thinks there are teachers who combine the project model with a listening test even though the county administration has decided differently. Another informant, Silje, told me that they have chosen a quite different method than the other schools. In this school the teacher has a meeting with the students at the beginning of the school year. In this meeting they discuss whether they want the project model or the question model. This is decided democratically, and the teacher writes a contract with the class representative in which they agree on which model to use. This year the students have chosen the project model. This is an interesting method since it gives all the power to the students and the teacher has to adapt to what the students want. As I will come back to in section 5.4 , there are several counties where the county administration gives the schools the option between the two models, but there is only in this school that they have chosen to let the students be in control of this.

Above I have gone through my findings concerning structure and format of the examinations. As mentioned, I have found that there are two commonly used models, the project model and the question model, but that within these models there are a number of interesting differences between the schools I have studied. In the following I will move on to the next theme, which is what is tested in the examination.

5.3 What is tested in the examination?

The second question I asked all my informants was; what is tested in the exam? When asking this question I wanted to find out what the teachers actually assess in an English oral exam. I wanted to find out whether it was language skills or content knowledge that was considered to be most important. In the following I will go through what I found out and give examples of what the informants told me.

Out of my 16 informants, 14 said that it is the competency aims from the subject curriculum that are assessed. The two teachers who did not mention the competency aims explicitly, instead referring to competences that are actually a part of the subject curriculum and the competency aims. The main finding is that it is obvious that the competency aims are important to all the teachers, and that is important for oral examinations in Norway. Many of the informants give examples of what is assessed, in the following I will present some of these examples.

To start with, one informant, Birgitte, told me that the students have to show that they can go in depth on the topic and show that they are capable of using sources in a critical and independent way. Furthermore, the students have to show that they can relate texts they have worked with earlier in the school year to the topic they present on the oral examination. Next they have to show that they are able to use the knowledge they have in a relevant manner for the exam. Birgitte also mentioned that it is important that the students are able to give their own opinions. To sum it up, she feels that knowledge, understanding and language skills are tested in the oral exam, and that they all should count the same when assessing. Several teachers mention the same elements.

One of the informants who used a listening test in the examination, said that this makes it possible to test both how well the students can communicate about topics they have gone through earlier, but also how well they can discuss new information and material through the listening test.

Several of the informants give examples of language skills are tested in the exam. For example Silje said: “The students have to be able to use an advanced and varied vocabulary and it is important that they use correct grammar. I also think it is an advantage if the students speak either British English or American English, rather than mixing them.” Silje is the only teacher that mentions this aspect. In addition she pays especially attention to the use of voiced s/z, to diphthongs, and to other phonological elements that can help place the students on the correct level in language knowledge and skills. Other informants also mention that vocabulary, fluency and grammar are important, while one informant emphasized the importance of intonation and pronunciation and said that it is also tested how well the students are able to listen and have a conversation in English as well as the skill of giving a presentation.

In addition to language skills, many informants mentioned content, reflection and communication skills as aspects that are tested in the exam. Mari said: “I think it is very important that both informal and formal communication skills are tested in the exam”. To ensure this she uses the listening test in groups in order to test their informal communication skills. Only one informant said that they also test how well made the presentation is made. I understand this the way that they evaluate the students' digital skills in the oral exam, and it is interesting that they do so, since the Directorate of Education has pointed out that it is the students' performance in the examination that is to be tested, not what they have done in the preparation period. Finally, several of the informants mentioned that they also assess how well the students solve their problem formulation.

Above I have given a selection of examples of what the informants said is assessed in the examination, and as we see there are several elements that according to the informants are assessed. Consequently, I think that what the above findings show is that it might not be completely clear to all parties what should be assessed in the examination. When the informants present so many things that should be assessed in the examination, it to me does not seem easy for the students to know what they are tested in. Some of the informants mentioned that the competency aims that are going to be tested are given to the students together with the topic for the preparation period, while others do not mention this. I would think that this is the case in more schools, but maybe not all. The fact that the informants mentioned so many different elements that are assessed in the examination shows that there might be confusion even among the teachers as to what should be assessed in an oral examination. My question is therefore, how should the students understand what they are tested in if the teachers do not really know? This is an interesting point that I will come back to in the discussion. In the following I will go through the informants' answers concerning who is responsible for making the examination.

5.4 Who is responsible for making the examination?

In this part I will go through the answers to my third question to my informants, which was; Who is responsible for making the examination? Is it made locally or has the county administration given any instructions for it? This was a question that was to find out whether the county administrations have handed all of the responsibility over to the local schools, or if they have kept some of the control over the oral exam.

Out of my 16 informants, all told me that it was the school that was responsible for making the tasks for each examination. However, there were differences as to how much control the counties have. Five of my informants did not mention the county administration at all when answering this question, and I interpret this as the responsibility for the oral examination in these five counties being given completely to the local schools and teachers. One of the informants, Tor, explicitly said this when he said that there are no guidelines given from the county administration. Furthermore, another two informants said that the examination is made in cooperation between the local teacher and the external examiner. For example, Lars said: "We send the tasks, the reading lists and the assessment criteria to be approved by the external examiner at least four working days before the examination." The remaining 11 informants told me about an examination that is made locally by the teachers, but with varying degrees of involvement from the county administration. In the following I will present some examples.

One of my informants, Birgitte, mentioned that the framework for the oral examination is given by the county administration, but the local teacher develops the examination tasks which is again approved by the external examiner. However, the county administration wants the schools to have as similar examinations as possible, and therefore the oral examination is discussed by county's forum for the English subject to make sure that it is as similar as possible in the different schools in the county. In the forum there is a representative from each school, and they have several meetings every year. All the schools' principals report to the forum what they want to be discussed in the meetings. I expect that the teachers of English are supposed to tell the principle in their school if they have points that they want to be discussed in the forum meetings. This is one of the cases where the examination is the most regulated by the county administration.

Another two informants told me that the county administrations in their counties have made two models that the schools can choose between for the oral examination. However, in these schools it is the teachers who make the examination tasks which are sent to the external examiner for comments and approval. The reason why it is like this, the informants said, is because the oral examination should be locally given according to the English subject curriculum. One of these informants, Ellen, said: "I think it is very important to make sure that the students are tested in material that they have actually been taught in school [...] the written exam becomes more and more general and I think that should be avoided in the oral exam”. As she points out, the tasks for the oral examination can be adapted to the material that has been taught when they are made locally, rather than asking general questions based on the curriculum.

Another seven informants all told me that there are guidelines given from the county administration on how the oral examination should be, while one informant told me that there are no such guidelines at this point, but that there has been done much work to try to make a common format template for the whole county. One of the seven informants I mentioned, said that the guidelines given from the county administration among other things present different models that can be used for the oral examination, while another of the informants said that the guidelines are very general and only says that the students should be given some preparation time, and that they should be able to choose between two given tasks connected to English language or literature. A third of these informants told me that the guidelines are discussed in the county's network for the English subject, and also in the English sections at the local schools. This case is the one that is the closest to what Birgitte told me about, as I presented above.

To sum up, I have provided a number of examples of what the informants told me about who is responsible for making the oral examination. The findings show that, to varying degrees, most of the county administrations have chosen to exert some degree of control over the oral examination, while a minority has given all of the responsibility to the local schools and teachers. This level of autonomy may well be a question of available resources within the counties, since it will certainly take both time and money to make guidelines and to have subject forums that are responsible for the examination. There might also be plans for making guidelines or common templates within the counties, but I have not received any information about this. The findings also show me that across the country there are many differences as to how the county administrations have chosen to administer their responsibility for the oral examinations, and that to a large extent the local schools are given much freedom to conduct the oral examinations as they wish, also in the counties where guidelines are available. As I will come back to in the discussion, this might be in accordance with the English subject curriculum, even if it does not ensure that the students are tested in the same way across the country. In the next section I will move on to my findings concerning assessment of the oral exams.

5.5 Assessment

In the process of deciding which questions I was going to ask in the interviews,I found it very important to find out how the students are assessed in the oral examination, and it especially interesting to find out whether or not there were used assessment criteria when assessing the students' performances. To find out about this I asked the informants these questions; How is the examination assessed? Who assesses? Do you have assessment criteria that have to be used by all examiners? In the following I will go through my findings and give examples of what the informants answered.

First of all, 12 of the informants said that the assessment is done in cooperation between the local and the external examiner, but that it is the external examiner who has the last word and decides the final grade, while two informants said that it is the external examiner alone who assesses the examination. Finally, there are two cases where the informants have not provided any information about this. When I further look at how the examination is assessed, seven of the informants said that there are not any common assessment criteria that are to be used on the examination, while the remaining nine informants told me about common assessment criteria being used. In the schools that use common assessment criteria, however, it varied whether these criteria were locally made or provided by the county administration. I will first look at the cases where there are no available assessment criteria.

5.5.1 Schools with no common assessment criteria

All in all, seven of my informants did not use common assessment criteria on the examination. One of them, Tor, said: “We don't use common assessment criteria, but I think it should be spent time on making them, because today it is mostly up the external examiner what grade the student gets.” Apart from this he said that it is the Act of Education and it's regulations that together with the competence aims from the English subject curriculum, are the starting point for the assessment. Another informant said that they do not have common assessment criteria and that the examiners assess based on experience and subjective opinion. Among the informants that said that no common assessment criteria are used, one said that the county's forum for the English subject is supposed to make a form with assessment criteria some time this term. For now, it is the competence aims that make the basis for assessment. One informant said that because they do not use any common assessment criteria the important thing is that assessment of the candidates is fair. However, this informant did not say anything about how this is ensured when there are no assessment criteria to be used. On the other hand, one of the informants told me that even if there are no common assessment criteria, there are network courses designed to make assessment across the county more coordinated. My informant, Petter, said that the tasks and the competence aims that the students are given for preparation also service as assessment criteria together with a description of student skills on different levels, these are made on the basis of the assessment criteria the Directorate of Education has made for the written examination. He said: “I don't think it is possible to have common assessment criteria, since the exam tasks are so different.” He feels the examinations have to be assessed one by one because they are so different.

In other words, there are different ways to solve the question of assessment in the counties that do not have common assessment criteria. However, it would seem that most counties want the students to be assessed in approximately the same way in all schools within the county, even if they have not made common assessment criteria to ensure this. In other counties this is ensured by assessment criteria made by the county administration.

5.5.2 Schools that use common assessment criteria

As mentioned above, it was found that nine of my informants said that common assessment criteria were used to assess the students' performances in the oral examination. In the following I will give examples of these. I found that some of the schools use locally made assessment criteria and some use assessment criteria made by the county administration. Five of the nine informants told me about assessments criteria that were provided by the county administrations. For example, Tonje said: “An individual assessment is the basis for assessment, but the county's network for the English subject has made competence aim descriptions that are supposed to be used for the oral exam.” Furthermore, another informant said that the county's network group for the English subject has made assessment criteria that have to be used by the examiners, while the other three informants only mentioned that it is the county administration that has provided the assessment criteria that are being used. On the other hand, four informants told me that the assessment criteria that are used for the oral examination are made locally. One of these four said that they use an assessment form for oral presentations that is used by all teachers in the school, but there are not common criteria for the whole county. Kari said that there are locally made assessment criteria in addition to the competence aims being the basis for assessment. Furthermore, she said: “these [the assessment criteria] are locally made, but are based on statements from the county administration that the presentation should be clear and systematic, and that the grade should be built on the students' competence in English”. Yet another informant, Helga, told me that currently the school uses locally made assessment criteria, but that there has been done much work to make common criteria for the entire county, however, this work has been stopped by the county administration for unknown reasons. Finally, Mari mentioned that their school uses locally made assessment criteria that focus especially on pronunciation, vocabulary, grammar an correct use of formal and informal language. These criteria are used by both examiners and are handed out to the students as well.

5.5.3 Summing up

As we have seen above there are extensive variations when it comes to the use of assessment criteria for the oral examinations. Some counties have criteria that are to be used by all schools, while in other counties the schools can make the assessment criteria themselves. On the other hand, there are schools where they do not use any common criteria, and leave the assessment entirely to the examiners. Some of the informants who report that there are no common assessment criteria, express a wish for these to be made, while others do not. In the following I will look at another question I asked my informants concerning assessment.

5.6 The five most important criteria when assessing an oral exam

In the following I will go through what was originally the last question I asked my informants. The reason why I choose to do it in this order is because it is so closely connected to the question I discussed above. I asked the teachers to do this; “If you are to make a prioritized list, what are the five most important elements you look for in a candidate at the oral examination?” This was the question I added after having written the pilot study because I found that it would be very interesting to find out what the informants find important when they are to assess a student at the oral examination. As you will see, the variation in the answers is large, and for that reason I have chosen to present the answers in a table. In the table you see the fictitious names of the teachers and the answers they gave me.

