What is to be assessed?

Teachers’ understanding of constructs in an oral English examination in Norway

Henrik Bøhn

Thesis submitted for the degree of Philosophiae Doctor

Department of Teacher Education and School Research
Faculty of Educational Sciences
UNIVERSITY OF OSLO

2016
Summary

The present thesis has investigated EFL teachers’ rating orientations in an oral English examination at the upper secondary level in Norway. As part of this investigation, aspects of the teachers’ scoring behaviour, i.e. grades given, have also been studied. In addition, comparisons were made between what the teachers understand as relevant performance aspects to be tested and what the English subject curriculum and accompanying government documents define as construct-relevant.

The thesis is article-based and comprises three articles and an extended abstract. The extended abstract provides a background for the investigation, a theoretical framework, a literature review, a presentation of the research design and methods used, as well as a discussion of the main findings. The articles present the three individual studies which have been conducted. A major concern throughout the thesis has been the lack of a common national rating scale in the upper secondary school context and how this may affect the validity and reliability of the scores.

In Study 1 (Article 1) semi-structured interviews were used to explore 24 Norwegian EFL teachers’ general understanding of the constructs to be tested. The study found that the teachers focused on two main constructs, namely ‘communication’ and ‘content’, which in turn comprised a number of sub-constructs. Overall, the teachers understood the main constructs in the same way, but they disagreed on some of the more specific performance aspects, such as ‘pronunciation’. In addition, the study found that teachers weighted the content construct differently. The teachers working in the general studies programme put more emphasis on content than did the teachers in the vocational studies programmes. There was also evidence that some teachers focused on construct-irrelevant performance features, such as effort. Beyond this, the investigation of scoring behaviour indicated that there was fairly good agreement in the scoring of performance.

Study 2 (Article 2) used semi-structured interviews and questionnaires to investigate the rating orientations of 70 EFL teachers regarding aspects of the pronunciation construct. These aspects included native speaker pronunciation and intelligibility, as well as the pronunciation of segmentals (individual sounds), word stress, sentence stress and intonation. The results showed that the teachers had widely differing views on native speaker pronunciation, but that they strongly agreed that intelligibility was important for a high-scoring performance. In addition, they largely agreed that segmentals, word stress and sentence stress were important features to be assessed. As for intonation, however, the
findings indicated that the teachers were either not as concerned with this feature, or unsure of its relevance.

Study 3 (Article 3) employed verbal protocol analysis and semi-structured interviews to explore 10 EFL teachers’ understanding of the content construct. This construct was mainly analysed in terms of a subject matter dimension and a skills and abilities dimension. Comparisons were also made between the teachers’ perceptions of content and aspects of content identified in the subject curriculum and accompanying government documents. The results showed that the teachers had a very general understanding of subject matter, largely interpreting it in terms of ‘general world knowledge’, which may be said to correspond well with what the subject curriculum stipulates. In addition, the study found that the teachers were more concerned with the skills and abilities dimension than with the subject matter dimension, stressing the importance of higher-order thinking skills for a top-scoring performance. There was also evidence that the teachers largely had the same understanding of the construct, but that some of them disagreed on what kind of performance was indicative of the different achievement levels. These differences were largely attributed to study programme affiliation, the vocational studies teachers being more lenient in their assessment orientations than the general studies teachers.

In sum, the three articles provide empirical evidence of what kind of performance aspects teacher raters attend to in a curriculum-based, oral EFL assessment context at the upper-intermediate level (Common European Framework of Reference B1/B2 level) where no common rating scale exists. Overall, the results showed that the teachers had a similar understanding of the main constructs to be tested, but that they disagreed on the more narrow performance aspects. The study also indicated that constructs such as pronunciation and content are somewhat elusive and need to be better defined in order to provide scoring outcomes that are valid and dependable. In addition, the findings suggested that the Norwegian educational authorities should consider introducing common rating scale guidelines, as well as more coherent rater training, in order to guide teachers in their assessment of oral English performance.
Acknowledgements

This thesis would never have been finalized without the help of a number of people. First of all, I am greatly indebted to my main supervisor, Professor Glenn Ole Hellekjær, for all his excellent advice, helpful comments, genuine care and pragmatic guidance when I got lost in details along the way, which I constantly did. I am also indebted to my co-supervisor, Professor Hilde Hasselgård, for her great repertoire of practical advice, razor-sharp analyses of my drafts, eminent suggestions for text improvement and wonderful sense of humor.

I also heartily thank the student who had the courage to let me film her during her oral exam. Without her, this project would have been very different. Similarly, the 80 teachers who agreed to watch the video-clip and answer my questions as interview informants and questionnaire respondents deserve credit for sharing their precious time. Special thanks go to the teachers Christian Andresen at Jessheim Upper Secondary School, Phil Grey at Bjørkelangen Upper Secondary School and Margrethe Hall Christensen at Ås Upper Secondary School for granting me access to study participants.

Likewise, a number of good colleagues at Østfold University College must be acknowledged. First of all, I want to thank Associate Professor Magne Dypedahl for his rock-solid support through all these years. His untiring encouragement, insightful analyses and pointed comments on different text versions have been particularly helpful in this process. I am also indebted to Assistant Professor Thomas Hansen for agreeing to co-author the second article of this thesis with me, for helping me to analyse data and critique drafts and for his admirable good spirits. Professor Roald Jensen also deserves great thanks for sharing his profound knowledge of learning and assessment theory and for his thorough and helpful feedback on two of the chapters. My sincere thanks are furthermore due to Professor Julianne Cheek, whose competence in the area of research methodology is truly of the highest academic standard. Her lucid criticism of parts of this text proved immensely valuable to me. I am also indebted to my former Dean, Associate Professor Eva Lambertsson Björk, for her backing and counselling in the very early stages of the project.

In the same manner I owe a great deal to helpful colleagues at the University of Oslo. I am particularly grateful to the members of the SISCO research group, who will leave a lasting impression on my academic career, because of all the things they taught me about good research. Professor Kirsti Klette and Professor emerita Frøydis Hertzberg, who chaired this group as I joined it in 2013, and who gave me such a warm welcome, combine the highest academic standards and a human touch in an exemplary way. The same can be said of
Professor Andreas Lund, who generously set aside time to read and provide excellent comments on my PhD project proposal before I was admitted to the Faculty of Education. Similarly, Associate Professor Ulrikke Rindahl should be credited for kind-heartedly providing very valuable feedback on various versions of the articles and the extended abstract. Also, former Master’s student, Caroline Borch-Nielsen, contributed insightful ideas on the development of the questionnaire which I used for my pilot.

Furthermore, I want to express my gratitude to Associate Professor Therese N. Hopfenbeck at the University of Oxford, for her superb comments on various parts of the thesis in the final stages of the project, and to Mathilda Burman and Kim Buxton at the Norwegian Directorate for Education and Training who kindly shared their time to inform me of assessment policies in the Norwegian context. I must also thank the library staff at Østfold University College, for their excellent and speedy service, always willing to make an extra effort to provide me with sources before deadlines. In the same vein I am obliged to Geir Jarberg and Anne Grethe Bremnes, at the ICT technical support unit at Østfold University College, who always jumped to their feet to assist me whenever my computer jammed, when a software license expired or when my lack of digital skill prevented me from getting my writing done.

Finally, and most importantly, I want to thank my family. I am truly grateful to my parents, Tor Inge and Hildegunn, for their unwavering support in all respects, to my sister, Anne Marie, whose experience as an English teacher brought fruitful ideas into this project, and to my children, Eva Julie and Simon, who always cheer me up. I dedicate this thesis to my wife, colleague and best friend, Gro-Anita, whose love, compassion, academic agility and moral and practical support got me through this long process. You have my heart.

Halden, 30 May 2016
Henrik Bøhn
Table of contents

Summary ................................................................................................................................ iii
Acknowledgements .................................................................................................................... v
Table of contents .................................................................................................................. vii
List of tables ........................................................................................................................ x
List of figures ........................................................................................................................ x
Abbreviations ...................................................................................................................... xi

Part I: Extended abstract

Chapter 1: Introduction .......................................................................................................... 1
1.1 General background ....................................................................................................... 1
1.2 Assessment paradigms in the educational domain ........................................................ 2
1.3 The Norwegian context ............................................................................................... 4
   1.3.1 Assessment in Norway ...................................................................................... 4
   1.3.2 English in Norway ......................................................................................... 5
   1.3.3 The Norwegian school context ...................................................................... 6
1.4 Research purpose and aims......................................................................................... 8
1.5 A note on terminology ............................................................................................... 9
1.6 The structure of the thesis......................................................................................... 10

Chapter 2: Theoretical framework ...................................................................................... 11
2.1 Introduction to the chapter ....................................................................................... 11
2.2 Constructs: Operationalization of underlying abilities .............................................. 11
2.3 Validity and validation ............................................................................................... 14
   2.3.1 Evaluating evidence about score interpretations ............................................. 14
   2.3.2 The unitary theory of validity and argument-based validity approaches ......... 15
   2.3.3 The social consequences of test use ............................................................. 18
Chapter 3: Literature review

3.1 Introduction to the chapter
3.2 International research on rater orientations and rater behaviour
   3.2.1 Rater variability and the focus of the present investigation
   3.2.2 Rater orientations
   3.2.3 Rater behaviour
3.3 Assessment research in the Norwegian context
3.4 Short summary

Chapter 4: Methodology

4.1 Introduction to the chapter
4.2 The phases of the research process
   4.2.1 General overview
   4.2.2 The pilot phase
   4.2.3 Study 1 (Article 1)
   4.2.4 Study 2 (Article 2)
   4.2.5 Study 3 (Article 3)
4.3 Data collection
   4.3.1 Study 1: Teachers’ overall understanding of constructs
   4.3.2 Study 2: Teachers’ orientations towards pronunciation
   4.3.3 Study 3: Teachers’ understanding of content
4.4 Participants and procedure
4.5 Data analyses
4.6 Research validity
4.7 Ethical considerations
4.8 Generalizability................................................................. 56
4.9 Short summary................................................................. 57

Chapter 5: Summary and discussion ........................................ 58
5.1 Introduction to the chapter.................................................. 58
5.2 Summary of the articles..................................................... 58
   5.2.1 Article 1................................................................. 58
   5.2.2 Article 2................................................................. 61
   5.2.3 Article 3................................................................. 63
5.3 Research contribution ....................................................... 66
5.4 Implications for the Norwegian educational context.................. 68
5.5 Concluding remarks......................................................... 71

Part II: Articles


List of tables

Table 1  CEFR production scale for overall spoken production. ........................................  22
Table 2  CEFR production scale for overall spoken interaction  ........................................  23
Table 3  Overview of the central elements in the research process .................................  50