Table 4: The five most important criteria when assessing an oral examination

Teacher	The five most important criteria when assessing an oral exam
Elisabeth	- The student knows his material well and speaks independent of manuscript - The student goes more in depth than the textbooks, uses other sources -There is a clear connection in the presented material - The student is able to answer direct questions beyond what has been gone through, shows the ability to reflect - The student uses the time at his disposal well
Birgitte	- The general impression - The ability to express oneself with fluency and coherence - Knowledge/understanding/reflection - Relevant vocabulary - Correct grammar
Tor	- The general impression - The student is willing to speak, takes initiative - Vocabulary - Good pronunciation - Good intonation - Grammar
Ellen	- Ability to communicate, use advanced language and correct grammar - Ability to speak about known material in a good way - Ability to give his own interpretations - Ability to speak about the unknown listening test in a satisfying manner
Silje	- Knowledge about the subject - Grammar - Interdisciplinary initiative - Pronunciation and vocabulary - Level of communication
Tonje	- Communicative competence - Language competence - Subject competence - Structure
Liv	- Independence - Structure - Insight into the subject - Language - Will to show what he knows
Karine	Difficult to make a prioritized list. Since both language and knowledge of subject are very important, it is impossible to range one above another. However, important elements are: idiomatic language, being independent of the manuscript, using the presentation to back up what the student says.
Mari	- Fluency in language and communication - Relevant content connected to the competence aim from the curriculum that is tested - Independent discussion of the examination question - Correct pronunciation - Language accuracy
Helga	It is difficult to prioritize because there are many things that are tested in the exam. The competence aims of the curriculum is the basis for my assessment and usually I do not range them after importance
Ingrid	- Synthesis between language and content - Structure - Fluency - Variation - Correct grammar
Kari	- The candidate answers and argues well, is convincing and shows knowledge

	- The candidate masters a nuanced and relevant vocabulary - The candidate is able to relate the topic to other topics gone through in school - The topic is presented in a well articulated and clear manner - The candidate is thoughtful, keeps eye contact with examiners, independent of manuscript and masters grammar expected at this level
Erik	- Whether the student has answered the problem formulation or not - Ability to express oneself - Ability to have a conversation - Pronunciation, vocabulary, intonation - Ability to answer questions from examiners
Lars	Refers to the assessment criteria which among other elements mention: pronunciation, intonation, idiomatic language, being able to participate in oral interaction, produce language adapted to the situation and topic, being able to analyze and discuss the knowledge in context.
Hanne	- Ability to communicate - To what degree the student communicates with the examiners - To what the degree the student gives a precise and thorough account for the topic - Relevant and precise vocabulary - Pronunciation and language The general impression is the most important
Petter	- The general impression - Correct language; grammar - Pronunciation and intonation - Topical knowledge - The student takes initiative in the conversation, not just waits for questions

As displayed in the table there is great variation in the informants' answers of this question. Many informants present five elements that in reality comprise more than five elements. This may indicate that it was difficult for the informants to choose five. The reason for this might be that, as many of the informants mentioned, they find the general impression so important. However, I found that there are elements that are commonly mentioned by the informants. In the following I will present the most frequently mentioned elements in five categories that I have called; language competence, communicative competence, subject competence, the ability to reflect and discuss independently, and the ability to speak freely and independent of the manuscript.

First there are the elements that are a part of language competence such as grammar, pronunciation, intonation and vocabulary. These are elements that all reflect the language skills the students have. Next, it is frequently mentioned that the ability to communicate or good communication skills, the ability to have a conversation and communicate with the examiners are important elements. These elements can all be summed up in the term communicative competence. A third category that is mentioned in various ways is subject competence. For example it is mentioned that it is important that the students show knowledge about their topic, that the students give answer to the problem formulation, that the content is relevant and that the students have insight to his topic. The fourth category that is often mentioned, is how well the students are able to reflect and discuss the topic in an independent way. Last, the fifth category is that the students speak freely and independent of manuscript. The five categories of elements that the teachers mostly mention as important when assessing an oral examination, not in prioritized order, are:

- Language competence
- Communicative competence
- Subject competence
- Ability to reflect and discuss independently
- Ability to speak freely and independent of manuscript

Even if these are the most frequently mentioned elements, I also want to give examples that are not so common.

One example is that there is only one of the informants who mentioned interdisciplinary initiative. It is interesting because this is not mentioned in the Act of Education or its regulations for the oral examinations. In addition to interdisciplinary initiative, this informant also mentions the more common elements such as subject knowledge, grammar, vocabulary, pronunciation and communication skills. Another element that is mentioned by only one informant, is that the students have used more sources than just the textbook. If we look at the competence aims of the curriculum it is found that it explicitly says that the student should be able to select and use content from different sources, independently, critically and responsibly, so in that manner it is interesting that not more informants mention this. However, as I have mentioned earlier, it has been pointed out by the Directorate of Education, that it is only the students' performance during the examination that is to be assessed, and not their work in the preparation period. Even so, I expect that more of my informants find this important, but that it did not find its way to the top five elements. Moreover, there is only one informant who uses the categories I have used to sum up which elements the teachers find important.

The examples I have given above are of course only a selection of what my informants have said, and there are more that could have been mentioned. However, I feel the most important are included here and because you find all the elements in the table, these are the ones I choose to mention here.

To sum up, the answers to this question show that, even though the teachers mention various elements that are important for them when assessing an oral exam, the differences may not be so big after all, since many of the elements can be categorized the way I did above. In the following I will go through the last two questions that I asked my informants, which reveal some of the teachers' opinions about the oral examination and the system of the oral examination.

5.7 Teachers' opinion: Is it a fair exam

In the following I will go through the question that concerns whether or not the informants think the examination is fair for the students. In this question I wanted to find out if the informants were really conscious about the possibilities of making a fairer oral examination for the students than the one we have today, or if they are satisfied with the way it is now. What I found was that most of my informants did not interpret the question as I did, since most of them answered according to how the examination in their school is, and not at a national level. This is understandable, and the answers to that question are also interesting. That is why I did not ask any more about the exam at a national level to most of the informants. It is very interesting to note that not one of the informants said that the examination is not fair. Indeed, all informants are rather confident that the examination in their school is fair. There are, however, some informants who mention how the examination could have been made fairer, and there are some informants that mention the problem of fairness at a national level. In the following I will give examples of their reasons for saying so.

I will start by giving examples of the reasons the informants who think the oral examination is fair have given for their point of view. Among the informants who use the project model, one of the informants said that the examination is fair because the students get 48 hours to prepare for the exam, and then they have the possibility to get guidance from their teacher and others, to use all available sources, and they have all possibilities to prepare to get the best grade possible in the exam. On the other hand, Birgitte, said: “I was at first skeptical to the project model because it is possible for the students to get help from others. But, my experiences have shown that this works well. […] I think that the students who have the chance to get much help home will have an advantage no matter which format is used […] I have not experienced any big differences between the quality of the presentation and the rest of the exam”. This is of course interesting, because I have heard others speak about this as a problem a long time before I started writing my thesis. They have said that they have experienced students who have given a brilliant presentation in all ways, but when they are to interact with the examiners and answer unprepared questions they almost fall apart. This can be proof that they have received so much help with their presentation that it is not representative of their level, or it can of course be a result of nervousness.

However, Birgitte is not the only one that feels that this is not a problem. For example Tonje said: “You might think it is unfair that the students can receive help to prepare the presentation, but they are well taught in how to do this, and it is the last part of the exam that counts the most if there are differences in quality”. On the other hand, Erik thinks that because of the differences in how much help the students receive the examination is only fair to a certain extent. Another of the informants who use the project model argues that the examination is fair because the students have the chance to show much of their skills and competences with this examination model. However, this informant felt that the old examination model tested more than the one they use today. I expect that this school earlier used the question model, and that this is what the informants referred to. Finally, Kari finds that the examination is fair because they assess all students according to the same assessment criteria and the same competence aims.

Furthermore, the two informants who use the question model are confident that the examination in their school is fair. One of them thinks this model is fairer than the project model, because of the fact that the students cannot access the same resources in the preparation period in the project model. This informant also finds the examination fair because it tests several elements when they have both a listening test and a questioning. The other informant who uses the project model, Ellen, is also confident that the examination is fair and said: “I especially think the combination of a locally given oral exam and a centrally given written exam is very good” and that this helps ensure fair examinations for the students.

Above I showed examples of informants who felt that the examination is fair, and as we see there are several reasons why they feel it is a fair examination. Most of these informants have not considered the problem of fairness at a national level, or at least they did not mention it in their answers. In the following I will give examples of the informants who have the opinion that the exam could have been made more fair, and the teachers that mention the problem with fairness at local and national level.

One of the informants that feel the examination is only fair to a certain extent, Tor, said that the reason for this is that even though the students are tested according to the same competence aims, he feels that there should be assessment criteria that all examiners use and have knowledge about. Furthermore, he said that there probably should be centrally given guidelines for the oral examination. At the same time he points out that most students get better results in the oral than in the written examination. In my opinion this does not necessarily mean that the oral examination is fair, but it is interesting that he mentions it. Another informant who also feels that the examination is fair to a certain extent, Silje, said: “I often have students who are stronger written than orally, and then the oral exam does not reflect their real level in English”. Because of this, she questions the fairness of the examination. I, however, suspect that this is not a fairness problem that could have been solved by changing the format of the examination in any way, but rather a problem that should be solved in the English classroom.

There are three informants that mentioned the problem of fairness at both a local and national level. Mari said that the examination is fair within the school, but not at a county or national level because of the big differences in how the schools conduct and assess the oral examinations. She feels that the examination format they use in their school may be more difficult than what other schools do, and this is not fair for the students. Another of these three informants, Helga, said that the exam is fair within the school, but not at a national level. Petter, on the other hand, said: “I don't think the differences across the country are so big that it creates a fairness problem. If you make that a problem you could also say that the final grades are not fair either, because there will always be differences between teachers […] Maybe I am more strict than you for example”. However, this informant thinks there is something else that may cause a fairness issue, and that is the draw of the subject you are to have an examination in. In his opinion it is not fair that the students have examinations in different subjects. It is interesting to see that the informants who have mentioned the possible differences at a national and local level do not agree. This is an issue I will look closer at in the discussion.

Above I have gone through the question of whether the informants think the oral examination is fair for the students or not. This showed me that most of my informants find the examination in their school fair, however, some questioned if there could have been done more to ensure a fair examination and if it is really fair at a national level. It is interesting that not more of the informants mention this as a possible problem, however, it could be because of the way they interpreted my question. In the following I will go through the last question that might also shed some light on this issue.

5.8 Teachers' opinion: Should there be a locally or centrally given oral examination?

The last question I will go through whether my informants think there should be a locally or centrally given oral examination in English. In the previous section I mentioned that there were informants who did not mention any problem of fairness at a national level when asked if they thought the exam was fair or not. However, I thought that the last question might make more of the teachers think about this as a possible problem.

I found that seven of my informants wanted the county to be responsible for the oral examination. Interesting, that was exactly the same number as those who wanted a centrally given oral examination, also seven informants, However, one informant did not have an opinion about this question. There was also one informant that I cannot really place in any of these groups, because this informant thinks that it is important with a template that is common for the entire county, and possibly for the whole country, but she feels that it is very important that the exam tasks are made locally to ensure some pedagogical freedom for the teachers. In the following I will give examples of how the seven teachers who want a locally given oral exam argue for their point of view, and further I will look at the argumentation the other seven teachers give for a centrally given exam.

Two of the seven informants who favor a locally given examination, say that it might be good with national guidelines to make sure that the differences are not too big, however, they do not want the Directorate of Education to control all the details concerning the examination tasks. The reason for this is that they find the curriculum so wide that it would be difficult to give centrally exam tasks that would be covered by all schools across the country. Another informant, Elisabeth, said, “I think that the best solution is that the county administration is responsible for the oral exam to make sure that local differences are taken into account when making the tasks. […] Different schools use different school books and may have focused on different parts of the curriculum even if the competence aims are the same”. Ellen also strongly feels for a locally given oral exam and said: “I remember the last time the Directorate of Education gave a suggestion about the oral exam, with a 48-hour preparation to a given topic, and pointed out that we could not ask questions outside the topic.[...] This was all wrong, and now it is specified that the students are to be tested in more than one competence aim.” Moreover, Ellen feels that there might be a problem in private schools where teachers are assessed, because this might create pressure to give good grades to the students. She suggests that there could be an inspector who comes to schools unannounced as an extra examiner to ensure that the assessment is done correctly. The three remaining informants also favor a locally given oral examination. One of them cannot imagine how a centrally given oral examination could be conducted, while the other is just satisfied the way it is today. Finally, the last of these informants, Petter feels that a locally given examination is the best because then the teachers have the possibility to test the students in what they have taught them. He feels that a centrally given oral exam would destroy the connection there is today between the teacher, the students and the material that has been gone through during the school year.

As we can see the informants who favor a locally given oral examination in some way or another do so because they do not want to lose control over it. They feel that a centrally given examination will take away the possibilities they today have at giving an oral examination that fits the way and what material they teach during the school year. Many of them point out that the material that has been gone through might vary from school to school as a result of the subject curriculum being so wide and general. That would of course mean that an oral examination given by the Directorate of Education would have to be more general in content than the locally given examination could be. Below I will look at some of the reasons the other informants give for wanting a centrally given exam.

Among the seven informants who favor a centrally given oral examination, it is repeated that there is a need to make sure the oral examinations are equal for all students across the country. One of the informants, Hanne, said, that it would be very good with centrally given guidelines for the examination, but that the most important is that the exam is adjusted to the students. It should be possible for them to show their competence and have their competence assessed in the same way, no matter in what school or part of country they are in, or which examiner they have. Furthermore, she feels that centrally given assessment criteria would ensure this, but that it is not possible to make completely objective criteria, and that professional opinions will always be important. All the six other informants who favor a centrally given oral examination, does so because it would ensure equal examinations across the country. Helga said “[...] It is especially wanted to have centrally given assessment criteria to ensure a fair exam. I think there are very big differences between schools and counties, and fear that the reliability of the exam is low considering that the students are competing for entrance to the same higher education with their exam results”. Mari agreed with Helga and said: “[...] the more central the better an exam would be, but the most important is the assessment criteria.” The other informants that favor a centrally given examination give the same reasons as mentioned above, that they want to ensure an exam that is the same across the country because students would then be treated more equally than they are today.

As we see above the informants are split on this issue, and maybe that this is to be expected. It is, however, interesting to see that some of the informants who favor a locally given examination, are still open for, and think that it might be a good solution to have centrally given guidelines to make the differences smaller than they are today. I would therefore say that most of the informants are interested in the students being assessed as equally as possible. However, at the same time there is a fear of losing control over what the students are to be tested in, and about having to administer an examination that you are not familiar with. This is understandable because these teachers have probably experienced this before and know how it feels to have to conduct tests and teach material they are not in control over. At the same time I think there some worry among informants about their not having gone through the same material as all other schools, to a large extent because the curriculum is considered so open. It is because of this they want to ensure that the students are tested in material that ensures that they are able to show their competence. All of this is understandable, but still there are those teachers who look beyond this and want all students across the country to be tested the same way to ensure a more reliable exam for the students. In any case, there are good reasons both for having a locally and a centrally given oral exam, and therefore it becomes a question of what is most important, to ensure that all students are tested equally or to ensure that all students have equal opportunities to show their competence. I will come back to this in my discussion.