List of figures

Figure 1  CEFR’s model of communicative language competence, Council of Europe (2001) ................................................................. 21
Figure 2  Competence aims underlying the two constructs to be tested; taken from the English subject curriculum, GSP1/VSP2 level ......................... 25
Figure 3  Overview of the research design ................................................................. 37
## List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AfL</td>
<td>Assessment for Learning</td>
</tr>
<tr>
<td>CBI</td>
<td>Content-Based Instruction</td>
</tr>
<tr>
<td>CLA</td>
<td>Communicative Language Ability</td>
</tr>
<tr>
<td>CEFR</td>
<td>Common European Framework of Reference for Languages</td>
</tr>
<tr>
<td>EAP</td>
<td>English for Academic Purposes</td>
</tr>
<tr>
<td>EFL</td>
<td>English as a Foreign Language</td>
</tr>
<tr>
<td>ESL</td>
<td>English as a Second Language</td>
</tr>
<tr>
<td>FYR</td>
<td>Fellesfag, yrkesretting, relevans [Vocational Education and Training Promotion]</td>
</tr>
<tr>
<td>GSP</td>
<td>General Studies Programme</td>
</tr>
<tr>
<td>GSP1</td>
<td>General Studies Programme, year 1 at the upper secondary school level</td>
</tr>
<tr>
<td>IELTS</td>
<td>The International English Language Testing System</td>
</tr>
<tr>
<td>LEA</td>
<td>Local Educational Authorities</td>
</tr>
<tr>
<td>LFC</td>
<td>Lingua Franca Core</td>
</tr>
<tr>
<td>LK-06</td>
<td>Læreplanverket for kunnskapsløftet [The Knowledge Promotion curriculum reform]</td>
</tr>
<tr>
<td>L1</td>
<td>First language</td>
</tr>
<tr>
<td>L2</td>
<td>Foreign or second language</td>
</tr>
<tr>
<td>NKVS</td>
<td>Nasjonalt kvalitetsvurderingssystem [National Quality Assessment System]</td>
</tr>
<tr>
<td>PIRLS</td>
<td>Progress in International Reading Literacy Study</td>
</tr>
<tr>
<td>PISA</td>
<td>Programme for International Student Assessment</td>
</tr>
<tr>
<td>RQ</td>
<td>Research question</td>
</tr>
<tr>
<td>TIMMS</td>
<td>Trends in International Mathematics and Science Study</td>
</tr>
<tr>
<td>TOEFL</td>
<td>Test of English as a Foreign Language</td>
</tr>
<tr>
<td>UDIR</td>
<td>Utdanningsdirektoratet [Norwegian Directorate for Education and Training]</td>
</tr>
<tr>
<td>VPA</td>
<td>Verbal Protocol Analysis</td>
</tr>
<tr>
<td>VSP</td>
<td>Vocational Studies Programme</td>
</tr>
<tr>
<td>VSP2</td>
<td>Vocational Studies Programme, year 2 at the upper secondary school level</td>
</tr>
</tbody>
</table>
Part I

Extended abstract
Chapter 1: Introduction

Speaking is [...] the most difficult skill to assess reliably.¹

1.1 General background

Educational assessment can be defined as “the planned and systematic process of gathering and interpreting evidence about learning in order to make a judgement about that learning” (Tina Isaacs, Zara, Herbert, Coombs, & Smith, 2013). This practice has a fundamental role in education and involves a number of political, philosophical, social, ethical and technical issues. In the past two decades assessment has received increasing attention from researchers, policy makers, teaching practitioners and the general public, not least because of the importance that has been attributed to the role of formative assessment in the advancement of student learning (Black & Wiliam, 1998; Hattie, 2009; Sadler, 1998; Stiggins, 2005). Other types of assessment practices have also seen an upsurge, especially in the form of national and international large-scale language testing, which may have diagnostic, comparative and accountability functions (Bennett & Gitomer, 2008; Hopfenbeck, 2014; Kunnan, 2008; Stobart & Eggen, 2012). In addition, school exams and other summative assessments continue to occupy an important place in school systems, serving achievement record and certification functions with potential high-stakes consequences. This thesis investigates the latter domain, with a focus on the rating process in an oral English exam.

Having worked as an English teacher at the upper secondary school level in Norway for three years, and as a lecturer at the tertiary level for more than 15 years, I have been involved in numerous assessment situations, many of which have been oral exams. My first encounter with an oral English school exam as a young teacher in the late 1990s came to epitomize some of the challenges of the Norwegian educational assessment system, which I encountered when embarking on this PhD-project many years later. Not only was there no interlocutor training to help me ask the right kinds of questions during the examination, there were no rater training and no rating scale to guide me in the rating process either. With no experience at all I felt very much dependent on the judgement of my more experienced co-assessor. In retrospect, as I became increasingly aware of the complexities of assessment, I have been pondering the arbitrariness of the system that I encountered, which no doubt sparked my subsequent interest in assessment research.

In the almost 20 years that have passed since then, the Norwegian educational system has undergone important changes, and international advances in testing and assessment research have improved our understanding of the nature of assessment and of how to improve testing practices. Despite this, nationally administered rater training and rating scales for oral exams are still non-existent in Norway. Moreover, the international research community continues to grapple with a number of unsolved challenges. For example, there are problems related to the issues of reliability (as indicated by Alderson and Bachman in the introductory quote), assessment design, score interpretation and test use. Other dilemmas relate to the nature of language ability, the appropriateness of the scoring system and the ethical uses of language assessments (Bachman, 2014; Davies, 2014; Green, 2014).

With regard to the appropriateness of the scoring system, the question of validity is particularly important. Very simply put, validity can be understood as the extent to which the inferences drawn from assessment scores are ‘sound’ (Cronbach, 1971; Fulcher, 2015). In order to ensure that inferences are sound, raters must have a common conceptualization of what is to be assessed. This ‘what’ is regularly referred to as the construct (O'Sullivan, 2014). If raters do not have a shared understanding of the construct, it will negatively affect validity (and reliability) (Jones, 2012; Taylor & Galaczi, 2011). Validity is therefore a fundamental aspect of the quality of the assessment procedure (Newton & Shaw, 2014; Wiliam, 2008).

In this thesis I address the issues of score interpretation and the appropriateness of the scoring system. The focus is on the assessment of spoken English as a Foreign Language (EFL) in an oral examination at the upper secondary school level. My main aim is to identify what aspects of performance teachers pay attention to in the rating process. As part of this investigation I compare their notions of relevant performance aspects with what the curriculum and other defining documents identify as relevant features to be assessed. In addition, I examine aspects of teachers’ scoring behaviour.

1.2 Assessment paradigms in the educational domain
Assessment is not undertaken in a vacuum. Ontological and epistemological assumptions, tradition, values and ideologies all affect the way assessment is looked upon, researched, designed, implemented and appraised. Taken together on a general level, facets such as these may be said to form a paradigm, or “a set of interrelated concepts which provide the framework within which we see and understand a particular problem or activity” (Gipps, 1994, p. 1). Not infrequently, ideas attributable to different paradigms may exist side by side,
creating tensions in societal systems. In educational assessment such tensions can be observed in views and practices stemming from two overarching paradigms, namely the measurement, or psychometrics, paradigm and the assessment paradigm. In order to understand the oral English exam under investigation here, it is relevant to discuss some of the main features of these paradigms and how they are reflected in educational assessment practices.

The measurement paradigm originated in the field of psychology in the 19th century and was traditionally associated with a positivist epistemological outlook (Baird, Hopfenbeck, Newton, Stobart, & Steen-Utheim, 2014; Broadfoot, 2007). A basic assumption in this paradigm is the idea that abilities are fixed individual properties which can be ‘measured’, or ‘tested’, quantitatively. Norm-referenced test practices are frequent, and reliability and standardization are of major concern (Baird et al., 2014; Broadfoot, 2007; Gipps, 1994). In order to enhance reliability, externally defined criteria, or standards, are commonly preferred. In terms of learning, behaviouristic and cognitive models are frequently drawn upon (Inbar-Lourie, 2008), and knowledge is often believed to exist separately from the learner. From this perspective, tests can be designed to objectively assess the amount of knowledge that a student has acquired (Serafini, 2001). Thus, an important purpose of tests in education is to monitor learning (Inbar-Lourie, 2008). Other important purposes are ranking, reporting, surveillance and the certification of competence (Black & Jones, 2006; Inbar-Lourie, 2008).

The assessment paradigm, on the other hand, which was developed in the late 20th century, is sometimes seen as a reaction against the psychometrics tradition (Throndsen, Hopfenbeck, Lie, & Dale, 2009). Based on interpretivist and constructivist epistemological positions, this paradigm typically sees abilities as evolving and contextually sensitive (Inbar-Lourie, 2008). On this view, learning is typically understood as knowledge construction, rather than something which is objectively acquired (Hargreaves, 2005). Moreover, there is a preference for criterion-referenced forms of assessment, whereas reliability and standardization are de-emphasized (Gipps, 1994; Inbar-Lourie, 2008). The main purpose of assessment is to promote learning, and in this process the teacher has a prominent role. Hence, criteria can legitimately be designed and implemented on the local level. Engh (2011) goes so far as to say that:

---

2 Other labels have been used to describe these paradigms. Inbar-Lourie (2008), for example, refers to them as “testing culture” and “assessment culture” (p. 285).

3 This may be a problematic stance in validity frameworks which incorporate reliability into validity (cf. section 2.3.3). However, in some approaches it may seem possible to have validity without reliability (Moss, 1994).
The teachers’ technical expertise is used, among other things, to assess student competence. This type of assessment is to be carried out on the basis of the teachers’ professional judgement. Only in exceptional cases is it possible or pedagogically sensible to use standards for assessing student performance. In most cases, what we are assessing is quality, and quality cannot be assessed with the use of standards or other quantitative measures (p. 17, my translation).

The Assessment for Learning (AfL) approach, which has been influential in Norway (cf. section 1.3.1, below) draws heavily on this tradition. This is an approach which reflects social constructivist, cognitive and socio-cultural theories of learning (Black & Wiliam, 2009).

Features of these two paradigms are recognizable in a number of educational systems, Norway being no exception. Traces of the measurement paradigm, for example, are evident in national and international large-scale (external) testing practices, which generally have a quantitative orientation and where the monitoring of learning is one important function of such assessment. However, what is even more relevant for the present thesis is the way aspects of the two paradigms are reflected in various forms of school-based (internal) assessment. For example, assessments used for formative purposes draw largely on the assessment paradigm, whereas summative assessments, such as exams, tend to share more features with the measurement paradigm. Still, there are overlaps, and in the concrete design and implementation of different assessments some important questions relating to differing views from the two paradigms need to be asked, for instance: How standardized do examinations need to be? Are common rating scales required? Is rater training absolutely necessary? This thesis presents empirical findings which relate to these questions and discusses potential consequences of choosing some solutions over others.

1.3 The Norwegian context
1.3.1 Assessment in Norway

Since the mid-2000s, assessment has been a major area of attention for Norwegian educational authorities, reflecting recent international trends in education (Andreassen & Gamlem, 2011). This development was prompted by the low average results of Norwegian students on international tests such as PISA, TIMMS & PIRLS after the turn of the

---

4 I here follow Harlen (2012) who refers to formative assessment as assessment intended to “help learning” and summative assessment as assessment intended to “report learning” (p. 97). That being said, it is also clear that exams generally have a certification function.
millennium (Engh, 2011). On the release of the first PISA results in 2001, the authorities immediately initiated a number of research projects to find out why Norwegian students did not perform better. The results of this research identified a number of challenges in the area of assessment, particularly with regard to formative evaluation. For example, studies found that feedback practices were unsystematic and poorly related to learning objectives, indicating that teachers lacked assessment competence (Haug, 2004; Haugstveit, 2005; Hertzberg, 2003; Klette, 2003; Solstad & Rønning, 2003). The government concluded that there was a “weak assessment culture” in many Norwegian schools (Meld. St. 16 (2006-2007), 2007, p. 77, my translation). Consequently, a range of measures were initiated to improve the situation, several of which related specifically to assessment. Among these were the establishment of a national quality assessment system (the “NKVS”) in 2004 with a particular focus on accountability measures, the introduction of the Knowledge Promotion curriculum reform (LK-06) in 2006, a revision of the Regulations to the Education Act in 2009, introducing a distinction between formative and summative assessment, and focus on AfL as a prioritized area in education (Meld. St. 20 (2012-2013), 2013). In addition, calls for more research were made, concerning both theoretical analyses and empirical investigations of assessment practices in schools (Thronsen et al., 2009). The present thesis is a response to these calls.

1.3.2 English in Norway

English holds a strong position in Norwegian society. Since 1969 it has been a compulsory school subject for all. Norwegians are widely exposed to English both at school and in society at large, and people use it for a number of different purposes across a range of different contexts (Chvala & Graedler, 2010; Simensen, 2011). Moreover, studies have shown that the proficiency level of the population is generally high compared to other countries in which English is neither the first nor an official language (Education First, 2014, 2015). However, studies have also shown that the proficiency level of the population may be insufficient for meeting the communicative requirements in professional settings (Hellekjær, 2007, 2008, 2012). The educational authorities have given English special status in the subject curricula by no longer subsuming it under the label “foreign languages”. Despite this, they do not explicitly use the label “second language”. Whether English in Norway should be treated as a “foreign” or a “second” language seems to be a matter of preference. Some scholars base the distinction between them on the status accorded to the language in society (e.g. Graddol, 2006, p. 84), whereas others base it on whether the language in question is used as an L1 by
the majority population (e.g. Alderson et al., 2015, p. 71). Yet others see the distinction in itself as somewhat artificial and outdated (Celce-Murcia, 2014). However, in this thesis I will follow Simensen (2014) and refer to English in Norway as a foreign language.

1.3.3 The Norwegian school context

Norwegian children start school at the age of six. Schooling is compulsory at the primary (grades 1-7) and lower secondary level (grades 8-10). Upper secondary school (grades 11-13) is voluntary, but everyone has the legal right to attend. At the upper secondary level students can choose between a general studies programme (GSP) – for students whose primary goal is to continue to the tertiary level – and various vocational studies programmes (VSPs).

The English subject is compulsory in primary and secondary school. At the upper secondary level it is required for all GSP students in their first year and for all VSP students in their first and second years. Both student groups have the same curriculum, albeit with some adjustments for the different study programmes. For example, students in the Building and Construction programme will be expected to handle a specialized vocabulary related to the building and construction domain, in addition to “a wide general vocabulary” expected of students regardless of programme (Norwegian Ministry of Education and Research [KD], 2006/2013). The reason for having a common subject curriculum, which is generally rather academic in its orientation, has been to give all students the opportunity to qualify for tertiary education (Skjersli & Aamodt, 1997). However, this curriculum has repeatedly been criticized for being too ‘theoretical’ and poorly tailored to the VSP students’ needs (Høst, Seland, & Skålholt, 2013; Solberg, 2010; Tarrou, 2010). Traces of this criticism are found in the data gathered for the present investigation. Recently, however, the government has partly acknowledged the critique by initiating projects such as the Vocational Education and Training Promotion (“FYR”) aimed at making the common core subjects, such as mathematics, Norwegian and English, more relevant for vocational students (KD, 2015).

The current English subject curriculum, which was introduced in 2006 and revised in 2013, is loosely based on the Common European Framework of Reference (CEFR) (Simensen, 2010). It specifies a number of learning outcomes or “competence aims”, which guide instruction and define what is to be assessed. The aims are grouped into four “main subject areas”: Language Learning, Oral Communication, Written Communication, and
Culture, Society and Literature (KD, 2006/2013). However, as the aims are many, and some of them are rather general, they need to be operationalized in order to be assessed (Meld. St. nr. 30, 2004, p. 40). In addition to the competence aims, the subject curriculum defines five basic skills, which are common to all subjects in school and which are described as “fundamental prerequisites for learning and development in schools, in the workplace and in society at large” (Norwegian Directorate for Education and Training [UDIR], 2015, p. 5, my translation). The inclusion of oral skills as one of these five basic skills underscores the importance attributed to spoken proficiency in the Norwegian context.

Summative assessment in upper secondary school is predominantly given in the form of overall achievement marks. These marks are awarded by each subject teacher on the basis of various forms of classroom assessment. In the case of the English level studied here, i.e. first year GSP / second year VSP, approximately 20 per cent of the students are also randomly selected to take a written exam, and five per cent are selected to take an oral exam. The educational authorities give no explicit reason why exams administered to only a portion of the students are needed in addition to the overall achievement marks, but the practice may be explained historically as a matter of different assessment traditions existing side by side (see e.g. Lysne, 2006). As the marks awarded are decisive for admission to colleges and universities, the different forms of summative assessment must be regarded as high-stakes.

An interesting distinction between the oral and the written exam, which is of relevance here, regards their administration. The written exam is administered nationally by the Norwegian Directorate for Education and Training, which provides exam tasks, written rating scales and assessment guidelines nationwide. The oral exam, on the other hand, is managed by the local educational authorities (LEAs) through the county governors in each of the 19 counties. Some of these LEAs provide rating scales, exam tasks and rater training for teachers, but in many cases they leave it to the individual schools to decide in these matters. In turn, some schools leave it to the individual teachers to handle the assessment procedures. Consequently, there are no common national exam tasks or rating scale. This apparent incongruity between the written and the oral exam may be partly be explained in terms of a long tradition of the local level having a strong position in the management of school policies.

---

5 Minor revisions to the curriculum were made in 2013, just after the data for Article 1 had been collected. One of the most important ones was the division of the previous main area “Communication” into “Oral communication” and “Written communication”, thus emphasizing the importance of, and differences between, writing and speaking. The 2013 version can be found in Appendix 1, the 2006 version can be found in Appendix 2.
which was reinforced with the K-06 curriculum reform in 2006 (Sandberg & Aasen, 2008). In fact, the Norwegian Directorate for Education and Training emphasizes the importance of the subject curriculum being adapted locally in everyday teaching and assessment practices so as to “promote adapted education” (UDIR, 2014d, pp. 5, my translation). More generally, however, the difference in the administration between the oral and written exams may also be said to reflect the afore-mentioned tension between the measurement and the assessment paradigms. Arguments for more standardization, as manifested in the written exam, can be supported with reference to the measurement paradigm. Arguments for less standardization, on the other hand, which is demonstrated in the oral exam, can be supported by assessment paradigm thinking.

The lack of a national rating scale for oral English in the Norwegian system is of particular interest in the present thesis, as it appears to be taken as given in international language test design that a language test should be accompanied by a common rating scale (Fulcher, 2003; Ginther, 2013; Luoma, 2004). Rating scales are considered invaluable tools for raters in helping them to focus on those aspects of the performance which the test is intended to measure. As Fulcher (2012) has noted, the rating scale can be seen as the operationalization of the construct to be tested. For example, if the assessment is intended to test pronunciation, this should be specified in the rating scale. If not, it should be left out. A number of studies have investigated rater variability in test situations where rating scales exist; a considerably smaller number of investigations have studied tests without rating scales (e.g. Brown, Iwashita, & McNamara, 2005). In both cases, there is evidence that raters have somewhat different conceptions of the construct to be assessed. In any case, assessment contexts with no common scales are special, and they beg for closer scrutiny.

1.4 Research purpose and aims
The present thesis investigates rating processes and outcomes in an oral English exam at the upper secondary school level. This exam is administered to GSP students in their first year (GSP1) and VSP students in their second year (VSP2). The main focus is on teachers’ understanding of the constructs to be tested.6 As a part of this inquiry, rater behaviour in

---

6 This is intrinsically linked to the specific aspects of performance that the teachers focus on during assessment. In order to describe this focus I will use the terms “teacher (rater) orientations” “teacher (rater) perceptions” and “teacher (rater) cognition” (cf. Brown et al., 2005).
terms of grades awarded is also examined, as well as correspondence between the teachers’ notions of construct and the intended construct as specified in the subject curriculum and related documents. The three studies that have been undertaken have had the following foci:

- Article 1 has examined teachers’ general perceptions of what should be tested in the GSP1/VSP2 oral English exam. This has included a brief analysis of scoring behaviour (i.e. grading) and a comparison between aspects of teachers’ orientations and the construct to be tested according to the English subject curriculum and other defining documents.

- Article 2 has investigated teachers’ orientations towards the assessment of various aspects of pronunciation in the GSP1/VSP2 oral English exam.

- Article 3 has explored teachers’ understanding of how to assess subject content in the GSP1/VSP2 oral English exam, and compared their assessment foci with what the subject curriculum stipulates with regard to content.

The present study is a response to the calls for more assessment research in the Norwegian educational context (cf. section 1.3.1, above) by providing empirical evidence of what happens in the rating process. In this sense, the study contributes to the evaluation of assessment quality in the GSP1/VSP2 oral English exam. In addition, the study more generally provides information on rating processes in EFL school contexts at the upper-intermediate proficiency level.

1.5 A note on terminology

Bachman and Palmer (2010, pp 19-21) use the terms test, assessment, measurement and evaluation more or less synonymously to describe the practice of collecting and evaluating evidence about learning in order to make judgements (c.f. Tina Isaacs et al.’s, 2013, definition on p. 1, above). In this thesis I follow their use of the terminology. However, I am well aware that these terms are used with different meanings and connotations, not least because of their association with the two paradigms outlined in section 1.2, above. Hence, a brief explication of some definitions will follow.

Generally, it may be said that assessment is a broader term than test and evaluation, and that the former subsumes the latter two (Kunnan, 2004, p. 1). Tina Isaacs et al. (2013) explain assessment in relation to learning, but it could also be explained in relation to abilities
or behaviour generally, or even on a macro-level, such as an educational programme. A test, on the other hand, is typically seen as a more systematic and rigorous form of information gathering, normally restricted by a predetermined time frame (Green, 2014, p. 6). Evaluation is sometimes regarded as the “use” of assessment, for example in the evaluation of an educational programme, whereas measurement is often associated with the gathering of quantitative data according to explicit rules and procedures (Bachman, 2004, pp. 8-9). While recognizing these terminological differences, I would still argue that the collection and evaluation of evidence about students’ abilities and behaviour can be referred to as assessment, testing, measurement and evaluation in the context of this study, in line with Bachman and Palmer (2010).

1.6 The structure of the thesis

This thesis is divided into two main parts. Part I contains the extended abstract, and Part II comprises the three articles which report on the investigations undertaken. The extended abstract consists of five chapters. While the present chapter situates the study by providing a general introduction, Chapter 2 explains the theoretical framework for the thesis. This framework is largely based on theories from the fields of educational and psychological measurement and applied linguistics. The rationale for using these theories is that they provide relevant conceptualizations for understanding the nature of the phenomena being studied, i.e. assessment processes and outcomes in the GSP1/VSP2 oral English exam. Chapter 3 reviews relevant research literature on rater cognition and rater behaviour, both in educational and non-educational contexts, and both internationally and in Norway. The purpose of the review is to identify the space in which the present study gives a research contribution. In Chapter 4 the research design and the methods used are outlined, including a presentation of the research questions, participants, data and analyses, as well as a discussion on the appropriateness of the methods chosen for the different research questions. In addition, I consider aspects of research validity and ethical considerations regarding the investigation. Finally, in Chapter 5 I discuss the main findings of the three articles, including their interrelatedness and the extent to which they have responded to the overall research aims and purpose of the study. The chapter ends with a number of implications for assessment and instruction and some suggestions for future research.
Chapter 2: Theoretical framework

2.1 Introduction to the chapter
In this chapter I discuss the theoretical framework of this thesis. As the main focus is on rating processes in a high-stakes, oral examination, I have found it relevant to use a number of conceptualizations developed in the field of educational and psychological measurement, typically applied in standardized, large-scale testing. Examples of such conceptualizations are ‘construct’, ‘validity’ and ‘reliability’. The use of this terminology reflects a pragmatic stance on the relevance of these conceptualizations for the object of study. This is also related to the fact that the summative nature of the oral exam makes it “test-like” (Erickson, 2014, p. 50). Thus, the decision to use this theoretical framing is consistent with a pragmatist epistemological position, which holds that concepts are to be understood as tools for understanding the phenomena we want to study (Hookway, 2015).

In the following I will start by explaining the concept of construct, before moving on to an exposition of the notions of validity and validation. I continue by outlining some perspectives on reliability and standardization. Finally, I discuss the concepts of oral communication and content, as they represent the main components of what should be tested in the oral exam under scrutiny.

2.2 Constructs: Operationalization of underlying abilities
As established in section 1.1, assessment can be seen as the collection and interpretation of evidence about learning in order to form a judgement about that learning. Regardless of test purpose, a very central concern in testing and assessment is what one is trying to form a judgement about. In test theory this ‘what’ is commonly referred to as “attributes”, “traits” or “constructs” (Fulcher, 2015, p. 127; Kane, 2006, p. 30; Newton & Shaw, 2014, p. 10). According to Weir (2005), constructs are the “underlying […] abilities we wish to measure in students” (p. 1). An example of such an ability, taken from the CEFR, is lexical competence (Council of Europe, 2001, p. 110).

The choice of label for these abilities is a contentious issue (Fulcher, 2015; Kane, 2012). In this thesis I use the term “construct”, rather than “attribute” or “trait”, because I find that it aptly points to the constructed and abstract nature of the phenomena being investigated,
such as, for instance, lexical competence.\textsuperscript{7} The justification for this view is that an unobservable concept such as lexical competence is an abstract notion, which can only be assessed after having been operationalized (Fulcher & Davidson, 2007, pp. 369-370). Thus, in order to assess lexical competence, one would have to identify observable properties which can serve as indicators of this construct. Examples of such properties are “sentential formulae” (“How do you do?”, “Good morning!”) and “phrasal idioms” (“He kicked the bucket”, “It’s a long shot”) (Council of Europe, 2001, p. 110).