5.9 Chapter summary

In this chapter I have presented the findings from my interviews. I have gone through each of the questions I asked my informants one by one presenting and giving examples of what they said. The questions that the teachers were asked concerned these elements:

- structure of the examination
- what is tested in the examination
- who is responsible for making the examination
- how the examination is assessed
- what five elements the informants find most important when assessing
- whether the informants find the exam fair or not
- whether the informants prefer a locally or centrally given oral examination

My findings show that there are differences between schools, counties, and informants in all these elements. In the following I will sum up and present the main findings before I move on to the discussion of my findings in connection with the theory presented in chapters 2 and 3.

Concerning format and structure of the oral examination, it was found that schools structure the oral examination in two different ways, using either the project or the question model. The project model gives the students a 48-hour preparation period to prepare a presentation on a given topic. On the examination day, the students are to give their presentation and this is followed up with a conversation between the student and the examiner where the topic of the presentation is discussed more in detail and other topics from the curriculum may also be discussed. The question model gives the students a 30-minute preparation period to prepare for a given topic. During the examination the student is to answer questions about the topic and elaborate and discuss the given topic. In addition, other topics from the curriculum may also be discussed. Both of the models are sometimes combined with a listening test, either one by one or in groups. It was found that the most common model is the project model, which is used by 12 of my informants. Only two of my informants said that they use the question model. Even so, it is in many counties possible for the schools to use either the project model or the question model, even if most choose the project model. Furthermore, it was also found that there are variations in details among the schools that use the same models. Within the schools that use the project model, it varies how long the prepared presentation the students give is. This varies from five to 15 minutes. It also varies whether the examination is two-parted or three-parted. In the cases where the examination is two-parted, it consists of the students giving their prepared presentation and afterwards this is followed up with a conversation between the student and the examiner. In the cases where the examination is three-parted this most commonly consists of a listening test the students are asked questions to, and afterwards they give their prepared presentation and the examination is ended with a following up conversation between the student and the examiner. Another variation between the schools that use the project is the way the students are given the topic for the preparation period. In some schools the students are given an already made problem formulation, while in other schools they are only given a topic and have to make a problem formulation themselves.

Furthermore, it was found that six schools use a listening test in combination with either the project model or the question model. Five of these give a listening test prior to the examination and during the examination the students are asked questions to the listening test. However, one informant presented an approach to this, quite different than the others. This informant practices a listening test in groups. The students listen to a text in groups and after that they are to have a conversation together about what they have listened to. This is included to test the students' ability to make informal conversation about an unfamiliar topic. I will come back to this in the discussion.

The second question concerned what is tested in the examination. It was found that most of the informants refer to the competence aims from the English subject curriculum in answering this question. However, when they answer more in detail, we can see that the opinions about what elements are important when assessing in an oral exam vary. This might indicate that it is not clear from the competence aims what should be assessed. The findings also seem to show that there is confusion among the informants as to what should be assessed. I have argued that this may cause confusion for the students as well.

Closely connected to the findings presented above are the findings of which five elements the informants find most important when assessing an oral examination. As the findings presented above indicate, there are differences in what the informants find important. However, it was found that the most frequently mentioned elements can be summed up in the five following categories:

- Language competence
- Communicative competence

- Subject competence
- Ability to reflect and discuss independently
- Ability to speak freely and independent of manuscript

The last two categories are self-explanatory, however, the three first need to be explained in some more detail. Language competence comprises elements such as grammar, pronunciation, intonation and vocabulary and are elements that reflect the language skills the students have. Communicative competence comprises elements such as the ability to communicate or good communication skills, the ability to have a conversation and communicate with the examiners are important elements. Subject competence can for example be explained as the importance of the students showing knowledge about their topic, that the students give answer to the problem formulation, that the content is relevant, and that the students have insight to his topic. Even if the most frequently mentioned can be categorized as above, it is important to mention that there are also many elements the informants mention which cannot be categorized in this way. These elements show the variations that might indicate confusion among the informants as to what is to be assessed in an oral examination.

Furthermore, there are also great differences when it comes to whether or not assessment criteria are used in the oral examination. Seven of the informants said that there does not exist any common assessment criteria that are to be used on the examination, while the remaining nine informants told me about common assessment criteria being used. In the schools that use common assessment criteria it, however, varied whether these criteria were locally made or provided by the county administration

In the schools where common assessment criteria are not used, the informants in varying degree say that the assessment is based on the competence aims and the tasks given to the students. Others say that assessment is based on the examiners' experience and personal opinion. Some of these informants express a wish for common assessment criteria to be made, while others explicitly say that common assessment criteria are not possible to make because each examination has to be assessed individually. The examples presented here show that there is great variation when it comes to how oral examinations are assessed in the schools that do not use common assessment criteria. However, there are also variations among the schools that use common assessment criteria. In five counties the assessment criteria are provided by the county administration, while in the four other counties that use assessment criteria, these are made locally. As we see the variations in the use of assessment criteria are extensive and I will in the discussion argue that this may cause issues concerning the reliability of the oral examinations.

Moreover, I asked my informants whether they think the oral examination is fair or not. All my informants are rather confident that the examination they practice is fair for the students. However, a few teachers question the fairness on a national level, even if they think that the examination is fair at a local level. It is interesting that not more of the informants mentioned a possible issue with fairness at a national level, and it may seem that this is something they have not considered at all. However, my findings show that more of the informants are aware of this possible issue when they are asked if they prefer a locally or centrally given oral examination. It is interesting, if not surprising, that the teachers are split in two groups when answering this question. One half wants a locally given exam, while the other half wants a centrally given exam. The informants who favor a locally given oral examination all express a fear of losing control over the examination they are going conduct. A few of them are, however, open for the use of centrally given guidelines for the oral examination, but express that they do not want all details of the examination to be decided at a national level. The informants who favor a centrally given oral examination, argue that this is necessary to ensure that the students are assessed more equally than they are today. These findings at least show that it is necessary to look closer at this question and to discuss what is most important, to ensure equal assessment of students across the country, or to make it possible to adapt the exam according to local differences. I will come back to this in the discussion.

In other words, the findings from my research show that there is considerable variation when it comes to both the format of the oral examinations and how these are assessed. Furthermore, I feel that the findings indicate that these differences might cause issues concerning both construct validity and reliability. This will be discussed in detail in the following chapter.

6 Discussion

6.1 Introduction

In the previous chapter I went through the findings of my project, and in this chapter I will discuss the findings of my project. I start by going back to my research questions.

6.1.1 Revisiting the research questions

My research statement for the present study was: *How are English oral examinations at VG1 level in Norway designed, carried out and assessed across the country?* This constituted 3 research questions that I wanted to look closer at in the study. These questions were:

1) How are oral examinations designed, how is the format?
2) How do teachers assess the oral examinations with regards to assessment criteria and rating scales?
3) Are the examinations and the results of the examinations valid and reliable?

The answer to question one was presented in the previous chapter. I found that there are two examination formats that are the most common across the country; the question model and the project model. The project model gives the student a 48 hour preparation period to prepare a presentation on a given topic. On the examination day, the students are to give their presentation and this is followed up with a conversation between the student and the examiner where the topic of the presentation is discussed more in detail and other topics from the curriculum may also be discussed. The question model gives the students a 30 minute preparation period to prepare for a given topic. During the examination the student is to answer questions about the topic and elaborate and discuss the given topic. In addition other topics from the curriculum may also be discussed. Both of the models are sometimes combined with a listening test, either one by one or in groups. I will come back to these models again when I discuss construct validity in section 6.2.

With regards to research question 2, the results showed that how the examinations were assessed varied more than the formats. First of all, I found that in most cases the assessment is done by the internal and external examiners in cooperation, but if there is disagreement between the two it is the external examiner who decides the final grade. Furthermore, I found that from my sample of 16 teachers, seven of the teachers said that there were no common assessment criteria being used. In these cases, the examiners assessed based on experience and subjective opinion of what is a good performance and not. However, some of these seven teachers reported that common assessment criteria were being developed, but had not been ready to use yet. Only nine teachers reported that they used assessment criteria that were common for both examiners when assessing the performance of the oral examinations. In addition, it varied whether these assessment criteria were locally made or provided by the county administration. I will discuss this more in detail in section 6.3. Furthermore, I discovered that when asked what is tested in the exam, most teachers referred to the competence aims of the subject curriculum. However, when they were going to exemplify this by saying what they find the five most important criteria when assessing in the oral examination, it showed that the answers varied extensively. It did not seem that it was completely clear to the teachers what should be tested and assessed in the examination.

The third research question will hopefully, be answered in the discussion of construct validity and reliability in this chapter. What the results do provide, however, is what the teachers think about whether the examination is fair or not, and this is closely connected to the question of reliability. When asked about the fairness of the examination, all of my informants are confident that the examination they conduct is fair. When interpreting the answers given by informants it is clear that they answer this question thinking only about the examination at a local level. However, there are teachers who mention a possible problem of fairness at a national level. Two teachers mention that there might be a fairness problem at a national level because of the big differences in examination format and assessment practices that exist across the country. This number is, however, surprisingly low in my opinion, but is somewhat altered when we look at the whether the teachers prefer a locally or centrally given oral examination. The answers to this question showed that seven of the informants prefer a locally given oral examination, while seven informants prefer a centrally given one. Two of the informants did not answer this question in way that makes it possible to place them in one group or another. The seven informants who argued for a centrally given examination do so with reference to ensuring more equal examinations across the country, which is one method of ensuring more reliable examinations for the students. Even among the informants who argue for a locally given examination, there are several who are open for centrally given guidelines for the examinations, again because this would help ensure more equal examinations across the country. These findings show that there is uncertainty among some of the teachers about the reliability of the oral examination. This, together with the lack of assessment criteria and the lack of rater courses being given shows that there might be issues concerning reliability of the oral examinations, which I will discuss more in detail in section 6.3. The second part of research question 3 concerns the construct validity of the oral examinations, which I will discuss this in detail in section 6.2. The differences in examination format that I presented above may, however, indicate that there are issues regarding the construct validity as well as reliability.

Based on these indications, I will in the next sections especially discuss the validity of the English VG1 oral examination, connected to research question 1 and 3, and reliability of the examination, connected to research question 2 and 3. Concerning validity I will try to see if the examinations presented by my informants are valid, meaning if they test what they are supposed to test according to the construct provided by the English subject curriculum that I presented in chapter 3. In the discussion of reliability I will try to see if the assessment process the informants presented to me, correspond with the theory about what can be done to ensure a reliable examination. Hopefully this discussion will provide answer to the last research question.

6.2 Construct validity

As presented in chapter 3, construct validity thinking is today based on the work of Messick (1989). He defines construct validity as

> "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment"
>
> (Messick 1989 quoted in Fulcher and Davidson 2007, p. 12).

Furthermore, Fulcher & Davidson say that with this view, construct validity "is not a property of a test or assessment but the degree to which we are justified in making an inference to a construct from a test score" (Fulcher & Davidson 2007, p. 12). As Fulcher & Davidson say there is no absolute answer to the question of construct validity (Fulcher & Davidson 2007).

However, Luoma (2004) gives suggestions as to what can be done to ensure construct validity in speaking tests. She says that "the validity of speaking scores is grounded in the purpose that the scores are intended to serve " (Luoma 2004, p. 185). This means that the first thing one has to do is to specify what is the purpose of the test. Next, it has to be defined what kind of speaking the test is supposed to assess, and what the construct is. When the purpose of the test has been specified and the construct defined, the next step to ensure validity is to show evidence from the test development process that the test actually implements the construct. (Luoma 2004).

The first question at hand is what is the construct to be tested in the English VG1 level oral examination in Norway? According to the Regulations to the Act of Education it is the curriculum that is to be tested in the examination (Forskrift til Opplæringslova § 3-3, §3-17, § 3-25). It also says that as many of the competence aims as possible should be tested in the examination (Forskrift til Opplæringslova § 3-30). Consequently the competence aims from the English subject curriculum is the construct that the oral examination is supposed to test. Based on what Luoma (2004) says must be done to ensure a reliable speaking test, we have to discuss whether the examination formats that have been presented in the results actually implement the construct, which is the competence aims from the subject curriculum.

In chapter 3 I presented the competence aims that make out the construct. They are taken from the main areas *Communication*, *Culture, society and literature* and *Language learning.* Even if I have already presented these in chapter 3, I will here list the competence aims from these areas that constitute the construct to be tested. I find it necessary to do so because the following discussion is so closely connected to these competence aims:

The student shall be able to

- understand and use a wide general vocabulary and an academic vocabulary related to his/her own education programme
- understand oral and written presentations about general and specialised themes related to his/her own education programme
- express him/herself in writing and orally in a varied, differentiated and precise manner, with good progression and coherence
- select and use appropriate reading and listening strategies to locate information in oral and written texts

- select and use appropriate writing and speaking strategies that are adapted to a purpose, situation and genre
- take the initiative to begin, end and keep a conversation going
- discuss social and cultural conditions and values from a number of English-speaking countries
- present and discuss international news topics and current events
- give an account of the use of English as a universal world language
- discuss and elaborate on English texts from a selection of different genres, poems, short stories, novels, films and theatre plays from different epochs and parts of the world
- discuss literature by and about indigenous peoples in the English-speaking world
- describe and evaluate the effects of different verbal forms of expression
- assess and comment on his/her progress in learning English

(Utdanningsdirektoratet 2006/2010)

When we know that this is the construct, the question is whether the examinations that the informants presented to me actually test this construct. This is related to both the format of the examination and the question of what the informants say is tested in the examination and what they think are important criteria to assess in an oral examination. I will start by looking at the format of the examination.