Quite frequently, a construct and its observable indicators will form a larger whole, together with other constructs and indicators in a more or less unified theory. In the CEFR, for instance, lexical competence is a construct within a model of communicating language competence, which is logically linked to a number of related constructs, such as grammatical competence, semantic competence and phonological competence (Council of Europe, 2001, p. 109). In turn, this model is built on theories of communicative competence (North, 2014).

One of the reasons for the disagreement over the use of terminology is that the term “construct” is used with so many different meanings that it may be difficult to know what it refers to (Kane, 2012, p. 67). Moreover, theorists disagree on the ontological nature of constructs. Measurement specialists who subscribe to a realist world view, for example, typically see constructs as psychologically real entities, which exist in the minds of individuals and which may cause variation in behaviour (e.g. Borsboom, Cramer, Kievit, Scholten, & Franić, 2009, p. 150). Theorists who subscribe to an antirealist position, on the other hand, question the existence of constructs as ‘real’ attributes of the mind. To antirealists, they are first of all theoretical ideas, constructed by the research community, which are meant to describe and explain patterns of behaviour (Newton & Shaw, 2014, p. 164). From a realist perspective, there is no point in trying to measure constructs (i.e. theoretical ideas) since they cannot cause variation in behaviour if they do not exist. Borsboom et al. (2009), for example, therefore suggest that the construct label is replaced by the term “psychological attribute”, which can be regarded as a property that “plays a role in psychological reality” (pp. 150, 152).

A third ontological position, referred to as pragmatic realism (Fulcher, 2015), holds that a construct (such as lexical competence) is real if the operationalizations of the construct “can be observed, and if they vary in ways predicted” (Fulcher, 2014, p. 1447). According to this view, a construct can be seen as:

\textsuperscript{7} In Article 1 I use the term “criterion” as an auxiliary concept. This is defined as “aspects of performance to be assessed”. The reader is referred there for a further discussion on the use of this term.
The abstract name for a complex idea derived from observations of co-occurring phenomena, the purpose of which is to explain the coherence of our perceptions and make predictions about the likelihood of future states or events. The names are ‘the signs of our ideas only’, but no less real for that. (Fulcher, 2015, pp. 129-130)

By keeping a dual focus on the existence of both theoretical constructs and observable indicators, Fulcher occupies a middle position between (extreme) realist and antirealist positions.

It follows from my pragmatist epistemological position (cf. section 2.1) that it would not greatly matter which label I choose for the aspects that teachers attend to when assessing performance. Still, I find that the notion of underlying theoretical constructs, operationalized in terms of observable properties (Fulcher & Davidson, 2007), aptly describes what is to be assessed in the Norwegian context. As the English subject curriculum – which forms the basis for instruction and assessment – is based on theories of communicative competence through its influence from the CEFR (cf. section 1.3.3), it explicitly and implicitly uses theoretical constructs which need to be operationalized. Examples of such constructs are “speaking strategies” and “fluency” (Norwegian Ministry of Education and Research [KD], 2006/2013). For example, fluency cannot be observed directly in students, but must be inferred on the basis of properties such as “pauses”, “fillers”, “false starts” etc. (Brown et al., 2005, p. 23).

Furthermore, the use of constructs as analytical tools in the present context fits well with “Bachman & Palmer’s (2010) claim that a construct is defined on the basis of a “frame of reference” (pp. 212-213). This frame of reference may be a theory of language, a syllabus, a needs analysis, or a combination of the three. As already mentioned, it is the English subject curriculum which above all informs teaching and assessment in the Norwegian context. However, communicative theories also play a part through their influence on curriculum development. In addition, there are government documents, such as circulars, which specify what goes into the construct and what does not. For instance, in Norway a circular specifically states that a student’s “effort” is not to be assessed (UDIR, 2010, p. 13). Hence, in Norway the frame of reference for the construct definition is the English subject curriculum, communicative theories and government directives.

---

8 This circular was replaced by a revised one in 2014, after the main bulk of the data for this thesis had been collected. In the new circular (UDIR, 2014b), the reference to “effort” has been omitted.
2.3 Validity and validation

2.3.1 Evaluating evidence about score interpretations

In section 1.1 I pointed out that validity is commonly regarded as a fundamental concern in assessment, sometimes referred to as the quality or ‘soundness’ of an assessment procedure. However, the concept is multifaceted and complex and its meaning has evolved over the years. Although some agreement can be found today, not all theorists interpret the concept in the same way (Newton & Shaw, 2014, pp. 7-9). In addition, it should be noted that the concept of validity in this thesis is discussed against an educational backdrop, where there may be said to be tensions between assessment and learning (Baird et al., 2014, p. 97). More broadly, these tensions are echoed in the measurement and the assessment paradigms (cf. section 1.2), which affect assessment theory and practices in different ways.

The classic definition of test validity concerned the extent to which a test “measures what it purports to measure” (McCall, 1922, quoted in Anthony Green, 2014, p. 75). According to this view, validity is seen as a property of the test itself. Some measurement specialists (e.g. Borsboom et al., 2009) still adhere to this notion of validity, but to most authors it is no longer tenable (Bachman, 2014; Fulcher, 2015; Anthony Green, 2014; Kane, 2013; Newton & Shaw, 2014). A typical argument for rejecting the classical definition is that no matter how well-developed a test is, it would not measure what it is supposed to measure if it is poorly administered or used in contexts for which it was not intended (Newton, 2012, p. 3). Hence, the ‘consensus’ view today holds that validity is not a property of a test, but of the inferences that are made from assessment results (Fulcher, 2015; Newton & Shaw, 2014; Wiliam, 2008). In the Standards for Educational and Psychological Testing (henceforth: Standards) the concept is defined in the following way:

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


In passing, it is worth commenting on the term “validation” in the above quotes. According to Davies and Elder (2005) validity refers to the theoretical aspect of assessment quality, whereas validation is the actual practice of evaluating the quality of a test.
Two aspects of the Standards definition are particularly important for the present thesis. The first relates to the formulation ‘interpretations of test scores’, which essentially concerns score meaning⁹. The notion of score meaning raises a host of questions: For example, in the oral English exam under investigation one may ask what the mark 3 means. According to the Norwegian Directorate for Education and Training the numerical mark 3 means “fair degree of competence in the subject” (UDIR, 2009, p. 2). However, one could continue to inquire: “In relation to what? That is, what kind of competence has been assessed?” According to the Regulations to the Education Act, it is the competence aims of the subject curriculum which form the basis for assessment (KD, 2006/2015). However, not all of these competence aims are relevant for the oral exam (e.g. those that relate to written proficiency). One may therefore continue to probe: “Which competence aims have been tested? How have they been operationalized? What kind of performance has the student given that is indicative of goal attainment with regard to the competence aims being tested?” etc.

The second aspect to consider in the Standards definition concerns the importance attributed to the collection and interpretation of evidence as a central element in validation. In order to make sure that the interpretations from test scores are valid, one has to gather and analyse information about the different aspects of the assessment process, such as task design, scoring procedure, rater bias etc.

These two aspects of validity and validation, i.e. score meaning and the collection and interpretation of evidence, are of direct concern in the present thesis. In all three articles I have provided evidence of the teacher raters’ perceptions of score meaning in terms of the constructs to be assessed. In addition, I have gathered and analysed data from the English subject curriculum and accompanying government documents concerning the intended meaning of the scores. Finally, in Article 1 I have also investigated rater consistency by asking teachers to score student performance.

2.3.2 The unitary theory of validity and argument-based validity approaches
In order to understand more fully how the interpretation of scores can be evaluated, it is worth considering two frameworks which have been very influential in educational and psychological assessment. These are the unitary theory of validity and the argument-based validity approach.

⁹ In this thesis I follow Messick (1989), who uses the term “score” in a very general sense (p. 13). This means that it does not only reflect numerical ratings, but also, for example, verbal descriptions of scoring outcomes.
In his unitary theory of validity, Messick (1989) developed a validity framework which puts primary emphasis on the construct to be assessed. Messick defines validity as:

>a]n 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, p. 13, emphasis in the original)

In order to be able to draw sound inferences from score meaning, Messick maintains, it is important that the results represent, as accurately as possible, the intended constructs. However, this is difficult in practice, as measurements will typically “either leave out something that should be included according to the construct theory or else include something that should be left out, or both” (p. 34). To conceptualize this, he borrowed two terms introduced by Cook and Campbell (1979). Aspects of the construct which are left out are labelled construct underrepresentation, and aspects which are not supposed to be included are referred to as construct-irrelevant test variance (Messick, 1989, p. 34). Returning to the example of lexical competence, one could say that failure to assess a student’s knowledge of idioms would indicate construct underrepresentation, whereas the assessment of the ability to use adjectives and adverbs correctly would signify construct-irrelevant variance. Validation studies should therefore collect and analyse evidence to identify such validity threats. This evidence could come from virtually any source, and Messick advocates the collection and interpretation of as many sources of evidence as possible (Messick, 1989, p. 35).

One potential problem with the unified theory of validity, however, is its applicability, since it is conceptually very complex. As Messick himself attested to, a consequence of this complexity is that validation studies will require paying attention to a very large number of different questions in order to provide good validation evidence (Messick, 1996, p. 7). The broadening of the scope of validation to such an extent has led some validation practitioners to regard the theory as impractical for application in test evaluation (see e.g. Baird et al., 2014, p. 79). Alternative approaches have therefore been developed. One such approach is argument-based validation (Cronbach, 1988; Kane, 2006; Mislevy, Almond, & Lucas, 2003).

Argument-based validation distinguishes itself from the unitary theory of validity in not taking theory-based constructs as the starting point for the evaluation of inferences. Rather, as the name suggests, it uses an argument to clarify the reasoning of the proposed interpretations and uses of scores (Kane, 2013, p. 8). This argument typically consists of claims for the inferences to be made from the scores, warrants to support the inferences,
backing evidence, and alternative hypotheses as rebuttals to the claims. Simply put, this means that it is possible to infer directly from an observation to a claim, without reference to a construct. For instance, on the basis of the observation that a test taker’s speech is unintelligible, one may make the claim that he or she is unfit for working as a teaching assistant at university.

However, as Kane makes clear, the structure of the argument would depend on the attribute to be assessed, and on the claim to be stated (Kane, 2012, p. 68). In cases where the trait is a theoretical construct, defined in terms of an underlying theory, the argument-based approach will be similar to construct validity approaches, such as the unified theory of validity (Kane, 2013, p. 9). However, in most cases, there is no need to invoke notions of construct, since the trait to be assessed will be an “observable attribute” that is not dependent on an underlying theory (Kane, 2012, p. 68). This attribute can then be assessed directly. As Kane (2013) points out, it would be possible to test someone’s skill in servicing computers without making assumptions about an underlying “‘computer-servicing’ trait” (p. 21).

Kane (2006) distinguishes four main types of inferences in an interpretive argument: scoring inference, generalization inference, extrapolation inference and implication inference. The first one is the most relevant in the present thesis. The scoring inference concerns the assignment of scores to a test taker’s performance according to a “scoring rule, which provides the warrant for the scoring inference” (Kane, 2006, p. 34). The scoring rule includes aspects such as the purpose of the assessment, the types of tasks included and the criteria to be applied. In order to evaluate the appropriateness of the scoring inference several kinds of evidence can be used. For example, empirical evidence from the rating process can be used “to check on the consistency (e.g., inter-rater reliability) and accuracy (e.g. quality control data) of scoring” (p. 34). The specification of different types of inferences in this way is meant to provide guidance on what kinds of evidence that are needed for validation.

To sum up, both the unitary theory of validity and the argument-based approach provide valuable conceptualizations for investigating rater perceptions (and behaviour) in the GSP1/VSP2 oral English exam. Both emphasize the collection and analysis of evidence for the evaluation of the inferences that are made from score meaning. In addition, the unitary theory of validity brings to the analysis the concepts of construct underrepresentation and construct-irrelevant variance, which are appropriate for analysing the extent to which the teachers attend to the things that they should attend to. The argument-based approach helps narrow the focus by pointing to the types of evidence needed for the analysis of the scoring process. In addition, even if the argument-based validity approach de-emphasizes the role of
construct, it does not eliminate it, provided that it is based on some underlying theory. In the case of communicative language assessment, such as in Norway, there is reason to argue that there is an underlying theory for the attributes to be tested (cf. section 2.2).

The features of validity and validation discussed in this section concern the technical aspects of assessment quality, and are by far the most important for the present thesis. However, both the unitary validity theory and the argument-based approach include a broader concern as well – which is also reflected in the analyses I have undertaken – namely the social consequences of test use. In the next section I therefore turn to this issue.

2.3.3 The social consequences of test use

Messick’s (1989) concern for the social consequences of test use implied an evaluation of the value implications of score interpretation, as well as issues such as fairness and test impact. For example, in a validation study it could be relevant to evaluate the impact that the use of a test would have on teaching, learning and instructional materials (Cumming, 2013, p. 6008). Although this issue is not directly addressed in the present study, the question of values surfaces in the analyses of my data. For example, some of the teachers interviewed report that they experience the assessment system as being unfair to some students. Consequently, they deliberately score the students that they perceive as being disadvantaged more leniently than other students. This type of value judgement influences the interpretations that the teachers make, in the sense that it negatively affects the consistency, or reliability, of the scores.

This last point is also interesting in the sense that both Messick and Kane integrate the issue of reliability, or score consistency, into their frameworks (Kane, 2006; Messick, 1989). Since reliability may be affected by test administration and rating procedures and is relevant in a discussion on the amount of standardization needed in an assessment context, I now turn to a consideration of these issues.

2.4 Reliability and standardization

According to Harlen (2012), assessment used for summative purposes requires some form of quality assurance procedures in order to ensure that the scores are reliable, or dependable. “The more weight that is given to the summative judgment”, she says, “the more stringent the quality assurance needs to be” (p. 97). This would mean that some form of standardization of the assessment procedure is required. However, as Harlen’s quote indicates, such
standardization is a matter of degree. The question is then: How standardized do the assessment procedures need to be?

In large-scale testing, rating scales and rater training are regarded an important part of test development and administration and rarely, if ever, dispensed with. As Standard 4.18 of the Standards specifies: “[I]n scoring more complex responses, test developers must provide detailed rubrics [i.e. rating scales] and training in their use” (p. 91, italics added). However, the same Standards also make clear that “more flexibility in the assessment procedures” may be considered in order to better capture complex constructs “that may be otherwise difficult to assess”, even if this jeopardizes reliability (AERA, APA, & NCME, 2014, p. 36). Although the Standards here do not mention the abandonment of rating scales or rater training as examples of greater flexibility, they do not exclude the possibility of dispensing with them either, or as in the Norwegian setting, to create locally developed scales. The fact that assessment in a curriculum-based context is intrinsically linked to teaching and learning makes the oral exam studied here somewhat different from testing in non-educational contexts. In such contexts, researchers have advocated greater procedural flexibility. According to Moss (1994), for example, privileging standardization in order to enhance reliability may come into conflict with good teaching and learning, since there are “certain intellectual activities that standardized assessment can neither document nor promote” (p. 6).

Finally, in the discussion of standardization and the use of a common rating scale, it may be argued that although rating scales provide a means for guiding raters in the assessment process, they do not necessarily reflect the complexity of the constructs to be tested (Taylor & Galaczi, 2011). Logically enough, rating scales need to be simplified in order to make them usable to raters. This seems particularly obvious in the case of assessment based on a comprehensive subject curriculum. Hence, rating scales may be criticized for not giving a realistic representation of the features to be assessed (Lumley, 2005). As the Standards alludes to, there is a tension here, reflecting preferred emphasis: should reliability be stressed, or is it better to go for a looser structure in order to better capture the constructs?

2.5 Constructs to be tested: Oral communication and content

The oral English exam under investigation is markedly different from traditional language tests in the sense that the English subject curriculum, in addition to language-specific constructs, also specifies a number of content-related issues that are to be tested (cf. section 1.3.3). This means that two main constructs may be identified: oral communication and
In the following, I will briefly outline some theoretical perspectives on these two constructs and link them to the corresponding competence aims of the subject curriculum.

2.5.1 Models of communicative competence and the oral communication construct

Models of communicative competence generally aim to describe and explain what it means to know and use a language for communication (Purpura, 2008). Several communicative models have been developed. The perhaps most well-established ones are Canale and Swain’s model (Canale, 1983; Canale & Swain, 1980) and Bachman’s model (Bachman, 1990; Bachman & Palmer, 1996). The former describes communicative competence as consisting of four components: grammatical competence, sociolinguistic competence, pragmatic competence and strategic competence. Bachman’s framework is an expansion of Canale and Swain’s and other earlier models in that it explicitly “attempts to characterize the processes by which the various components interact with each other and with the context in which language use occurs” (Bachman, 1990, p. 81). Communicative competence, or “communicative language ability” (CLA), is described by Bachman as comprising two main components: (i) language knowledge, such as grammatical knowledge, textual knowledge, functional knowledge and sociolinguistic knowledge; and, (ii) strategic competence, defined as “a set of metacognitive components, or strategies, which can be thought of as higher order executive processes that provide a cognitive management function in language use” (Bachman & Palmer, 1996, p. 70).

Despite their currency, these models have attracted criticism from various quarters. Some authors have claimed that they are too simplistic, not accounting sufficiently well for all the different elements which affect communication, particularly as regards contextual factors (McNamara, 2003). Bachman’s model, for example, taking a cognitive perspective on language ability, sees the construct as something residing in the individual. From an interactionist point of view, however, this is a too narrow perspective, as communicative competence is understood as being more explicitly shaped by contextual features, such as physiological setting, participants and tasks (Chalhoub-Deville, 2003; Chapelle, 1998; He & Young, 1998). As Chalhoub-Deville (2003) argues, “the ability components the language user brings to the situation or context interact with situational facets to change those facets as well as to be changed by them” (p. 372, italics added). On this view, language ability is seen as co-constructed by the participants of the interaction in local settings. This is an interesting position which is relevant in a discussion on oral exams in Norway, where the exams are

---

10 Cf. appendices 1 and 2.
administered on the local level, and where no national rating scales exist. However, as both Bachman (2007) and Chalhoub-Deville (2003) have pointed out, such a position may cause problems in language assessment, as it makes generalizations across contexts difficult.

In addition to the criticism levelled against communicative models for their lack of complexity, there are also those who have found the models too complex for application in test situations (c.f. Harding, 2014). As a practical solution to this problem, Harding points to the use of frameworks such as the CEFR, which may function “as an accessible de facto theory of communicative language ability” (p. 191).\textsuperscript{11} For this reason, and because CEFR has been influential on curriculum development in Norway, it is worth briefly considering some of its main features relevant for the assessment of oral communication.

As was mentioned in section 2.2 the CEFR lists a number of factors which contribute to a language user’s “communicative language competence”. The three main features are: (i) linguistic competences, (ii) socio-linguistic competences; and (iii) pragmatic competences. In Figure 1 these three competences and their sub-components have been listed.

\textbf{Figure 1.} CEFR’s model of communicative language competence (Council of Europe, 2001).

\textsuperscript{11} Cf. Fulcher’s (2004, 2010) critique against CEFR’s lack of a theoretical basis.
In addition to the elements listed in Figure 1, mention is also made of *fluency* and *propositional precision*, the latter defined as “generic qualitative factors which determine the functional success of the learner/user” (p. 128). Beyond this, the CEFR makes clear that there are additional competences contributing to a language user’s ability to communicate. Three such competences are mentioned: “knowledge of the world”, “sociocultural knowledge” and “practical skills and know-how”. Furthermore, the CEFR refers to an additional concept labelled “production strategies”, which involves “mobilising resources, balancing between different competences – exploiting strengths and underplaying weaknesses – in order to match the available potential to the nature of the task” (Council of Europe, 2001, p. 63). Thus seen, production strategies have affinities with the notion of strategic competence in Bachman & Palmer’s (1996) model.

Beyond the descriptions of communicative competence, the CEFR is relevant in the present thesis in that it provides descriptions of “oral production” and “oral interaction” in a number of proficiency scales. These describe oral ability at different levels of competence. In Table 1 and Table 2 extracts from the two scales describing overall production and interaction have been listed. Note that only the B1 and B2 levels are included, as these correspond to the average proficiency level of the Norwegian students in this investigation.

**Table 1. CEFR production scale for overall spoken production.**

| B2 | Can give clear, systematically developed descriptions and presentations, with appropriate highlighting of significant points, and relevant supporting detail. Can give clear, detailed descriptions and presentations on a wide range of subjects related to his/her field of interest, expanding and supporting ideas with subsidiary points and relevant examples. |
| B1 | Can reasonably fluently sustain a straightforward description of one of a variety of subjects within his/her field of interest, presenting it as a linear sequence of points. |
Table 2. CEFR production scale for overall spoken interaction.

<table>
<thead>
<tr>
<th>OVERALL SPOKEN INTERACTION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B2</strong></td>
</tr>
<tr>
<td>Can use the language fluently, accurately and effectively on a wide range of general, academic, vocational or leisure topics, marking clearly the relationships between ideas. Can communicate spontaneously with good grammatical control without much sign of having to restrict what he/she wants to say, adopting a level of formality appropriate to the circumstances. Can interact with a degree of fluency and spontaneity that makes regular interaction, and sustained relationships with native speakers quite possible without imposing strain on either party. Can highlight the personal significance of events and experiences, account for and sustain views clearly by providing relevant explanations and arguments.</td>
</tr>
</tbody>
</table>

| **B1**                     |
| Can communicate with some confidence on familiar routine and non-routine matters related to his/her interests and professional field. Can exchange, check and confirm information, deal with less routine situations and explain why something is a problem. Can express thoughts on more abstract, cultural topics such as films, books, music etc. |
| Can exploit a wide range of simple language to deal with most situations likely to arise whilst travelling. Can enter unprepared into conversation on familiar topics, express personal opinions and exchange information on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events). |

The descriptors listed in Table 1 and Table 2 indicate that a number of communicative language competences are involved in oral production and interaction. For example, the references to coherence, fluency, grammatical control and level of formality point to linguistic, socio-linguistic and pragmatic competences (cf. Figure 1). Moreover, the references to different types of topics indicate that other competences such as knowledge of the world and sociocultural knowledge are also important in actual language use. In addition to these two overall scales, the CEFR includes a number of more detailed scales for “public announcements”, “conversation”, “informal discussions” and “phonological control”, to mention a few.

2.5.2 Theoretical perspectives on content

According to Met (1998), language education may be regarded as a continuum from language-driven approaches to content-driven approaches. In Norway the English subject taught at the GSP1/VSP2 level may be located somewhere in the middle of this continuum. Competence aims such as “[The student shall be able to] discuss literature by and about indigenous peoples in the English-speaking world” (cf. Figure 2, below) attest to this claim.
In other words, they reveal that content is an important part of the construct to be taught and tested.

In models of communicative competence language aspects are distinctly more elaborately described than are content aspects. Apart from sketchy descriptions of “topical knowledge” in Bachman and Palmer (1996) and general references to “world knowledge” and “socio-cultural knowledge” in the CEFR (cf. section 2.5.1), there is, in fact, little theoretical support to be found for the analysis of content in the GSP1/VSP1 oral English exam. However, relevant conceptualizations can be found in the field of content-based instruction (CBI), which is common in curriculum-based, second language instruction in the U.S. In CBI subject-specific curricular content – such as language arts or social science – is typically taught alongside a second language in order to improve students’ skills and abilities in both areas (e.g. Brinton, Snow, & Wesche, 2003; Chamot, 2009; Snow & Katz, 2014). In addition, valuable theoretical perspectives on content may be found in Bloom’s revised taxonomy, which has affinities with Norwegian subject curricula through its focus on learning objectives (Anderson & Kratwohl, 2001).