6.2.1 Examination format

What I want to discuss here is whether the format of the examinations ensures that the construct is tested. As I presented in the results there are two formats that are common for the examinations; the project model and the question model. The project model gives the students a 48 hours preparation period to prepare a presentation on a given topic to give on the examination day. The examination usually starts with the students giving their presentation, and this is followed up by a conversation between the student and the examiner where both the presentation and other topics from the curriculum are discussed. The question model gives the students a 30 minutes preparation period to prepare for a given topic. The examination is then a 30 minutes dialogue between the student and the examiner where the student is to answer questions about the given topic. In addition it is possible to combine both these models with a listening test, which is done in varying degree.

The project model comprises, as mentioned above, a 48 hours preparation period where the students are to prepare a presentation on a given topic to give on the examination day. On the examination day the students usually start by giving their presentation, and this is followed up by a conversation between the student and the examiner where both the presentation and other topics from the curriculum are discussed. This model gives the students the possibility to prepare well for the examination and give a good presentation on the given topic. It is of course of interest which topic they are given for the preparation period since this shows which of the competence aims from the main area *Culture, society and literature* the students are tested in. However, my study has not provided information about this so I will not look further at this. The point, however, is that the project model gives good opportunity to test the competence aims concerning culture, society and literature if the topics given are well formulated. When it comes to the competence aims from the main area *Communication* I want to take a closer look at these.

When a student is given a topic to prepare for in 48 hours it is very much possible for him or her to prepare for using both a wide general vocabulary and an academic vocabulary related to the given topic that is, hopefully, connected to his or her education program, in the presentation. The following up conversation will moreover, test if the student masters such vocabulary in regular conversation. The following up conversation will also, to a certain degree, test if the student is able to use speaking strategies adapted to the situation and genre. The examination is a rather formal situation and the students have to show that he or she manages to speak in an appropriate manner for this situation. Moreover, the prepared presentation tests whether the student is able to use speaking strategies appropriate to the presentation genre. When it comes to if the student is able to express him/herself orally in a varied, differentiated and precise manner, with good progression and coherence, I think both the prepared presentation and the following up conversation are good chances to test this.

As we can see I do think the project model covers some of the competence aims that constitute the construct that is to be tested. However, there are competence aims from the area *Communication* that I do not think the project model covers. For example I do not think the project model covers the competence aim that says that the student should be able to understand oral and written presentations about general and specialized themes related to his/her own education program. This might be covered in the preparation, but since the Directorate of Education points out that it is the students' performance at the examination that should be assessed and not the work done in the preparation period (Utdanningsdirektoratet 2010) I do not think this is covered in the project model. Neither do I think the competence aim that says that the student should be able to select and use appropriate reading and listening strategies to locate information in oral and written texts is covered in the project model, based on the same argument as above. Furthermore, I think that the project model with the presentation and the following up conversation does not necessarily cover the competence aim that says that the student should be able to take the initiative to begin, end and keep a conversation going. The reason for this is that the conversation between the student and the examiner is not a normal conversation because it is an examination situation. The nature of this conversation is not the same as a regular conversation, and because of this I argue that the student will not be able to show his or her conversational skills in this kind of conversation. I do think, however, that these three competence aims that I argue are not covered by the project model, can be covered if the project model is combined with for example a listening test. I will come back to this after I have discussed the question model.

The question model gives the students a 30 minutes preparation period to prepare for a given topic. The examination is then a 30 minutes dialogue between the student and the examiner where the student is to answer questions about the given topic. This gives the student much less time to prepare for the examination. As with the project model, which topic the students are given will be important as to which of the competence aims from *Culture, society and literature* that are actually tested. Since I do not have information about this, it cannot be discussed here. However, I do think the question model to a certain extent can cover the competence aims concerning culture, society and literature depending on which topics that are given to the students in the 30 minutes preparation and the questions asked during the examination. At the same time I think it will be difficult for the question model to cover the competence aims that use the verbs *present* and *discuss*. The students can to a certain extent elaborate about topics they are asked about and do some discussion, but I would argue that this is easier to do in the project model.

Moving on to the competence aims from *Communication* I think that the question model gives the opportunity for the students to show if they master a wide general vocabulary and an academic vocabulary connected to their education program, even if they do not have the same chance to prepare vocabulary for a specific academic area. This model, as well as the project model, covers the competence aim that states that the student should be able to express him/herself orally in a varied, differentiated and precise manner, with good progression and coherence, and the competence aim stating that the student should be able to select and use appropriate speaking strategies that are adapted to a purpose, situation and genre. However, I think the question model tests fewer situations and purposes than the project model since it only tests one communication situation, the question-answer conversation between the student and the examiner. This is, as mentioned earlier, a formal situation that requires the student to use speaking strategies that are appropriate to a formal conversation.

As with the project model, there are competence aims from the area *Communication* that I do not think the project model covers. First of all, I do not, as I argued above, think that the question model tests the ability to use appropriate speaking strategies as well as the project model does, because the question model tests less situations and purposes than the project model since it only tests one communication situation. In addition it is again the competence aims that state that the student should be able to understand oral and written presentations about general and specialized themes related to his/her own education program, to select and use appropriate reading and listening strategies to locate information in oral and written texts, and to take the initiative to begin, end and keep a conversation going, that I do not think is covered well enough by the question model. This is based on the same argument as when I discussed these competence aims in relation to the project model. The question model does not test the students' ability to understand oral and written presentations, nor the students' ability to start, end and keep a conversation going.

Even if I do not think the two models cover all the competence aims that constitute the construct, I think it would be possible to make this better by combining these models with for example a listening test. This is done by several of the informants as you can see in the presentation of my results. Giving the students a text to listen to before the examination, and then asking them questions or asking them to explain or discuss what they have heard is a way to cover both the competence aim that says the student should be able to understand oral and written presentations about general and specialized themes related to his/her own education program and the one that says the student should be able to select and use appropriate reading and listening strategies to locate information in oral and written texts. Another opportunity is to give the students an unprepared text related to their educational program and use this in the same way as a listening test. However, you would not then test their listening skills.

Still, there is one competence aim that is not covered, and that is the one saying that the student should be able to start, end and keep a conversation going. This, I think, is actually covered by one of the informants I have talked with. Mari uses a listening test in groups in addition to the prepared presentation and the following up conversation. The students listen to a text in groups and after that they are to have a conversation together about what they have listened to. By doing this you can test the students' conversational skills. It makes it possible to see how well they master turn-taking, discussion, how to start argumentation, how to end a conversation and other elements in a conversation. At the same time as this format covers the competence aim that none of the other formats do, it also gives a better coverage of the competence aim that says the student should be able to use appropriate speaking strategies because this format also tests the students' ability to make informal conversation about an unfamiliar topic in conversation with peers. This is a quite different situation than a conversation with the examiner. Consequently, I think the examination format used by my informant Mari, the project model in combination with a listening test in groups is the examination format that covers the competence aims best. This means the examination format I find ensures construct validity in the best way, is the one used by Mari, the project model in combination with a listening test in groups.

In addition to the discussion above, there is one more element related to the examination format that I want to discuss. It is connected to the project model and concerns the fact that it varies how the students receive the topic for preparation. In some schools the students are given a topic or one of the competency aims from the subject curriculum, and are expected to make their own problem formulation. In other schools the students draw a problem formulation that has been made by the examiner or can choose from a selection of already made problem formulations. I think it is less demanding to make a presentation to a given problem formulation than to have to make one of your own and then start to make the presentation that answers your questions. This is a question of what you want to test your students in, and it might not be fair that it is expected more of some students than others in this aspect. In addition, as I said above, the Directorate of Education states that the students should be assessed based on their performance at the examination and not the work done in the preparation period. Because of this, I would ask whether an already made problem formulation, chosen or drawn from a selection, would be the best way to test what we are supposed to test, since the curriculum does not mention anything of making problem formulations and the Directorate explicitly says that it is the performance at the examination that should be assessed.

To sum up I think the most valid examination format, among the ones presented in the results section, is the project model in which the students are given an already made problem formulation to make a presentation for, combined with a listening test in groups. In the following I will look at another question related to the construct validity of the oral examination. That is the question of what the informants say is tested and what they think are important criteria to assess in an oral examination.

6.2.2 What is the content being tested in the oral examinations?

One thing is to say that it is the construct, the competence aims that are tested in the exam. Another is to break these down to elements possible for both examiners and students to understand and use. In the results I found that most informants referred to the competence aims when asked what is tested in the exam. However, it was not so clear what was tested when they exemplified this. The results showed me there were many elements that the informants felt were tested in the examination. However, the answers did not always correspond with each other. This made me think that it is not necessarily clear from the competence aims which elements should be tested and not. This again made me think that if it is not clear for the teachers what is tested, how is it then going to be clear for the students? In the following I will look at what my informants said was tested in the examination and what they find important when they assess in an oral examination. I will try to see whether this correspond with the competence aims that is the construct.

As you can see in the result chapter, my informants mention many elements that are connected to language competence, such as vocabulary, grammar, pronunciation, fluency and intonation. If we look at the competence aims you see that except from vocabulary, that has its own competence aim, none of these elements are explicitly mentioned in the competence aims. However, I think most of these are covered in the competence aim that says the student should be able to express him/herself in writing and orally in a varied, differentiated and precise manner, with good progression and coherence. You cannot be able to do this without mastering language skills as grammar, pronunciation, fluency and intonation. This means that the language skills that my informants say are tested, should also according to the competence aims be tested.

Furthermore, the informants commonly mentioned elements connected to communicative competence, such as showing good communication skills, being able to communicate with the examiners and the ability to have a conversation. Communicative knowledge can also be found in the competence aims that say that the student should be able to select and use appropriate speaking strategies that are adapted to a purpose, situation and genre and that the student should be able to take the initiative to begin, end and keep a conversation going. Also the competence aims related to vocabulary and language skills can be relevant for communicative competence. As we see communicative competence is also something that should, according to the competence aims, be tested in the oral examination.

Also the third group of elements that my informants commonly mentioned, subject knowledge, is clearly something that should be tested in the oral examination. Elements mentioned by my informants are among others that the students show knowledge of their topic, if the student gives answer to the problem formulation, that the content is relevant and that the student has insight to his topic. Subject knowledge is found in all the competence aims from the main area *Culture, society and literature* and should therefore be tested in the exam.

The fourth group of elements that the informants mention as important when assessing in the oral examination is the students' ability to reflect and discuss independently. These elements can be found in several of the competence aims in the main area *Culture, society and literature* and should therefore for good reasons be tested in the exam.

The last of the most common elements the informants mentioned as important when assessing in the oral examination, is the ability to speak freely and independent of manuscript. This is not that clearly found in the competence aims, but I would argue that it is covered by the competence aim saying that the student should be able to use appropriate speaking strategies adapted to purpose and situation.

Above I have showed that the elements of what is important when assessing in the oral examination most commonly mentioned by my informants, are covered by the competence aims from the English subject curriculum that constitute the construct for the oral examination. However, there are elements mentioned by informants that cannot be found in the construct for the oral exam. One example is the ability to go in depth in the topic and show that they are capable of using sources in a critical and independent way. The ability to use sources in a critical and independent way is a part of the competence aims, but I do not think it is covered by the competence aims that constitute the construct for the oral examination. The reason for this, as I have already mentioned several times, is that the Directorate of Education says that it is the performance at the examination that is to be assessed, not the preparation. Consequently, the use of sources should not be tested in the oral examination. Another element that is however only mentioned by one of my informants, is how well the presentation is made, which would be an assessment of the students' digital skills, which can also not be tested in the oral examination because it is the performance at the examination that is to be tested. A third example of an element mentioned by informants that is not part of the construct for the oral examination, is interdisciplinary initiative. This is not mentioned in the construct definition and should therefore not be tested in the examination.

To sum up, most of the elements that my informants think are important in the oral examination is also part of the construct definition, and should correctly be tested in the examination. However, there are elements that are mentioned that are not part of the construct, and consequently should not be tested. It can be discussed whether these elements should be a part of the construct, but since my project does not aim at analyzing the construct the examinations are build on, I will not discuss this more in detail. The fact that teachers test elements that are not part of the construct together with the fact that the teachers use so many different formulations to explain what should be tested in the examination shows that it is not clear to the teachers what should be tested in the examination. This is supported by research showing that raters have different understanding of and interpret differently the construct to be measured (Ang-Aw & Goh 2011) and that it might be difficult to operationalize the curriculum goals, and because of this teachers develop and use different assessment criteria (Prøitz and Borgen 2010; Throndsen et al. 2009). Since there is confusion among teachers as to what is going to be tested in the exam, I would think the confusion among students is even bigger.

6.3 Reliability

As presented in chapter 3, the classic definition of reliability was provided by Lado (1961):

> “Does a test yield the same scores one day and the next if there has been no instruction intervening? That is, does the test yield dependable scores in the sense that they will not fluctuate very much so that we may know that the score obtained by a student is pretty close to the score he would obtain if we gave the test again? If it does, the test is reliable”
>
> (Lado 1961, as cited in Fulcher 2010)

Said in other words reliability has to do with the extent to which scores are consistent (Brown & Hudson 2002; Henning 1987; Luoma 2004). This is important, because if test scores are reliable we can rely on them in decision making, but if scores are not reliable, it can have serious consequences for the students (Luoma 2004). Of course, there are factors that may threaten the reliability of a test. These factors are called measurement errors, and can for example variation in administration, quality of a test, differences in test forms, changes in test takers over time, differences in scoring and differences in raters may influence the reliability. When it comes to reliability in speaking tests in particular, Luoma (2004) says that there are three types of reliability that especially relevant to the assessment of speaking; intra-rater reliability, inter-rater reliability and parallel form reliability. Intra-rater reliability is the consistence of one rater’s assessment while inter-rater reliability concern whether or not different raters agree on scores given. Parallel form reliability can be measured if there is more than one test format that is going to test the same construct. Because reliability is so important there becomes a question of how to ensure reliability of tests. She also says that the most common methods to ensure reliability is the use of assessment criteria, rater training, the setting of cut scores and of course to make sure that the raters are aware of what might lower the reliability so that they focus on being consistent in their assessment.