In Chamot’s (2009) CBI-framework content is referred to as facts, concepts, laws, principles and theories. An interesting point that she makes, which I have also found traces of in statements from the teachers in my material, is that understanding subject matter concepts and the relationships between them is a prerequisite for developing academic knowledge (p. 20). Moreover, Chamot stresses the importance of higher-order thinking skills, such as being able to analyse, reflect, predict and synthesize (p. 30). A similar emphasis is found in the English subject curriculum, where a number of competence aims in the content area include terms such as “assess”, “discuss” and “elaborate on” (cf. Figure 2).

Chamot’s exposition points to a separation of the content construct into two main components. On the one hand, there are the subject matter elements, such as facts, concepts, laws etc., and on the other, there are the thinking skills or abilities, which are needed to handle subject matter. This perspective is even clearer in Bloom’s revised taxonomy, where learning outcome objectives are arranged along a knowledge dimension and a skills and processes dimension (Anderson & Kratwohl, 2001). Accordingly, the content construct consists of a what-dimension (subject matter knowledge) and a how-dimension (skills, abilities, processes). The former is divided into “factual knowledge”, “conceptual knowledge”, “procedural knowledge” and “metacognitive knowledge”. The latter is structured hierarchically from the simple to the complex in the following order: “remember”, “understand”, “apply”, “analyse”, “evaluate” and “create” (Kratwohl, 2002, p. 216).
2.5.3 Competence aims which relate to oral communication and content

Turning to the subject curriculum, there are a number of competence aims which underlie the two overall constructs to be tested in the oral English exam. Figure 2 gives an illustration.

**ORAL COMMUNICATION**

*(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 [...] presentations about general and specialized themes related to his/her own education programme
- express him/herself [...] orally in a varied, differentiated and precise manner, with good progression and coherence
- select and use appropriate [...] listening strategies to locate information in oral and written texts
- select and use appropriate [...] speaking strategies that are adapted to a purpose, situation and genre
- take the initiative to begin, end and keep a conversation going

**CONTENT**

*(The student shall be able to:)*

- exploit and assess various situations, working methods and strategies for learning English
- describe and evaluate the effects of different verbal forms of expression
- assess and comment on his/her progress in learning English
- select and use content from different sources independently, critically and responsibly
- select an in-depth study topic within his/her own education programme and present this
- use technical and mathematical information in the media
- discuss social/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

**Figure 2.** Competence aims underlying the two constructs to be tested; taken from the English subject curriculum, GSP1/VSP2 level.
It should be noted here that the competences aims listed in Figure 2 are taken from the previous version of the subject curriculum (2006), which was the governing document when the first data was collected in 2012. The curriculum was slightly revised in 2013. As can be seen in Figure 2, the oral communication construct draws on competence aims which not only include production skills, but also reception skills, i.e. listening. In this sense the Norwegian subject curriculum follows the CEFR, which treats these skills as overlapping (Council of Europe, 2001, p. 92). Despite the rather general formulations characterizing these competence aims, the references to “vocabulary”, “coherence” and “speaking strategies” give clear indications of linguistic, socio-linguistic and pragmatic competences as defined by the CEFR (cf. Figure 1, above). In addition, expressions such as “precise manner” and “progression” (i.e. fluency) point to the notions of fluency and propositional precision mentioned in section 2.5.1, above.

In terms of the content construct, the competence aims relating to this trait have a very wide scope, ranging from rather conventional English subject content matter, such as being able to “discuss and elaborate on English texts from a selection of different genres” to rather unconventional ones, such as “exploit and assess various situations, working methods and strategies for learning English”. However, in line with Bloom’s revised taxonomy, they both relate to skills or processes (other than communicative language competence), on the one hand, and subject matter knowledge, on the other. For example, being able to assess strategies for learning English, has to do with subject matter knowledge in the field of metacognitive strategies. And, in order to be able to assess such strategies, one also has to be able to understand, use, apply and analyse them (cf. section 2.5.2, above).

Two more comments are worth making. Firstly, there are eight competence aims listed in the subject curriculum, which have not been included in Figure 2. These involve skills such as reading, writing and using digital aids (cf. Appendix 2). Secondly, it is, of course, virtually impossible to test all these aims in a single oral exam. Hence, some aims may be excluded when the construct is to be operationalized. This may be a potential problem if teachers, who both design test tasks and evaluate performance, consistently disregard some of these aims.

12 Cf. footnote 5, p. 7 and appendices 1 and 2.

13 «Progression» is an erroneous translation of the Norwegian word “flyt” (cf. the Norwegian version of the curriculum, Appendix 3, bullet point 5 under the heading “Kommunikasjon”).
2.6 Short summary
The theoretical aspects of assessment discussed in this chapter provide a relevant framework for understanding the phenomena studied in this thesis, i.e. teacher rater perceptions (and behaviour) in the Norwegian EFL context. Moreover, they also afford valuable tools for evaluating the quality of the rating processes. Overall, the notion of “construct” offers a useful conceptualization for understanding what aspects teachers attend to while rating oral performance. In addition, the concept of “validity” and its related notions of “construct underrepresentation” and “construct-irrelevant variance” guides the attention towards the aspects of student performance that should be attended to according to the curriculum and accompanying guiding documents. This is further supported by the idea of scoring inference, which emphasizes the importance of clarifying the reasoning of the proposed interpretations of the scores. Furthermore, as the analyses undertaken in this thesis give indications of social value judgements which affect the teachers’ ratings, it is also relevant to discuss this feature in light of the social consequences of test use. Finally, the theoretical perspectives on oral communication and content provide a useful backdrop for analysing the teachers’ orientations towards the two main constructs to be assessed in the GSP1/VSP2 oral English exam.
Chapter 3: Literature review

3.1 Introduction to the chapter
In this chapter I situate the present investigation in the assessment research context. The review supplements and expands on the review sections of each of the three articles included in Part II of the thesis. The choice of literature has been guided by: (i) the focus on (teacher) rater orientations (and to a lesser degree on rater behaviour and the validity of the scoring inference); (ii) the fact that the oral exam focuses not only on aspects of language, but also on subject content; (iii) the lack of a common rating scale for the teachers in this exam; and (iv) the Norwegian context.

A systematic search for relevant studies, based on these criteria, was undertaken by browsing databases such as Web of Science, Academic Search Premier, ERIC and Idunn, journals such as Language Assessment Quarterly and Language Teaching, and reference works such as The companion to language assessment (Kunnan, 2014). Both international and Norwegian sources were explored. In addition, Google Scholar searches were conducted to see whether additional studies could be identified. Moreover, the web pages of the Norwegian Directorate for Education and Training and print monographs on language testing and assessment were searched (e.g. Fulcher, 2003).

Beyond the studies located from searches in the sources mentioned, other studies were identified in the references lists of the literature which was examined. An expanded overview of sources explored and search terms used can be found in Appendix 4. All in all, the search yielded a total of 153 sources, 51 of which are considered in the present review. In the following I will first address international rater cognition and rater behaviour research, and then go on to present findings from the Norwegian educational context. As all the three articles include literary review sections, the reader is also referred to those sections for more specific information pertaining to the three studies.

3.2 International research on rater orientations and rater behaviour
3.2.1 Rater variability and the focus of the present investigation
It has long been known that substantial variation in scoring outcomes may be caused by raters rather than by the performance of the test takers (Bejar, 2012). Such variation is commonly referred to as rater variability (McNamara, 1996). Rater variability may take different forms,
such as: (i) differences in the kinds of performance aspects that raters pay attention to; (ii) variations in how raters interpret and apply assessment criteria; (iii) differences in how raters understand and use rating scale categories; (iv) variability in terms of rater severity or leniency; and (v) interactions between raters and test takers, tasks or other facets of the assessment situation (Eckes, 2005, p. 44). In the present thesis it is the first three forms which are of particular concern, as they relate to the question of rater orientations, or rater cognition. However, point number (iv) will be also considered, as the investigation touches on rater behaviour in terms of score consistency and the severity of the raters. The review will predominantly focus on oral assessment research, but where relevant, mention will also be made of written assessment research. Three other points are worth keeping in mind. Firstly, the vast majority of studies reviewed have investigated assessment in general proficiency speaking tests, rather than in curriculum-based achievement tests, as is the case in Norway. Secondly, they have involved different kinds of proficiency levels. Thirdly, only three of them (Brown et al., 2005; Pollitt & Murray, 1996; Yildiz, 2011) have investigated rater variability in L2 speaking contexts without a common rating scale. Hence, the majority of these are not immediately relevant for the GSP1/VSP2 oral English exam investigated in the present thesis.

3.2.2 Rater orientations

International research on rater orientations, or rater cognition, in the area of general L2 oral proficiency assessment is limited (Brown et al., 2005). Early studies on rater orientations tended to show that raters typically paid more attention to linguistic features of performance, especially grammar, than to other aspects (Magnan, 1988; McNamara, 1990). More recently, however, studies have shown that assessors heed a broader range of factors, including content, discourse complexity and functional skills (Ang-Aw & Goh, 2011; Borger, 2014; Brown, 2000; Brown et al., 2005; Sato, 2012). On the whole, there is evidence that raters heed a number of different features of performance, both construct-relevant and construct-irrelevant ones (Douglas, 1994; C.-N. Hsieh, 2011; Joe, Harmes, & Hickerson, 2011; May, 2006; Orr, 2002; Pollitt & Murray, 1996). Examples of construct-irrelevant performance aspects that raters attend to are interest and personality (Ang-Aw & Goh, 2011), effort (Brown, 1995), voice quality (C.-N. Hsieh, 2011) and age and gender (Orr, 2002). In addition, raters may fail to attend to features that are included in the construct of interest, thus underrepresenting it (Ang-Aw & Goh, 2011; Cai, 2015). In the following, I will review the studies by Brown et al. (2005) and Pollitt and Murray (1996) in some more detail, as these are the only international
investigations I have located which explored rater orientations in oral assessment contexts where rating scales were not provided. Moreover, they investigated so-called singleton speaking examination formats, which are also similar to the Norwegian EFL exam format, in that there is only one test taker being examined at a time.¹⁴ I will then go on to identify some common problem areas discovered in related studies.

Pollitt and Murray (1996) investigated examiners’ assessment of performance in speech samples taken from the Cambridge Certificate of Proficiency in English test, which is a high proficiency level oral examination. Five trained raters were presented with pairs of speech samples obtained from five young adult non-native speakers. The raters were first asked to decide which of the performances in each pair was “better” (p. 81). Afterwards, they were asked to verbalize their perceptions of similarities and differences in the different pairs. The results showed that while the raters paid more attention to content, or what was being said, at the higher levels of performance, they attended more to linguistic features and associated notions of ‘correctness’ at the lower levels of performance. In addition, the results showed that the raters’ disagreed on the question of whether comprehension should be part of the spoken performance construct. Pollitt and Murray also found that the judges were influenced by non-relevant criterion elements in their evaluations, such as the personalities, physical attractiveness and cultural backgrounds of the examinees.

Brown et al.’s (2005) study had a dual focus. It explored raters’ perceptions of relevant assessment criteria and investigated the quality of the test takers’ discourse in two oral proficiency assessments. The starting point for the study was the piloting of a TOEFL (Test of English as a Foreign Language) speaking test, which is an advanced proficiency level examination focusing on English for Academic Purposes (EAP). The raters’ orientations (n=10) were investigated using a qualitative research design (i.e. verbal protocol analysis). Overall, the study found that the judges had the same ideas of the main constructs to be tested, but that there was some variation in terms of their attention to the more finely grained performance features. The main aspects that the raters heeded were: “linguistic resources”, “phonology”, “fluency” and “content” (p. 31). Interestingly, it was found that content was a major focus, only surpassed by linguistic factors. The authors speculated that the relative importance attributed to content could be explained by the nature and the level of the test, which was an EAP examination. This would agree with Pollitt and Murray’s (1996)

¹⁴ The singleton format can be contrasted with the paired speaking format, which involves more than one test taker. Hence, in the paired speaking format it is possible to test, for example, interactional skills (see e.g. Ducasse & Brown, 2009, which investigated such a format without a common rating scale).
conclusion that raters pay more attention to content at the higher proficiency levels. A final observation, which is relevant for Article 2 in this thesis, is that in terms of the phonological, syntactic and organizational aspects of the examinees’ speech production, the raters were more concerned with “comprehensibility” and “clarity” than with native-speaker “correctness” (p. 101).

Beyond Pollitt and Murray (1996) and Brown et al.’s (2005) studies, other investigations – where rating scales have been provided – have reported rater variability both in relation to the interpretation and use of criteria, and to the raters’ comprehension and use of the rating scale categories (cf. points number (ii) and (iii) mentioned by Eckes, 2005, above). For example, in studies by Brown (2000), Kim (2015), May (2009) and Orr (2002) it was found that raters had conflicting views on the meaning of the criteria in the scales. Orr (2002) and Eckes (2009) also found that the raters comprehended the rating scale categories in different ways, thereby awarding the same score to different performances, as well as different scores to similar performances. A similar conclusion was reached by Douglas & Selinker (1993, 1994; cited in Douglas, 1994).

The exposition so far makes clear that some studies have documented considerable rater variability (Eckes, 2005; Orr, 2002), whereas others have demonstrated fairly good correspondence in the features that the raters attend to (Brown et al., 2005; see also Borger, 2014). Brown et al.’s (2005) study is of particular interest in this respect, as no scoring guidelines were provided. The question, then, is how can such differences be accounted for?

Different explanations for this variation have been suggested. For example, it may be a matter of rater background characteristics, such as professional background (Brown, 1995; Chalhoub-Deville, 1995), first language background (Kang, 2008; Y.-H. Kim, 2009) and rating experience (Lumley, 2005). However, it may also be related to a number of other variables, such as test tasks, test administration, rating scales and rater training. According to May (2009), for example, one explanation for the confusion over how to interpret the descriptors in the rating scales could be that the criteria in the scales are vague. As for the question of rater training, I will return to this issue in section 3.2.2 below.

More recently, efforts have been made to identify rater profiles in order to better understand rater variability, and to be able to tailor rater training to different needs. For example, Eckes (2009) classified raters into six different rater types, depending on the extent to which they focused on content, correctness, comprehensibility, description, completeness or overall performance. Similarly, Cai (2015) divided assessors into form-oriented, balanced and content-oriented, depending on the degree to which they focused mainly on linguistic
factors, content factors or both. These are interesting findings in the light of the strong focus on subject content in the English subject curriculum in Norway.

Before moving on to the issue of rater behaviour, I find it relevant to return briefly to Brown et al.’s (2005) finding that raters focused more on comprehensibility than on nativeness. A long-standing debate in the language teaching and assessment literature has been the question of speaker norms, and related issues of ‘correctness’ (Davies, 2003; Jenkins, Cogo, & Martin, 2011). Traditionally, the native speaker has been used as the model for production and the standard for assessment, both in terms of pronunciation, grammar, vocabulary and organization (Cook, 1999). However, with the advent of alternative approaches to L2 language teaching, such as World Englishes (Kachru, 1986), the Intercultural Model (Byram, 1997) and English as a Lingua Franca (Jenkins, 2000), there are indications that this is changing, with raters becoming more concerned with intelligibility than with correctness (Brown et al., 2005; Timmis, 2002). The extent of this change is not known, however, as a number of raters, perhaps particularly teachers, prefer a native speaker standard (Coskun, 2011; Deterding, 2010; Jenkins, 2007)

3.2.3 Rater behaviour

Although rater behaviour in terms of grades awarded is not a major focus of the present investigation, I still find it relevant to include a brief review of research in this area in order to provide a backdrop for the subsequent analysis of the Norwegian teachers’ scoring behaviour. Studies have shown that rater consistency in terms of inter-rater reliability is typically found to be high in oral performance tests (McNamara, 1996; Eckes, 2011; see also Fulcher, 2003). There is also evidence that rater training and experience will increase such reliability (Barnwell, 1989; Brown, 2012). In addition, training may improve intra-rater reliability (Weigle, 1998). As for the question of rater severity or leniency, however, studies are less conclusive as to the effect of training. Some investigations have found that it does have a positive impact, particularly with novice and excessively harsh or lenient raters (Davis, 2015), whereas others have found little effect (Eckes, 2011; Lumley & MacNamara, 1995). An additional point worth making here is that reliability ratings appear to increase when two or more raters are involved in the scoring (Henning, 1996). This is relevant in Norway, where two examiners are always involved in the scoring of performance in the oral exam.
3.3 Assessment research in the Norwegian context

As mentioned in section 1.3.1, studies of assessment practices generally in Norway in the early 2000s pointed towards a “weak assessment culture” in Norwegian schools, including inadequate focus on learning objectives and unsystematic feedback routines. Later studies have confirmed these findings, documenting inappropriate evaluation practices that have been contrary to government regulations. For example, there is evidence that many teachers have been prone to award achievement marks in a norm-referenced manner, despite that fact that the system has been criterion-referenced since 2001 (Galloway, Kirkebøen, & Rønning, 2011; Hægeland, Kirkebøen, Raam, & Salvanes, 2005). There are also indications that teachers have tended to award different grades for the same performance and that they heed construct-irrelevant performance aspects, such as effort (Nusche, Earl, Maxwell, & Shewbridge, 2012). This may be explained in terms of difficulties in the interpretation of competence aims and the problem of relating student performance to different levels of competence (Prøitz and Borgen, 2010; Throdsen et al., 2009). There are, however, also indications that the initiatives undertaken by the educational authorities to improve assessment literacy among Norwegian teachers have had an impact. Studies by Hodgson, Rønning, Skogvold, and Tomlinson (2010) and Sandvik and Buland (2014), for example, indicate that teachers have become more focused on implementing principles for good formative and summative assessment practices, although there are considerable differences across schools. Sandvik and Buland (2014) also show that teachers’ assessment literacy is typically of a general character, relating to overarching pedagogical principles rather than to the subject-specific nature of assessment in the different subjects.

Research on teacher rater orientations and behaviour in oral assessment in the English subject is extremely scarce, however. One reason for this may be that the area of oral skills in education, both as regards L1 and L2, is underresearched (Svenkerud, Klette, & Hertzberg, 2012). Only one study, a Master’s thesis, has been identified which specifically addresses rater perceptions in the English classroom (Yildiz, 2011). Using a qualitative research design involving interviews with 16 English teachers at the upper secondary school level, Yildiz examined both the administration of the oral exam across different counties, as well as what kind of performance features the teachers attend to. Overall, the study found that the teachers were concerned with five general criteria: (i) “Language competence”; (ii) “Communicative competence”, (iii) “Subject competence”; (iv) “Ability to reflect and discuss independently”; and (v) “Ability to speak freely and independent of manuscript”. However, there was also clear evidence that the teachers paid attention to different performance aspects, that they
considered construct-irrelevant criteria, such as effort, and that they weighted criteria differently.

3.4 Short summary
Overall, the rater orientation and behaviour studies reviewed in this chapter point to varying degrees of rater variability in oral EFL/ESL assessment. There is evidence that judges attend to different kinds of performance aspects, that they apply criteria differently, that they have problems distinguishing levels of performance and that they vary in severity/leniency. However, the number of studies conducted is limited, particularly in curriculum-based assessment contexts where no common rating scale exists. Consequently, there is need for more research to provide evidence of how the construct is understood in this context and what kind of features (teacher) raters heed when scoring performance.
Chapter 4: Methodology

4.1 Introduction to the chapter
In the present chapter I describe the research design and explain how it contributed to answering the research questions. I also discuss aspects of research validity, including strengths and weaknesses of the design in view of the research focus. In addition, I consider some ethical aspects of the research process. The chapter supplements the sections on methodology in each of the three articles and justifies their relationship in the research project as a whole.

4.2 The phases of the research process
4.2.1 General overview
This investigation was primarily designed to investigate teacher raters’ understanding of the constructs to be tested in the GSP1/VSP2 oral English exam. In addition, it was devised to study aspects of scoring behaviour and correspondence between the teachers’ understanding of constructs and the intended constructs to be tested as specified by the English subject curriculum and accompanying government documents. The study has an inductive theoretical drive (Morse & Niehaus, 2009), utilizing a “qualitative priority” (Creswell & Plano Clark, 2011, p. 65). This implies that the overall direction of the research was guided by the inductive analysis of the data which was gathered through the use of qualitative methods in the first phase of the project. The design may further be described as emergent, allowing for the ongoing reconsideration of how to collect and analyse information based on what has been learnt in earlier stages of the project (Creswell, 2013). As the overarching focus of the thesis is on raters’ understanding of constructs, the prime object of study is the teachers’ cognitive processes. Hence, I found it relevant to use introspective research methods in the study of these processes (Gass & Mackey, 2000; Grotjahn, 1987; Sasaki, 2014). Such methods may either be qualitative or quantitative (Richards, 2009; Sasaki, 2014), and in the present study I collected introspective data through semi-structured interviews, questionnaires and verbal protocols.15

The research process comprised a pilot phase and two main phases. In the pilot phase a questionnaire was administered to a group of EFL teachers and subsequently followed by a

---

15 Verbal protocols are used in verbal protocol analysis (VPA) methodology (Alison Green, 1998).
semi-structured interview with one additional teacher. The pilot studies aimed to examine the teachers’ general orientations towards the constructs to be tested in order to obtain preliminary data on the research topic and to evaluate the efficacy of these research instruments with regard to the object of study. On the basis of the pilots I decided that semi-structured interviews would be better suited for investigating the problem statement in the first main phase. Hence, in the first main phase I interviewed a group of teachers on their understanding of *constructs generally*. The main bulk of this interview data was analysed in Study 1 and reported in Article 1 (cf. Figure 3, below). The analysis pointed towards two constructs which needed to be investigated further, namely *pronunciation* and *content*. Therefore, in the second main phase I went on to investigate these two constructs more closely. Firstly, in Study 2 I used unexplored qualitative data from the interviews in Study 1, as well as a questionnaire designed exclusively for Study 2, to examine the teachers’ orientations towards aspects of the pronunciation construct. The results of this investigation were reported in Article 2. Secondly, in Study 3 I used verbal protocols and semi-structured interviews to analyse the teachers’ understanding of the content construct. The results from this analysis were presented in Article 3. Figure 3 provides an overview of the research design.
Overarching research focus:

*Teachers’ understanding of constructs*

(+ rater behaviour and correspondence between teachers’ understanding of constructs and intended constructs)

**Pilot phase**

**Phase 1**

**Study 1** (Article 1)

Main research focus:

*Teachers’ understanding of constructs generally*

Data collection method:

Semi-structured interviews

**Study 2** (Article 2)

Research focus:

*Teachers’ orientations towards aspects of pronunciation*

Data collection method:

- Semi-structured interviews
- Questionnaire

**Study 3** (Article 3)

Main research focus:

*Teachers’ understanding of the content construct*

Data collection method:

- Verbal protocols
- Semi-structured interviews

**Pilot studies:**

1. Questionnaire survey
2. Interview

**Informed**

**Figure 3. Overview of the research design.**
In addition to the main focus on teachers’ understanding of constructs, the ancillary focus on aspects of teacher rater behaviour, i.e. grading, and correspondence between the teachers’ understanding of constructs and the intended constructs, as specified by the English subject curriculum and accompanying government documents, were investigated in Study 1 and Study 3. In the following I will describe the phases of the research process in further detail.

4.2.2 The pilot phase

The pilot questionnaire was conducted in collaboration with a master’s student at the University of Oslo in the very early stages of the project. Its main aim was to investigate teachers’ understanding of the constructs in the GSP1/VSP2 oral English exam, in order to test the suitability of this research instrument for the object of study (Yin, 2016). 32 upper secondary EFL teachers were asked to watch a video-taped performance of two students taking a mock oral English exam and then to respond to a questionnaire designed to tap into their understanding of what should be tested. The items in the questionnaire were developed on the basis of an analysis of the competence aims in the English subject curriculum and the mock oral exam tasks. Furthermore, we included an item to investigate attitudes towards the question of whether native speaker pronunciation is a relevant assessment criterion. The reason for doing this was a then recently published study which had shown conflicting attitudes among Norwegian EFL teachers regarding this question (Hansen, 2011). In addition to the specific items which were constructed to elicit rater orientations, we also created space for feedback in the questionnaire regarding the relevance of the items (cf. Appendix 5).