As I said in the introduction of this chapter, I think that the results of my findings indicate that there might be issues concerning the reliability of the oral examinations in Norway. The results of my findings show a lack of the use of assessment criteria together with very few rater training courses being given. Together with argumentation from some of my informants that there is a low reliability, I find that this gives reason to discuss the reliability of the oral examination.

In this section I will therefore take a closer look at whether the examinations presented in the results and the assessment processes presented by the informants ensures a reliable examination or not. I will discuss the reliability of the examinations presented by the informants in connection with the assessment process in section 6.3.1. Reliability will further be part of the discussion of whether there should be a centrally or a locally given oral examination in section 6.3.2.

6.3.1 The assessment process

There are several issues that have to be taken into consideration when looking at the reliability of a speaking test. Here I will look at two of the types of reliability, presented in the theory chapter. Luoma (2004) presents both intra-rater reliability and inter-rater reliability. Intra-rater reliability is that raters agree with themselves, over a period of time, about the ratings they give. Inter-rater reliability means that different raters rate performances similarly. However, they do not have to agree completely even if the ratings should not be very different. These are issues I will look at when discussing the reliability of the assessment process the results of my project has showed.

As presented in the theory chapter, Luoma (2004) says that the most common ways to ensure reliable speaking tests, are the use of assessment criteria and rater training. Rater training can be given in courses to ensure that the raters, or examiners, have the same opinion of which performances should get which scores. This is often done by showing performances that are typical for the scores and further by the course participants rating performances in groups and discussing their scores to reach an agreed score. The courses are often ended with a qualifying test. Such rater training can help ensure both intra-rater reliability and inter-rater reliability in the way that it makes examiners more secure about their own assessment, intra-rater reliability, at the same time as it ensures that examiners have the same opinion of how a performance has to be to get a certain score, inter-rater reliability.

When I look at the results, it is very interesting to see that it is only one informant that mentions there being such courses in her county. There might of course be such courses in other counties even if the informants have not provided information about this. I did not ask specifically about such courses and must therefore take into consideration that the informants may not have provided information about this because I did not ask. However, I think the results indicate that there are not rater training courses in most counties. The reasons for this can be many. A rater training course is clearly both time consuming and economically expensive. The examiners are most of the time teachers who have teacher jobs in schools, and if they are to go to courses there has to be provided substitute teachers to teach their classes while they are gone. This is both expensive and sometimes also practically difficult. There also has to be found people who can conduct the courses and these people have to, in a sense, be experts on speaking assessment. There might not be such people in all counties and the people who have this kind of expertise may not be easy to locate. However, as I mentioned in the introduction, such courses are given for examiners of written exams. This shows that it is possible to conduct such courses. Since Luoma (2004) presents this as one of the most common and useful ways to ensure reliability, I argue that rater training courses should be given to all examiners of oral examination. Because it is the county administrations that are responsible for the oral examinations, these courses should probably be given at a county level.

When Luoma (2004) presents rater training courses, it seems that she takes for granted that there are assessment criteria or rating scales that are to be used when rating the performances showed. When looking at the results of my project we see, however, that there are only nine of the informants that say they use such assessment criteria for the examination. The remaining seven informants said that they do not use common assessment criteria in the assessment of the examinations. Among the nine informants who said that assessment criteria were used, it varied whether these were made locally or provided by the county administration. In other words, to the extent guidelines are available, they may be different from county to county.

This is to me very interesting because I think this is the source of the differences in assessment of oral examinations across the country. When there are no common assessment criteria this gives all the responsibility of the assessment to the examiners, and it leaves much room for subjectivity in the assessment. This again may cause an issue both concerning intra-rater reliability and inter-rater reliability. It may cause issues concerning intra-rater reliability because when the examiner does not use assessment criteria, it is much more difficult to be consistent in the assessment and let subjectivity control the assessment. When assessment criteria are not used, it may also cause issues concerning inter-rater reliability because the examiners do not agree about what is important when assessing in the oral examination. This is shown in the results presented in the previous chapter where we saw that the informants mentioned different elements when asked what is the five most important elements when assessing performances at the oral examination.

Furthermore, it varied among the informants who said that there were assessment criteria used, whether these were locally made or provided by the county administration. In five cases the assessment criteria are provided by the county administration, while in the four remaining cases they are made locally. As I presented in the theory chapters there are different types of rating scales or assessment criteria; holistic, primary trait and multiple-trait (Fulcher & Davidson 2007). As long as it is the county administration that is responsible for the oral examination and the tasks are designed locally, I would argue that a holistic rating scale provided by the county administration is the most useful in this case. A holistic rating scale is often quite general, but gives level descriptions of performance to a construct. Such a rating scale can be provided by the county administration based on the competence aims that constitute the construct. This can be used by all schools in the county. If this is done it would help ensure both the intra-rater reliability and the inter-rater reliability of the oral examination. It is also possible to make a holistic rating scale on a national level. This would make the assessment of oral examinations across the country much more consistent, and it can be used across the country even if it is the county administrations that are responsible for the examinations in their counties.

To sum up, I have discussed the reliability of the oral examinations presented in the results, based on what Luoma (2004) says can be done to ensure a reliable exam. I have argued that there can be given rater training courses and be made assessment criteria, either at a national or county level to help ensure higher reliability. This would help ensure both intra-rater reliability and inter-rater reliability.

6.3.2 A locally or a centrally given oral examination

In this section I will try to discuss a question that I also asked my informants: Should the oral examination be centrally or locally given? Today it is the county administrations that are responsible for the oral examinations. Furthermore, the findings of my project show that the county administrations to a large extent have handed the responsibility for the oral examinations to the local schools and the local teachers, and that the schools have much extensive leeway in deciding how they want to conduct the examinations. This is in accordance with the English subject curriculum which states that the oral examination should be prepared and graded locally. However, I question whether this is the best solution because it may not ensure that students are tested and assessed in the same way across the country. As I have stated earlier the English oral examination is high-stakes for the students (cf Broadfoot 2007) because the results of the examination might have consequences for what higher education the students are accepted to. For this reason I think it is important that the students are ensured that they are tested equally.

Moreover, I asked my informants what they thought about this. As we saw in the presentation of my results in the previous chapter, my informants were split in two when answering this question. Half of the informants argued for a locally given oral examination, while the other half argued for a centrally given one. Most of the teachers who argued for a locally given oral examination, did so because they think the English subject curriculum is so wide and open for differences in what the teachers choose to teach the students. Because the curriculum is so general, it varies what the teachers choose to focus on and it also varies what the text books focus on. Many argue that because of this it is necessary that the oral examination can be adapted to local differences. A centrally given oral examination would of course have to be much more general in content than a locally given exam can be, and it is argued by the informants that this would give the students less opportunity to show their competence.

On the other hand, there are several informants who argue for a centrally given oral examination. Most of these informants argue that a centrally given examination would ensure that the students are assessed equally across the country. Some tell that they fear for the reliability of the exam, considering that the students are competing for entrance to the same higher education with their examination results. The informants who favor a centrally given oral exam argue that this or at least centrally given guidelines and assessment criteria would ensure that the students were assessed the same way no matter where in the country they are. In any case, there are good reasons both for having a locally made or a centrally given oral examination, and as I said when presenting my results, this becomes a question of what is most important, to ensure that all students are tested equally or to ensure that all students have equal opportunities to show their competence.

6.4 Suggestions for what may be done to ensure a more valid and reliable oral examination

As I have presented and discussed above, I find that the findings of my results indicate that there may be issues with regards to both the construct validity and the reliability of the English VG1 level oral examination. There may be issues concerning the construct validity due to large differences in examination format and examination, and there may be issues concerning the reliability due to the lack of the use of assessment criteria and the lack of rater training courses, which results in assessment practices varying to a large extent. I have discussed these issues in sections 6.2 and 6.3. After having seen that there might be issues concerning these important concepts in assessment, it is necessary to take a look at what can be done to ensure a more valid an reliable oral examination.

Considering that the reason for questioning the construct validity is that there are so big differences with regards to the operationalizing of the oral examinations, I wonder whether a good solution might be to ensure that the format of the examination is decided either at a county level or even at a national level. It would of course be necessary to conduct further investigation to find which examination format would test the construct in the best way. Among the examination formats presented by my informants, I have suggested that the project model in which the students are given an already made problem formulation to make a presentation for, combined with a listening test in groups, is the format that best covers the competence aims from the English subject curriculum that constitute the construct. However, there may of course be other examination formats that ensure construct validity in a better way. In any case, I think this is something that needs to be looked into more in detail.

Concerning the reliability of the oral examinations, I have argued that there may be issues concerning this due to the lack of the use of assessment criteria and the lack of rater training courses, which results in assessment practices varying to a large extent. Rater training courses and the use of assessment criteria are the methods that Luoma (2004) say are the most common to ensure reliability of an oral examination. As my results showed, the use of assessment criteria varied much, and only one of my informants reported that there were given rater training courses in her county. A suggestion might be to conduct such rater training courses either on a county level or, preferably at a national level. When it comes to the use of assessment criteria, a suggestion is to ensure that all examiners use assessment criteria. To ensure that students are assessed equally, these could be made by the county or they could be made at a national if you wanted to ensure that students are assessed equally across the county.

The bottom line question, is whether we want a locally or a centrally given oral examination. As my results show, my informants are split in this question. And as I argued above, there are good reasons for both solutions and it becomes a question of what is most important, to ensure that all students are tested equally or to ensure that all students have equal opportunities to show their competence. This question is not easy to answer, however there might be a way to ensure both of these interests being looked after. If there were centrally given guidelines for the oral examination this could ensure that the examination format was the same across the country. Furthermore, there could be centrally given assessment criteria with level descriptions of each grade. These would have to be used by all examiners across the country and would help ensure that all students are assessed equally. To ensure that all students have equal opportunities to show their competence, the exam tasks could be made locally, but be based on the centrally given guidelines and of course the construct provided in the subject curriculum. This would take into account that there are local differences in what is being taught across the country because the English subject curriculum is so general. This is however, just a suggestion that might make all parties satisfied and ensure a more valid and reliable oral examination than the one we have today.

6.5 The validity of the present research project

In the last section of this chapter I would briefly like to return to the issue of the validity of my paper. As described in chapter four this is a small-scale qualitative study with 16 informants. Of course, 16 informants is a rather small sample, and because of that it can put limitations on the external validity of the paper (Johannessen et al. 2006). With such a small sample I cannot draw any decisive conclusions about the English VG1 level oral examinations and how they are conducted and assessed across the country. Because of the fact that there with small qualitative studies always will be limitations as to transferability of the results, I therefore cannot say that my findings are representative for all schools and all teachers in Norway. However, I have interviewed 16 teachers in 16 different schools in 16 different counties. This was done both to ensure a more valid and reliable study, but also because I wanted to look at the differences between counties in particular, because it is the county administrations that are responsible for the exam. Because I have informants from different schools in different counties, it is possible for me to argue that the patterns and tendencies I have found are not just coincidences, but may be patterns and tendencies that are valid for other schools and teachers as well.

On the other hand, my findings are supported by the findings of several studies both nationally and internationally (Galloway et al. 2011; Hægeland et al. 2005; Prøitz and Borgen 2010; Throndsen et al. 2009; Lumley & McNamara 1995; Orr 2002; Ang-Aw & Goh 2011).

The studies by Galloway et al. (2011), Hægeland et al. (2005) and Throndsen et al. (2009) show that many teachers assess performance in exams in a norm-referenced manner, even if this is not consistent with the Regulations to the Act of Education. Moreover, studies by Prøitz and Borgen (2010) and Throndsen et al. (2009) show that it might be difficult for teachers to operationalize the curriculum goals, and because of this, teachers develop and use different assessment criteria. Throndsen et al. (2009) further report that teachers say they find it difficult to describe competence at different levels. In their study, Ang-Aw & Goh (2011) report that raters have different understanding of and interpret differently the construct to be measured. Furthermore, it has been reported that an investigation by the County Governors in 2010 looking at exam practices in upper secondary education in Norway, shows that there is too much variation when it comes to exam formats (Bøhn 2011).

In other words, the above mentioned studies report findings that are comparable with those of my study; that there is a tendency with differences in the format and assessment of oral examinations. This means that there is a need to look more into the system of the oral examinations in Norway to find out if these differences threaten the construct validity and the reliability of the oral examinations.

In this chapter I have discussed the findings of my study in light of my research questions and in the light of relevant theory that I presented in chapters 2 and 3. I have given a short discussion of the validity of my study. In the next chapter I will move on to the conclusion.

7 Conclusion

In this chapter I will start by summing up the findings from my research and look at the main implications of these findings. Next, I will look at and suggest directions for further research in the area of English oral examinations in Norway. Last, I will look at possible implications for the English oral examinations based on the discoveries I have made in my study.

7.1 Have the research questions been answered

In chapter 1 I presented the research statement that I made for my project, which was "*How are English oral examinations at VG1 level in Norway designed, carried out and assessed across the country?"* I would argue that found undue variation concerning both format and assessment of the oral examinations. In addition, I have argued that these differences may cause issues concerning construct validity and reliability of the oral examination. As I will suggest in the following section, there a possible implications of this study. My suggestion is that there is made changes with the examination format that examiners need to be offered

7.2 Further research

As I have already said, this is a small-scale study, and results of it can not be generalized to oral examinations, even if they indicate certain tendencies. Therefore, it would be very interesting and useful to conduct a large-scale survey on the topic to see if my results can be confirmed. I think this would be useful because the oral examination is as important as it is, and because several studies suggest that there are variations both in examination formats and the way examiners assess oral performances.

There are of course several ways a large-scale survey on this topic could be conducted and I would think that both qualitative and quantitative methods would be appropriate. One possibility is to conduct a qualitative study on a much larger scale than I have done and interview teachers from a representative sample of schools from all 19 counties. This can give more in-depth information about the format of the examination. This should be combined with a collection of examination guidelines, assessment criteria and other formal examination specifications from county administrations where this exists. This would provide information about the format of the oral examinations and make it possible to evaluate the construct validity of the oral examinations.

Furthermore, the investigation of the use of assessment criteria and of which criteria examiners use when assessing oral performance at the oral examination, can be done by conducting a large-scale quantitative survey with a questionnaire distributed to a large sample of teachers. This can contribute to find out more about the reliability of the examination results. Moreover, it would be interesting to investigate inter-rater consistency more closely. This could possibly be done by using video-material from authentic oral examinations, and distributing these to a sample of examiners to assess the performances. For this to be valid, the sample of examiners has to be rather large. A comparison of the responses can provide indications of inter-rater reliability.

A method as described above would have the benefits of both a qualitative and a quantitative method, and would be very time- and resource consuming, but I think such a study would provide valuable information about the oral examinations in English and could possibly give good indications about what could be done to improve the oral examinations and the system of oral examinations that we have today.

7.3 Possible implications of my study

The findings of my research suggest that there are many variations concerning both design and assessment of the English oral examinations at the VG1 level. The question is then, what should be done about this.

First of all, I think it is important to ensure that all counties use an examination format that in an appropriate way tests the construct to be tested, which according to the Regulations to the Act of Education, is the competence aims of the English subject curriculum. Furthermore, the Regulations state that as many of the competence aims as possible should be tested in the oral examination. This can be ensured if there are steps taken to use an appropriate examination format.

Secondly, I think it is necessary to ensure that all examiners use the same assessment criteria when assessing performances in the oral examination. This would ensure to a large extent that students received the equal assessment that they are entitled to. If the same assessment criteria are not used, it increases the risk that the assessment will vary from rater to rater. Common assessment criteria would make assessment easier for examiners at the same time as it would help ensure that the results of the examinations are correct, and that they can be trusted to give a true picture of the students' competence in English.

Finally, I think it is necessary to ensure a more reliable assessment by having examiners trained in doing so. It is obvious that if there is uncertainty among examiners as to what is to be assessed, and how they should assess it, there will be variations in the way this is done. Rater training courses can therefore help ensure that the examiners assess oral performance in an examination in as uniform way as possible. This again would contribute to a higher inter-rater reliability and intra-rater reliability.

Above I have suggested three implications of the results of my study. However, I think the most important implication is that there has to be made more research on the topic, to ensure that the necessary steps can be taken to improve English oral examinations. It goes without saying that it would be interesting to investigate oral examinations in other subjects as well.

References

Alderson, C. J., Clapham, C., &Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press

Ang-Aw, H.T. & Goh, C.C. (2011). Understanding discrepancies in Rater Judgement on National-Level Oral Examination Tasks. *RELC Journal*, 42(1), 31-51.

Bachman, L. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition 10*(2), 149-164

Bachman, L. (1990). *Fundamental considerations in Language Testing.* Oxford: Oxford University Press

Bachman, L. (2004). *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press

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

Bakke, M. H. (2010). *Teaching reading in EFL-instruction(1)*, Master's thesis. Oslo: University of Oslo

Broadfoot, P. (2007). *An Introduction To Assessment.* New York: Continuum

Brown, A. (2000). An investigation of the rating process in the IELTs oral interview. *IELTS Research Reports*, 3, 49-84.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing 20*, 1-25

Brown, J. D. (1996) *Testing in Language programs*. Prentice Hall, inc.

Brown, J. D. & Hudson, T. (2002). *Criterion-referenced Language Testing.* Cambridge: Cambridge University Press

Brown, G. & Yule, G. (1983). *Teaching the Spoken Language: an approach based on the analysis of conversational English.* Cambridge: Cambridge University Press

Bygate, M. (1987). *Speaking.* Oxford: Oxford University Press

Bøhn, H. (2011) *Summative assessment of student's oral EFL ability. Validity and reliability considerations in oral exams at the upper secondary level in Norway* (PhD project description). Østfold: Østfold University College

Council of Europe (2001). *Common European Framework of Reference for Languages*: *Learning, Teaching, Assessment.* Strasbourg, France: Author. Retrieved from: http://www.coe.int/t/dg4/linguistic/cadre_en.asp

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125-44.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Faye-Schjøll, L. H. (2009). *Reading in Upper Secondary; what do they read, how is it taught and what are the teachers' attitudes towards the teaching of reading?* Master's thesis. Oslo: University of Oslo

Forskrift til Opplæringslova, Hefte 9 (2006) retrieved from http://lovdata.no/cgi-wift/wiftldles?doc=/app/gratis/www/docroot/for/sf/kd/kd-20060623-0724.html&emne=forskrift%20til%20opplæringslov*&&

Fulcher, G. (1997). Ch. 8: The testing of Speaking a Second Language. In (Ed.) Clapham, C., Corson, D., *Encyclopedia of Language and Education. Language Testing and Assessment* 7 1997, pp. 75-85. Kluwer Academic Publishers

Fulcher, G. (2010). *Practical Language Testing.* London: Hadder Education An Hachette UK Company

Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment an advanced resource book.* London: Routledge

Galloway, T. A., Kirkebøen, L.J., Rønning, M. (2011). *Karakterpraksis i grunnskoler: Sammenheng mellom standpunkt- og eksamenskarakterer.* Oslo-Kongsvinger: Statistics Norway

Hasselgren, A. (1998). *Smallwords and Valid Testing.* PhD thesis. Bergen: University of Bergen

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research.* Cambridge, MA: Newbury House

Hughes, A. (1989). *Testing for Language Teachers,* 1st ed. Cambridge: Cambridge University Press

Hægeland, T., Kirkebøen, L. J., Raum, O., Salvanes, K. J. (2005). Familiebakgrunn, skoleressurser og avgangskarakterer i norsk grunnskole. I *Utdanning 2005 – deltakelse og kompetanse.* Oslo-Kongsvinger: Statistics Norway

Ingram, E. (1968). Attainment and diagnostic testing. In Davies, A. (ed.), *Language Testing Symposium: APsycholinguistic Approach.* Oxford: Oxford University Press

Iwashita, N. (1999). The validity of the paired interview format in oral-performance assessment. In *Melbourne Papers in Language Testing 8*(1), 51-66.

Johannessen, A., Tufte, P.A., Kristoffersen, L. (2005) *Introduksjon til samfunnsvitenskapelig metode.* Oslo: Abstrakt forlag

Lado, R. (1961). *Language Testing.* London: Longman

Lazaraton, A. (1992). The structural organization of a language interview: a conversation analytic perspective, *System 20*, 373-386

Lier, L. van (1989). Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly 23*, 489-503

Lumley, T. (1998). Perceptions of language-trained rater and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes, 17*(4), 347 -67

Lumley, T. & McNamara, T.F. (1995) Rater characteristics and rater bias: implications for training. *Language Testing 12*(1), 54-71.

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

McNamara, T.F. (1996). *Measuring second language performance.* London: Longman.

Nikula, T. (1996). *Pragmatic Force Modifiers: a study in interlanguage pragmatics.* PhD thesis. Jyväskylä : University of Jyväskylä

North, B. (1996/2000). *The Development of a Common Framework Scale of Language Proficiency.* PhD thesis. London: Thames Valley University. Published in 2000 as *The Development of a Common Framework Scale of Language Proficiency.* New York: Peter Lang.

Orr, M. (2002). The FCE speaking test: using rater reports to help interpret test scores. *System, 30*(2), 143-54

Pollitt, A. & Murray, N.L. (1996). What raters really pay attention to. In M. Milanovic & A.J. Kunnan (Eds.), *Studies in Language Testing 3.* Cambridge: Cambridge University Press

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

Ragin, C. (1994). *Construction Social Research.* Thousand Oaks, California: Pine Forge Press

Read, J. (2000) *Assessing Vocabulary.* Cambridge: Cambridge University Press

Regjeringen, Kunnskapsdepartementet. *NOU 18 Vurdering og dokumentasjon av læringsutbytte.* Retrived from: http://www.regjeringen.no/nb/dep/kd/dok/nouer/2003/nou-2003-16/19.html? id=370792

Ruch, G. M. (1924). *The Improvement of the Written Examination*. Chicago: Scott, Foresman and Company

Savignon, S. (1985). Evaluation of communicative competence: the ACTFL provisional proficiency guidelines. *The Modern Language Journal 69*, 129-134

Simensen, A.M. (1998). *Teaching a Foreign Language – Principals an Procedures.* Bergen: Fagbokforlaget

Simensen, A. M. (2010). Fluency: An aim in teaching and a criterion in assessment. *Acta Didactica Norge, 4*(1), Art. 2.

Stokke, K. H., Throndsen, I., Lie, S. & Dale, E. L. (2008). *Evaluering av vurdering for læring. Underveisrapport fra følgeforskningen. Evaluering av modeller for kjennetegn på måloppnåelse i fag.* Retrieved from http://www.udir.no/upload/Rapporter/2008/Evaluering_laring.pdf

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing 18*(3), 275-302

Throndsen, I.S., Hophenbeck, T. N., Lie, S., Dale, E. L. (2009). *Bedre vurdering for læring: Rapport fra "Evaluering av modeller for kjennetegn på måloppnåelse i fag".* Oslo: University of Oslo

Towell, R., Hawkins, R. & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics 17*, 84-119

Utdanningsdirektoratet (2006/2010). *English Subject Curriculum.* Oslo: Author. Retrieved from http://www.udir.no/grep/Laereplan/?laereplanid=1097084

Utdanningsdirektoratet (2010). Individuell vurdering i grunnskolen og videregående opplæring etter forskrift til opplæringsloven kapittel 3. Rundskriv, Udir-1-2010. Retrieved from http://skolenettet.no/nyupload/Vurdering%20for%20laring/Dokumenter/Hva%20er%20vurdering%20for%20laering/Vurdering%20for%201%C3%%A6ring%20i%20forskriften/Udir_1_2010_Individuell_vurdering_i_grunnskolen_og_videregaende_opplaring.pdf

Weigle, S. C. (2002) *Assessing Writing.* Cambridge: Cambridge University Press

Weir, C. (1993) *Understanding and Developing Language Tests.* New York: Prentice Hall.

Wigglesworth, G. (1994). Patterns of rater behavior in the assessment of an oral interaction test. *Australian Review of Applied Linguistics, 17*(2), 77-103.

Appendices

Appendix 1

Intervjuguiden

Spørsmål 1: Hvordan er muntlig eksamen strukturert hos dere? Hvordan gjennomføres den?

Spørsmål 2: Hva testes på eksamen?

Spørsmål 3: Hvem har ansvaret for å lage eksamen? Er det lokalt eller er det gitt noe fra fylkeskommunen? Hvorfor har man valgt å gjøre det på denne måten?

Spørsmål 4: Hvordan vurderes eksamen? Hvem vurderer? Har man vurderingskriterier som er felles for alle sensorer etc?

Spørsmål 5: Din vurdering av eksamen: Er det en rettferdig eksamen?

Spørsmål 6: Din mening: Er det bra at det er fylkeskommunen som har ansvaret for eksamen eller er det ønskelig med en statlig ordning der eksamen gis fra udir slik som skriftlig eksamen?

Spørsmål 7: Hvis du skal sette opp en prioritert liste, hvilke er de 5 viktigste elementene du ser etter hos en kandidat på muntlig eksamen?

Bakgrunnsspørsmål:

- Hvor gammel er du?

- Hva er din faglige bakgrunn/utdannelse?

- Hvor mange år har du vært i læreryrket?

Appendix 2

Table with an overview of informants' answers

Lærer	**Hvordan er muntlig eksamen strukturert hos dere?**	**Hva testes på eksamen?**	**Hvem har ansvaret for å lage eksamen? Lokalt gitt eller noe fra fylkeskommunen? Hvorfor?**	**Hvordan vurderes eksamen? Hvem vurderer? Har man felles vurderingskriterier for alle sensorer?**	**Din vurdering: Er det en rettferdig eksamen?**	**Din mening: Er det bra at det er fylkeskommunen som har ansvaret for eksamen eller er det ønskelig ned en statlig ordning?**	**Sett opp en prioritert liste over de 5 viktigste elementene du ser etter hos en kansdidat på muntlig eksamen.**
1 «Elisabeth»	48 timers forberedelsestid der elevene skal forberede en presentasjon til et gitt tema/kompetansemål . Eksaminasjonen varer i 30 min Først 15 minutter presentasjon av eleven Deretter 15 min samtale som tar utgangspunkt i presentasjonen men man kan også bli	Hovedsaklig kompetansemålet knyttet til foredraget eleven holder, men man kan også bli testet i andre kompetansemål gjennom spørsmålene i 15 minutters samtale med eksaminator	Hver faglærer velger ut tema/kompetansemål og lager eksamen.	Eksamen vurderes av eksaminator og ekstern sensor i samarbeid, men sensor har siste ordet. Man har ikke felles vurderingskriterier i fylket, men på denne skolen bruker man et skjema for muntlige foredrag som alle lærerne på denne skolen bruker.	Ja, etter min mening er dette en rettferdig eksamen fordi elevene får 48 timers forberedelse der de kan få veiledning fra faglærer, bruke alle hjelpemidler og har all mulighet til å forberede seg best mulig for å oppnå en god karakter.	Best at det er fylkeskommunen som har ansvaret slik at det er mulig og ta hensyn til lokale forhold og det som er vektlagt av pensum gjennom året. Skolene buker ulike lærebøker og derfor blir tema vektlagt ulikt, selvom kompetansemålene er de samme.	- At eleven kan stoffet sitt godt og er uavhengig av manus/powerpoint - At eleven har gått dypere inn i stoffet/utover det som står i lærebøkene og har oppsøkt andre kilder for å forsøke å gjøre en analyse av tema - At eleven har en klar sammenheng i stoffet/rød tråd gjennom alt som presenteres -At eleven kan svare direkte på spørsmål som gå utover det som er gjennomgått,

	spurt spørsmål knyttet til andre deler av pensum.						og viser evne til å reflektere - At eleven disponerer tiden godt
2 «Birgitte»	Prosjektmodellen bestemt av fylket. Elevene får oppgitt tema 48 timer før muntlig eksamen. Skal forberede 15 min muntlig presentasjon til en selvlaget problemstilling. Problemstillingen er godkjent av faglærer. Ramme på 30 min. 15 min presentasjon m/powepoint. 15 min elevene blir bedt forklare og utdype presentasjonen, samt blir eksaminert i relevant fagstoff knyttet til tema. Sensor og faglærer kan stille spørsmål. Denne skolen bruker individuell eksamen, men gruppe er også mulig	Temaet tar utgangspunkt i et eller fler mål fra læreplanen. Elevene må vise at de kan gå i dybden og er i stand til å bruke kilder selvstendig og kritisk. I tillegg dokumentere kunnskap om tekster/emner som er relevant for oppgitt tema. Altså tekster man har jobbet med tidligere. Også uttrykke egne meninger. Kunnskap, forståelse, språklige ferdigheter testes med nogenlunde lik vekt. Skal ikke lese fra manus Hvis sprik mellom presentasjon og eksaminasjon, veier	Rammen for muntlig eksamen er gitt av fylkeskommunen Faglærer utarbeider tema. Sensor godkjenner det Fylkeskommunen ønsker lik praksis ved alle skoler. Fylkets fagforum i engelsk drøfter derfor muntlig eksamen slik at det skal være tilnærmet likt på alle skoler. I fagforumet sitter er representant fra hver skole. Flere møter i året og rektorkollegiet gir beskjed om hva de ønsker skal tas opp.	Læreplanens mål ligger til grunn for vurderingen. Sensor og faglærer drøfter karaktersetting. Sensor har siste ordet. Ingen felles kriterier, men fagforumet skal utarbeide et skjema dette halvåret.	Ja. Skepsis til prosjektmodellen først, men har vist seg å fungere bra. Elever kan få hjelp, men disse elevene vil ha en fordel uansett eksamensform. Har ikke opplevd stort sprik mellom presentasjon og eksaminasjon.	Positivt med lokalt gitt muntlig eksamen, men man bør ha en felles ramme for hele landet slik at det ikke blir altfor store forskjeller. Statlig detaljstyring er ikke ønskelig.	Helhetsinntrykket teller, men: 1. Evnen til å uttrykke seg med flyt og sammenheng, kommunisere 2. Kunnskap/forståelse/ refleksjon 3. Relevant ordforråd 4. Korrekt grammatikk.

		sistnevnte tyngst.					
3 «Tor»	Lokalgitt. Opptil faglærer å strukturere den og samarbeide med senson om gjennomføring. Vanligst med lang forberedelse, 48 timer. Eleven får utdelt problemstilling/tema. Skal lage ca 10 min presentasjon. I tillegg ofte en ukjent lytteprøve, og også spørsmål fra kjent litteratur/film	Kompetansemålene i læreplanen	Lokalt gitt, uten noen føringer fra fylkeskommunen. Faglærer er ansvarlig for å lage den.	Utgangspunktet er opplæringslova, forskriftene til den og kompetansemålene. Sensor vurderer eksamen. Burde blitt brukt mer tid på å lage felles kriterier. Idag er det opp til sensor stort sett.	Vanskelig å svare på. Men på en måte ja, alle testes etter de samme kompetansemålene og omtrent samme rammer. Selvfølgelig er de t forskjell på elever, eksaminatorer og ikke minst sensorer. Burde være felles kriterier som eksaminator og sensor kjente til i god tid på forhånd. Burde nok være sentralgitte retningslinjer. Samtidig gjør de fleste elever det bedre på muntlig enn skriftlig eksamen.	Det burde være sentragitt muntlig eksamen.	Jeg ser etter kompetansemålene. Legger vekt på helhetlig inntrykk. Konkrete grunnleggende elementer: - Vilje til å snakke - Ordforråd - Rimelig god uttale - Rimelig god intonasjon - Grammatikk
4 «Ellen»	2 modeller i fylket. I språkfag brukes den med 30 min forberedelse. I denne tiden får elevene en ukjent lytteprøve.	En god del av læreplanmålene. Spesielt kompetansemål innenfor kommunikasjon og kultur, samfunn og littertatur.	Fylket har utarbeidet 2 modeller for muntlig eksamen. Faglærer lager forslag til spørsmål som sendes til sensor som kommer med	Ikke utarbeidet felles vurderingskriterier. Sensor kommer med forslag og er eksaminator enig, blir det karakterereren.	Ja. Det vil alltid være element av skjønn, men synes kombinasjonen av sentralgitt skriftlig og lokalgitt muntlig	IKKE sentralgitt muntlig eksamen. Husker sist direktoratet hadde sitt forslag til muntlig eksamen med oppgitt tema 48	- Evne til å kommunisere, avansert språk, riktig engelsk - Evne til å snakke om kjent stoff på en god måte - Evne til å komme

	Under eksaminasjonen får blir de også hørt i et verk (roman/novelle/dikt) dette er ikke bestemt om de får vite hvilket på de 30 minuttene forberedelse.	En blanding av hvor mye de kan snakke om gjennomgått stoff, samtidig som de må samtale om ukjent stoff gjennom lytteprøven.	sine kommentarer. Bakgrunn for dette er at muntlig eksamen skal være lokalt forankret. Synes dette er viktig slik at elevene blir testet i stoff de har blitt undervist tidligere i året. Sentralgitt skiftlig eksamen blir mer og mer generell, selvom det er viktig at skriflig eksamen er sentraltgitt.	Det som er viktig er en rettferdig karaktergivning av kandidatene.	eksamen er meget god.	timer før og presiserte at man ikke skulle spørre utenfor tema. Dette ble helt galt, og nå er det klart at elevene skal eksamineres i fler kompetansemål. Men viktig å være obs pga privatskoler der lærere vurderes, at det kan bli press på sensor om å gi gode karakterer. Man kunne for eksempel oppnevne en kontrollperson som gikk rundt på skolene uanmeldt som ekstra sensor.	med egne tolkninger - Evne til å snakke om den ukjente lytteprøven og dens innhold på en god måte.
5 «Silje»	Lokalgitt. Faglærer ansvarlig for spørsmålene i samarbeid med sensor. Klassen velger eksamensmodell. 2 modeller; spørsmålsmodellen: 30 min forberedelse med alle hjelpemidler tillat og	Kunnskap og språk testes. Elevene må vise evne til å reflektere over oppgitt tema og ved å tydelig avgrense oppgaven, vise innsikt og kanskje greie å trekke tverrfaglig e slutninger. Selvstendighet og fagkunnskap	Faglærer i samråd med sensor lager eksamen. Lages lokalt, men retningslinjer kommer fra fylket.	Vurderes av eksaminator og sensor på eksamensdagen. Ingen vurderingskriterier, men hensiktsmessig og dra på nettverkskurs, slik at fylket får en tilnærmet samkjørt vurdering i faget.	Til en viss grad rettferdig, men man får ofte elever som er sterkere skriftlig enn muntlig, og da vil ikke muntlig eksamen bidra til å reflektere elevens faktiske nivå.	Jeg kunne tenkt meg en statlig ordning, eller i det minste en fylkeskommunal ordning, slik at alle elevene som kom opp muntlig får de samme spørsmålene.	-Fagkunnskap - Grammatikk -Tverrfalig initiativ - Uttale og ordforråd - Grad av kommunikasjon Evne til å kommunisere med eksaminator er viktig, kommer inn både i punkt 1 og 5

	påfølgende 30 min eksaminasjon eller prosjektmodellen: 48 timer forberedelse med oppgitt tema. På eksamen ha en fremføring på ca 10 min, i tillegg være forberedt på å svare på spørsmål fra eksaminator og sensor. I år prosjektmodellen	vektlegges. Eleven må greie å bruke et avansert ordforråd og korrekt grammatikk. Fordel at eleven snakker britisk engelsk eller amerikansk engelsk, ikke begge deler. Prøver å lytte etter stemt/ustemt s, r/R, diftonger og andre lyder som plasserer eleven på et trinn i henhold til språkkunnskap.		Sensor må delta på et kurs i forkant av at de godkjennes som sensor.			
6 «Tonje»	Fylkeskommunen har en egen mal for muntlig eksamen som gjelder for alle fag. 48 timer forberedelse med oppgitt tema. Eleven har et lite foredrag som er basert på dette temaet fra læreplanen. Deretter samtaler elev og eksaminator om både foredraget og om det oppgitte temaet mer generelt.	Utvalgte kompetansemål fra hovedområdet kultur, samfunn og litteratur samt de fleste muntlige og generelle kommunikasjonsmål .	Enkelte faglærer har ansvar for å lage eksamensoppgaver til gitt mal. Fylkeskommunen pålegger oss å samarbeide om oppgavene for å kvalitetssikre dem. Det gjør vi,og det er nyttig. Alltid samme oppgave for alle klasser på skolen.	Individuell vurdering ligger til grunn for muntlig eksamen. Fylket fagnettverk for engelsk har utarbeidet kjennetegn for måloppnåelse som skal brukes.	Ja, det kan sees som urettferdig at elevene kan få hjelp av andre til å forberede foredraget, men de får god opplæring i dette på forhånd, foredraget tillegges mindre vekt enn samtalen etterpå, Skjønn vil alltid spille inn, men føler meg trygg på at elevene får en så objektiv og rettferdig vurdering som mulig	Viktig med felles mal for hele fylket. Gjerne felles mal for hele lande, men viktig at vi får lage oppgavene selv. Gir pedagogisk frihet.	- Kommunikativ kompetanse - Språklig kompetanse - Faglig kompetanse - Struktur

	Trekkes ofte inn litteratur og film og evt. fordypningsemne Åpent for bruk av lytteprøve, men ikke tatt i bruk på denne skolen.						
7 «Liv»	Delt eksamensform; lytteprøve i tillegg til å bli prøvet i pensum. Starter med samtale om lytteprøven. Deretter samtale om trukket emne fra pensum, som vil inneholde en oppgave som er delt mellom litteratur og kultur. I tillegg legges det opp til at eleven skal snakke om sitt fordypningsemne.	Tester ulike mål fra læreplanen. For eksempel fra kommunikasjon og fra kultur, samfunn og litteratur.	Utarbeides lokalt oftest av faglærer i samarbeid med sensor. Retningslinjene er fylkesgitte	Vurderes i samarbeid mellom eksaminator og sensor. Sensor har siste ordet. Fylkets nettverksgruppe i engelsk har utaerbeidet vurderingskriterier som brukes til vurderingen.	Ja. Eksamen inneholder flere elementer. Synes dette er bedre enn prosjektmodellen, for da har ikke alle tilgang på de samme ressursene.	Synes det er riktig med lokal gitt muntlig eksamen Sentrale retningslinjer vil være greit, men ikke sentralgitte eksamensoppgaver siden læreplanen er så vid. Lærere tolker og velge ulikt.	- Selvstendighet - Struktur - Faglig innsikt - Språk - Vilje til å vise det man kan
8 «Karine»	48 timer foreberedelse med oppgitt tema. Prøven er todelt; 10-15 min presentasjon med utgangspunkt i en selvformulert	Kompetansemål fra læreplanen. Oppgitt til elevene ved utdeling av tema: **-discuss social conditions and values in various**	Oppgavene er laget lokalt av lærerne som har elever oppe til muntlig eksamen. Alle elevene får samme oppgaver. Foreligger retningslinjer fra	Vurderingkriterine er de som er gitt elevene på forhånd. I tillegg blir det lagt ved en beskrivelse av elevferdigheter på ulike nivå. Utgangspunkt i	Ingen sterke meninger om dette. Viktig at skolene lager sine egne oppgaver. Viktig med mer generelle	Ikke tenkt så mye på dette, men det ville kanskje vært bedre om staten tok ansvar for dette for å unngå evt store forskjeller fra fylke til fylke.	Vanskelig å lage prioritert liste. Både muntlig ferdigheter og faglig kunnskap er viktig. Umulig å sette den ene over den andre. Viktige elementer:

	problemstilling. Deretter samtale mellom elev og eksaminator om aspekter ved temaet som kom frem i presentasjonen, men også andre aspekter.	**cultures in a number of English-speaking countries -analyse and discuss a film and a representative selection of literary texts in English from the genres poetry, short story, novel and drama** -master a wide vocabulary -use the forms and structures of the language in spoken and written presentations - extract essential information from spoken and written texts (…) - select and use content from different sources independently, critically and responsibly	fylkeskommunen.	vurderingskriterier gitt av udir til skriftlig eksamen. Hver elev fikk vite sin karakter før neste kandidat kom inn.	retningslinjer for fylket, slik at eksamensformen ikke blir for ulik fra skole til skole.		idiomatisk språk, uanhengig av manus, bruke presentasjonen til å støtte opp om det som blir sagt. Blooms taksonomi, presentere faktakunnskap, se sammenhenger, årsaksforhold, konsekvenser.
9 «Mari»	48 timer forberedelse til gitt emne. Elevene trekkerpoblemstilling i et utvalg av 8 og skal lage en powepoint til denne. Eksamen begynner med en lytteprøve i gruppe deres elevene	Kompetansemål fra læreplanens del kultur og samfunn og kompetansemålene knyttet til språk. Viktig å teste både formelle og uformelle	Den lokale skolen og faglærer i samarbeid med sensor er ansvarli g for eksamen	Det er lokalt gitte vurderingskriterier som brukes av både eksaminator og sensor. Disse er også utdelt til elevene Spesielt fokus på	Den er rettferdig innenfor skolen, men ikke på fylkesnivå elle landsnivå. Denne eksamensformen kan være noe vanskeligere enn det som praktiseres	Mener det ville vært best med en statlig gitt muntlig eksamen. Jo mer sentral jo bedre. Aller viktigst at vurderingskriterier er sentralgitte.	- Generell flyt i språket og kommunikasjon - Relevant innhold i forhold til læreplanmålet som eksamineres - Selvstendig drøfting og vurdering av

	skal samtale sammen om en tekst de hører. For å teste uformell kommunikasjonsevne Deretter den forberedte presentasjonen Deretter oppfølgende samtale	kommunikasjonsevner		uttale, ordforråd, grammatikk og riktig bruk av formell og uformellt språk	andre steder.		eksamensspørmsålet - Riktig uttale - Språklig presisjon
10 «Helga»	48 timers forberedelse til gitt tema Elevene får velge en problemstilling fra en liste eller de kan lage en selv. ca 15 min presentasjonen ca 15 min oppføgelgende samtale der eksaminator kan stille spørsmål knyttet til presentasjonen eller andre tema fra læreplanen.	Tester hvor godt elevene løser problemstillingen sin og muntlig produksjon knyttet til kompetansemålene	Den lokale skolen og faglærer er ansvarlig, men det er gjort mye jobb for å lage en felles mal for fylket	Det brukes lokalt lagede vurderingskriterier som både eksaminator og sensor bruker ved vurderingen. Det er gjort mye jobb for å lage felles kriterier for fylket, men dette er nå forkastet av fylkeskommunen.	Ja, synes det er en rettferdig eksamen innenfor dette skolen, men ikke på landsbasis. Viktig med den oppfølgende samtalen for å teste spontan kommunikasjonsevne	Ville vært svært ønskelig med en statlig gitt eksamen. Særlig er det ønskelig med sentralgitte vurderingskriterier for å sikre en rettferdig eksamen. Tror det er store forskjeller mellom skoler og fylker og frykter for eksamens reliabilitet med tanke på at elevene kjemper om de sammme studieplassene med eksamensresultatene sine	Vanskelig med prioritering. Mange ting som prøves ved en muntlig eksamen. Læreplanmålene danner utgangspunktet for vurderingen og jeg rangerer dem vanligvis ikke etter viktighet.
11 «Ingrid»	48 timers forberedelse. Elevene trekker en problemstilling.	Kompetansemålene fra læreplanen brukes når de lager problemstillinger.	Den lokale skolen er ansvarlig både for oppgaver og eksamensform.	Det er ingen vurderingskriterier Sensor vurderer utfra	Ja, synes det er en rettferdig eksamen . Elevene får vise mye	Har ingen mening om dette	- Syntese mellom språk og innhold - Struktur - Flyt

	Problemstillingene er knyttet til kompetansemålene. Forberedt presentasjon der eksaminator kan stille spørsmål underveis Oppfølgende samtale om presentasjonen, men kan også trekke inn andre kompetansemål knyttet til temaet.	Testes om de har den kompetansen som er nødvendig i forhold til kompetansemålene. Tester også innholdet av presentasjonen, muntlig produksjon. Tester også hvor bra laget presentasjone er.	Men det finnes retningslinjer laget av fylkeskommunen som blant annet presenterer ulike eksamensformer. Men det gis stor frihet til de lokale skolene.	egen erfaring og egne kriterier.	av sine ferdigheter og kompetanse. Samtidig føler hun at den gamle eksamensmodellen ville teste mer enn dagens.		- Variasjon - Grammatisk korrekthet
12 «Kari»	48 timer forberedelse Får utdelt oppgavene som de skal forberede og legge frem på eksamen, elevene kan velge mellom 2 oppgaver. Følges opp av en oppfølgende samtale etter presentasjonen.	Kunnskap i fohold til oppgaven de har forberedt Hvordan elevene evner å legge frem oppgaven Ordforråd, flyt grammatikk Hvordan eleven viser refleksjon og kan trekke inn annen relevant kunnskap	Eksamen lages lokalt av den enkelte faglærer som får opp en gruppe elever, i samarbeid med sensor. Finnes svært generelle retningslinjer fra fylket som sier at det skal elevene skal få utlevert oppgaver når forberedelsen starter og at de skal kunne velge mellom to oppgaver knyttet til engelsk språk eller litteratur.	Ekstern sensor har det siste ordet og godkjenner eksameninasjonen. Man bruker vurderingskriterier i tillegg til at kompetansemålene er styrende for vurderingen. Disse er lokalt laget, men basert på at fylket sier at fremleggingen skal være klar og systematisk, karakteren skal bygge på elevens	Ja, fordi man vurderer alle elever etter de samme vurderingskriteriene og kopmpetansemålene	Ønsker en statlig ordning slik at det blir like forhold over hele landet	- Kandidaten svarer godt og begrunnet på oppgaven, er overbevisende og viser kunnskap - At kandidaten behersker et nyansert og veltilpasset ordforråd til tema - At kandidaten kan relatere emne til andre emner som er gjennomgått i klassen - At emnet blir lagt frem på en velformulert og klar måte - At kandidaten er reflektert, har

				kompetanse i faget (kompetansemålene) oppmerksomhet mot elevens faglige innsikt og forståelse.			øyenkontakt med tilhører, frigjør seg fra manus og behersker grammatiske elementer tilpasset nivået.
13 «Erik»	48 timers forberedelse Foredrag over et tema som er trukket ut på forhånd der eleven selv har utarbeidet en problemstilling ca 10-15 min oppfølgende samtale som tar utgangspunkt i foredraget, men andre emner fra pensum kan også trekkes inn	Evnen til å holde et foredrag, lytte, føre en samtale på engelsk. Uttale, intonasjon, vokabular	Eksamen lages av faglærer lokalt.	Ekstern sensor setter den endelige karakteren i samråd med faglærer. Ja, det finnes felles vurderingskriterier.	Bare delvis rettferdig. Favoriserer til en viss grad elever som har muligheter til å få hjelp under forberedelsen	Kan ikke se for seg hvordan en sentralgitt muntlig eksamen skulle kunne gjennomføres	Om eleven har svart på egen problemstilling - Evne til å uttrykke seg - Evne til å føre en samtale - Uttale, ordforråd, intonasjon - Evne til å oppfatte spørsmål fra eksaminator/sensor
14 «Lars»	48 timers forberedelse der elevene skal forbereder en kort innledning til et gitt tema. På eksamensdagen gis elevene en lytteprøve 30 minutter før eksaminasjonen.	Alle lærplanmål kan testes, men det skal spesifiseres på oppgaven hvilke læremål som testes i denne oppgaven	Eksamen lages av faglærer i samarbeid med sensor Oppgaver, leselister, og vurderingskriterier sendes til sensor senest 4 dager før eksamen for godkjenning	Vurderingen skjer i samarbeid mellom eksaminator og sensor, men det er den eksterne sensoren som har det siste ordet. Det bruker felles vurderingskriterier som er laget til muntlig eksamen i	Mener at muntlig eksamen aldri vil bli helt rettferdig men at det at man har felles vurderingskriterier hjelper for gjøre eksamen mer rettferdig.	Synes den nåværende ordningen fungerer godt.	Lærer henviser her til vurderingskriteriene som blant annet nevner: uttale, intonasjon, formverk, idiomtisk språk, kunne delta i muntlig interaksjon,

	Eksaminasjonen begynner med spørsmål til lytteprøven (10 min) Deretter har de sin korte innledning til tema (5 min) 15 minutter står igjen til samtale med eksaminator om innledningen men også knyttet til hele temaet som innledningen er basert på.			engelsk. Disse har sensor fått oversendt i forkant			pridusere språk tilpasset situasjon og tema, kunne analysere, drøfte og vurdere og se kunnskapen i sammenheng
15 «Hanne»	48 timers forberedelse hvor elevene forbereder en presentasjon. Eksaminasjone er 3 delt og vare i ca 30 min 1) Lytteprøven (20 minutters forberedelse) 2) Elevens presentasjon 3) Samtale mellom elev og eksaminator om tema knyttet til presentasjonen	Hvor langt eleven er kommet i forhold til målene for faget. Innhold og evne til kommunikasjon og refleksjon	Faglærer har ansvaret for å utarbeide eksamensoppgaven, ofte i samarbeid med andre lærere og noen ganger sammen med sensor. Oppgaven utarbeides etter retningslinjer fra fylkeskommunen. Disse er diskutert i fylkesnettverket for engelskfaget og fagseksjon på egen skole	I samarbeid mellom faglærer og sensor. Vurderingen bygger på målene for faget og vurderingskriterer utarbeidet av fylkesnettverket og drøftet i fagseksjonen. Disse vurderingskriteriene er felles og må følges av alle sensorer.	Så rettferdig som det er mulig å få til. Favoriserer de elevene som er mest verbale og har lett for å vise sin komptanse muntlig og klarer og reflektere under press.	Det kunne kanskje vært greit med sentrale retningslinjer, men det viktigste er at eksamen legges til rette slik at eleven får vist sin kompetanse og at kompetansen vurderes likt uansett hvor i landet man er/hvilken skole/hvilken sensor. Betryggende at man må forholde seg til vurderingskriterier, men tror ikke det er mulig å lage helt	-Evne til kommunikasjons -I hvilken grad eleven når frem til samtalepartner/publikum -I hvilken grad eleven kan gi en presis og fyldig redegjørelse for et tema - Størrelse på ordforråd slik at eleven kan uttale seg presist og relevant i forhold til tema - Uttale og språk Disse er avhengig av hverandre og det er

						objektive vurderingskriterier. Profesjonelt skjønn vil alltid være viktig	helheten som er viktig til slutt.

Appendix 3

An historical overview of validity research

A historical overview

From the early days of investigating validity, it has traditionally been split up in three types. Cronbach and Meehl (1955) defined these types as

- Criterion-oriented validity
- Content validity
- Construct validity

I will in the following give a short introduction of these terms, based on the work of Fulcher and Davidson (2007).

Criterion-oriented validity is concerned with the "relationship between a particular test and a criterion to which we wish to make predictions" (Fulcher & Davidson 2007, p. 4). There are two types of criterion-oriented validity; predictive validity and concurrent validity.

Predictive validity is when the scores are used to predict future criterion (Fulcher & Davidson 2007). This can for example be the case when a test is measuring how well students in upper secondary handle reading academic English before entering higher education where it will be necessary for them to be able to read academic English.

Concurrent validity is when the scores are used to predict a criterion at the same time as the test is given (Fulcher & Davidson 2007). An example of this could be a shorter test than the one above, given to predict the scores on the longer test. In this case it is the longer test that is the criterion.

Content validity can be defined as "any attempt to show that the content of a test is a representative sample from the domain that is to be tested" (Fulcher and Davidson 2007, p. 6). I will continue to use the example of a language test to measure how well students in upper secondary handle reading academic English. In this case it would be necessary to ensure that the texts given on the test were texts that are typical for academic texts in higher education. Furthermore, the test should measure how well the students process the texts in comparison with how it is expected of students in higher education to process academic texts. Usually this means to extract key information, take notes and use the information extracted to write their own texts. (Fulcher and Davidson 2007).

Construct validity is the third type of validity Cronbach and Meehl (1955) defines. The first problem with construct validity is defining what a construct is. It is possible to define a construct as

> “ a concept that is defined so that it can be scientifically investigated. This means that it can be *operationalized* so that it can be measured. Constructs are usually identified by abstract nouns, such as “fluency”, that cannot be directly observed in themselves but about which we need to make inferences from observations.”
>
> (Fulcher & Davidson 2007, pp. 369-370).

Furthermore, Cronbach and Meehl (1955) say that

> “Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are the means of confirming or disconfirming the claim”
>
> (Cronbach & Meehl 1955 quoted in Fulcher & Davidson 2007, p. 8).

This represents the view on construct validity before 1989. However, as I mentioned above, the view on validity changed after this with Samuel Messick’s work from 1989 (Fulcher & Davidson 2007). In the following section I will look at the question of validity after 1989.

Validity after 1989

Up until 1989 the way to consider validity was to measure it in the three types that I have presented in the previous sections. However, Messick (1989) wrote that “because criterion-related and content-related evidence contribute to score meaning, they have come to be recognized as aspects of construct validity” (Messick 1989 quoted in Fulcher & Davidson

2007, p. 12). Messick meant that this left only one category; construct validity. He further described validity as:

> “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment”
>
> (Messick 1989 quoted in Fulcher and Davidson 2007, p. 12).

This changed the way we understand validity fundamentally. Fulcher & Davidson say that with this view, validity “is not a property of a test or assessment but the degree to which we are justified in making an inference to a construct from a test score” (Fulcher & Davidson 2007, p. 12).

Looking more at the development of validity thinking after 1989, Fulcher & Davidson (2007) present theories by Chapelle. In her works from 1998 and 1999 she has characterized three current approaches to validity:

- Trait theory
- New behaviorism
- Interactionist understanding of score meaning

Trait' means approximately the same as the term 'construct' used by Cronbach and Meehl presented above. In trait theory it is assumed that the construct to be tested is an attribute to the test taker, meaning that the test taker's knowledge and processes are considered stable and real, and that the test is supposed to measure these. This again means that “score meaning is established on the basis of correspondence between the score and the actuality of the construct in the test taker (Fulcher & Davidson 2007, p. 16).

When speaking about new behaviorism and a behaviorist approach, the test score is mostly affected by context. This means that because there in real world communication is always a context or a setting, this context has to be copied as best possible in the test, or else we cannot infer meaning form the test score to the real world criterion. (Fulcher & Davidson 2007). An example of a real world situation is to order food in a restaurant, and the conversation between waiter and customer. If you want to test how well students are able to communicate in this kind of situation, you have to make the test situation as realistic as possible. The validity check would be to see how well the test situation corresponds with the real world situation.

The third approach is an interactionist understanding of score meaning ,described by Chapelle (quoted in Fulcher and Davidson 2007, p.17) as “the result of traits, contextual features, and their interaction” where “performance is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts”. This means that one accept that a test is only a sample of a real world situation we want to generalize. The validity check then becomes a question of finding evidence showing that the sample is relevant to the real world situation.

As we can see from the development in validity theory presented above, validity theory is in itself changing and evolving. And as Fulcher & Davidson (2007) say there is no absolute answer to the validity question. In the following section I will look at validity in speaking tests in particular.