A main finding from this pilot survey concerned the usefulness of the research method. In written comments on the questionnaire four teachers pointed to the inappropriacy of several items, one remarking: “This has got nothing to do with the mastery of English or other foreign languages!” Consequently, we encountered hands-on a criticism sometimes voiced against questionnaires: The fact that the questions, or items, are developed by the researchers means that they to some extent measure the researchers’ understanding of the phenomena being studied (Jacobsen, 2005, p. 31). Although this position may be counter-critiqued, as questionnaires can be trialled and improved, it nevertheless prompted me to consider alternative methods of data collection in the first main phase of the investigation.

---

16 The mock exam consisted of a preplanned presentation (monologue) task and a discussion task based on a fictional text from the syllabus.
To continue the process, I decided to employ a qualitative research instrument, since qualitative methods are claimed to be particularly suitable for studying “the views and perspectives of a study’s participants” (Yin, 2016, p. 9). Hence, I created an interview guide based on the pilot questionnaire responses in order to test whether semi-structured interviews would better capture the teachers’ orientations towards the constructs (cf. Appendix 6). The interview guide was piloted on a GSP1/VSP2 English teacher, and on the basis of this pilot, I concluded that semi-structured interviewing would be a more appropriate data collection method for Study 1. In addition, I decided that a video-taped student performance used as prompt, or stimulus, would be useful for eliciting responses. However, rather than employing the video-clip from the mock exam presented to the survey pilot teachers, I decided to use a video-recording from an authentic GSP1/VSP2 oral English exam.

The video-prompt was obtained by recruiting a VSP2 student who agreed to be filmed as she was taking her GSP1/VSP2 oral English exam. The exam format comprised three tasks: (i) a pre-planned presentation task, followed by a discussion between the examiners and the student; (ii) an interview task based on a short story from the syllabus; (iii) an interview task based on an audio-taped listening comprehension sequence. In the first task the student had been given 48 hours in advance to plan a presentation on “a common health issue in today’s society [and] the problems it causes the individual and in society”. As for the second and third tasks, the short story was about eating disorders, and the listening comprehension exercise treated the issue of English as a world language.

4.2.3 Study 1 (Article 1)

In Study 1 I used a revised version of the pilot interview guide as a basis for interviewing a sample of 24 GSP1/VSP2 English teachers on their understanding of the constructs. First, I asked the teachers to score the video-taped student performance. Next, I interviewed them on what kind of performance features they had been paying attention to in the specific performance of the student in the video-clip, as well as which performance aspects they would heed generally in this exam. Beyond this, I compared the teachers’ understanding of constructs with aspects of the intended construct as stipulated by the subject curriculum and a government circular. These analyses served as evidence for the evaluation of the validity of the scoring inference. The results of the analyses were presented in Article 1 and formed the basis for identifying the phenomena to be studied more closely in Study 2 and Study 3.
Overall, the responses indicated that phonology was an important feature heeded by the teachers. Moreover, as the pilot survey had indicated that the teachers had widely differing views on the question of native speaker pronunciation, I decided to investigate the pronunciation construct in more detail in Study 2. In fact, a question on native speaker phonology had been included in the interview guide used in Study 1, but the responses to this question had been left unexplored. The reasons for this were partly space constraints, and partly because the strong focus on pronunciation by the teachers made me realize that the native speaker question could be further investigated in the second main phase of the project. In addition, as the teachers interviewed in Study 1 had diverging opinions on the question of how to evaluate content, I found it relevant to look more closely into this construct in Study 3.

4.2.4 Study 2 (Article 2)

Study 2 analysed the unexplored question of native speaker pronunciation which had been put to the 24 teacher informants in Study 1. In addition, on the basis of findings from pronunciation research, I developed a questionnaire which I distributed to another cohort of 46 teachers. The questionnaire items attempted to elicit both the teachers’ views on the native speaker question, as well as their orientations towards aspects of the pronunciation construct which has been found by research to be important for communication. The video prompt, which had been used in Study 1, was also used as stimulus in Study 2. The analyses of this data were presented in Article 2.

4.2.5 Study 3 (Article 3)

Finally, in Study 3 I collected concurrent verbal reports (Alison Green, 1998) and interview data from an additional group of 10 teachers in order to investigate their understanding of the content construct. The study used theoretical and empirical evidence from Study 1, as well as conceptualizations from the research literature, to develop a conceptual framework for analysing teacher statements pertaining to content. In addition, I compared the teachers’ understanding of the content construct with aspects of content identified in the English subject curriculum. This comparison provided evidence for the evaluation of the validity of the scoring inference. The verbal reports were gathered by letting the teachers watch the same video-prompt used in Study 1 and Study 2 and then simultaneously comment on which

17 In this thesis I use the terms phonology and pronunciation interchangeably to refer to both segmental (i.e. individual sounds) and suprasegmental (e.g. stress, intonation) aspects of pronunciation (cf. Talia Isaacs, 2014).
performance aspects they paid attention to. Subsequently, semi-structured interviews were carried out with the same group of teachers, who were asked to elaborate on their understanding of the content construct. The findings were reported in Article 3.

4.3 Data collection

4.3.1 Study 1: Teachers’ overall understanding of constructs

As mentioned above, the experiences from the pilot studies prompted the decision to use semi-structured interviews in Study 1. However, a number of issues had to be considered, as qualitative interviews may take a number of forms, depending on the researcher’s ontological and epistemological positions, the nature of study, the research questions, time constraints etc. (Brinkmann & Kvale, 2015; King & Horrocks, 2010; Mann, 2016). I will here briefly present some epistemological reflections on interviewing which are important for understanding my approach to collecting and analysing the interview data in this thesis.

According to Silverman (1993, 2011) and Alvesson (2003), interview research can broadly be said to be informed by three major paradigms, namely neopositivism, romanticism and localism. Neopositivism typically draws on quantitative ideals for data collection, seeing interviews as a channel for transmitting, more or less, objective knowledge. In order to obtain valid data, contextual factors must be controlled or minimized as they may ‘contaminate’ the information that is gathered. Objectivity and neutrality are important principles, and the researcher should therefore strive to keep a professional distance and follow rigorous procedures in order not to cause unnecessary ‘bias’ (Alvesson, 2003). Romanticism, on the other hand, is more inclined to regard the research interview as a meaning-making event, rather than a “pipeline for transmitting knowledge” (Gubrium & Holstein, 2003, p. 68). The focus is on creating a situation in which the interviewer can get access to the perceptions or experiences of the informant by establishing “rapport, trust and commitment”, rather than following a strict research protocol (Alvesson, 2003, p. 16). However, romanticism shares with neopositivism the idea that the perceptions and experiences of the research participant can be regarded as an “object […] located inside people’s heads” (Silverman, 2011, p. 18). Thus, the interview may be seen as a ‘tool’ or a ‘technique’ for collecting evidence about how individuals understand the world (Alvesson, 2003). Localism, on the other hand, is sceptical of this tool metaphor and stresses the importance of the social context in the creation of interview outcomes. As Alvesson (2003) puts it, the interview is seen as a localized accomplishment, in which the interviewees are producing “situated accounts, drawing upon
cultural resources” (p. 17). Also, the localist position takes an essentially critical stance on interviewing. It confronts the assumptions, purposes and arguments of those who aspire to use interviews instrumentally (Silverman, 1993).18

In section 4.2.1 I described the data collection methods used in this thesis as introspective. Although introspective research methods may be employed using different philosophical perspectives, there seems to be a tendency for them to be used by researchers adhering to the neopositivist paradigm. For example, according to Gass & Mackay (2000), introspective methods are “a means of eliciting data about thought processes involved in carrying out a task or activity, [working under the assumption that] humans have access to their internal thought processes at some level and can verbalize those thought processes” (p. 1). Such statements point to a neopositivist epistemology, evoking the pipeline metaphor of Gubrium and Holstein, quoted above. However, I believe a critical perspective on such a position is important, as a host of factors – other than interviewer bias – may influence the interview ‘output’. Examples of these are the expectations of the interviewee, the informant’s desire to express a certain identity (or identities), the use of professional jargon to describe the phenomenon being studied and the informant’s inclination to further his or her own political viewpoints (Alvesson, 2003). Alvesson, therefore, advocates a reflexive pragmatist view on interviews, in which it is important that the researcher recognizes the variety of meanings that may occur, interpreting these in an open and (self-) critical way. Based on my pragmatist epistemological position (cf. section 2.1), I found this approach to be highly relevant for the collection and analysis of data in this thesis. I do not completely reject the interview-as-tool metaphor (neither does Alvesson), but I believe it is important to critically scrutinize the data produced in the interviews, always keeping in mind alternative interpretations. This fits with the pragmatist notion that the outcome of research is not necessarily “true” knowledge, but knowledge that is “useful” (Brinkmann & Kvale, 2015, p. 65).

On a more practical note, I also considered collecting data in Study 1 by using verbal protocols. This is a form of introspective data collection method which has been popular in rater cognition studies (Borg, 2003; Cai, 2015; H. J. Kim, 2015; May, 2006; Orr, 2002). It assumes that it is possible to make inferences about individuals’ cognitive processes by asking them to verbalize their thoughts during or after the completion of a task (Ericsson & Simon, 1993; Alison Green, 1998). Verbalizations which are made during task completion are

---

18 Brinkmann and Kvale (2015) present an alternative description of interview research paradigms, illustrated in terms of an interviewer-as-miner metaphor and an interviewer-as-traveller metaphor. The former may be labelled (neo)positivist, the latter constructivist.
referred to as concurrent, whereas verbalizations that are produced afterwards are termed retrospective. The verbal data will constitute a ‘protocol’ or a ‘report’, which must be gathered in the form of an audio or video-recording in order to be subjected to scientific analysis. Verbal protocols can either be used with or without stimuli, or “memory aids” (Sasaki, 2014, p. 1344). Protocols using stimuli are also referred to as stimulated recall.

One of the advantages of verbal protocols compared to interviews is the lower level of personal reactivity involved (Hammersley, 2008a), which means that the research participants are less likely to be affected by the researcher’s involvement, for example as interviewer. However, there are also disadvantages, especially regarding concurrent reports, relating to issues such as taciturn participants, ambiguous and ‘superficial’ comments and the missed opportunity on the part of the researcher to ask for clarification and to probe further into the participants’ understanding of the phenomena.

Despite the reputed advantages of verbal protocols in rater cognition studies, I decided to employ qualitative interviews in Study 1. Even though I recognized the challenges of reactivity, I wanted to have the opportunity to ask for clarification and to delve deeper into the teachers’ understanding of constructs, which is the main focus of this project. This included their understanding of the overall meaning of the scores, their perceptions of which criteria they regarded the most important, and the question of native speaker pronunciation (cf. interview guide, Article 1, Appendix B). On the basis of these aims, I chose a semi-structured interview format, which implied an overall inductive approach, where open-ended questions at the beginning of the interview were meant to make the teachers describe in their own words their understanding of what to be tested. However, deductive elements are apparent in the specific questions relating to, for example, native speaker pronunciation.

It may of course be objected that I could have combined interviewing and verbal protocols in this study, as I did in Study 3, in order to obtain richer data. This possibility was indeed considered. However, the recruitment of students for the video prompt proved very time-consuming, since the majority of the local educational authorities, school principals and teachers that I contacted turned down my request for authorization to film. Therefore, I decided to opt for interviewing only. The research questions which were addressed in Study 1 are listed in Table 3, below.

---

19 This presupposes a ‘non-mediated’ VPA procedure, where the researcher refrains from intervening during verbalization. Alison Green (1998), however, also describes ‘mediated’ alternatives, where the researcher actively prompts participants to make comments when they fall silent, or even asks questions for clarification.
4.3.2 Study 2: Teachers’ orientations towards pronunciation

As mentioned in section 4.2.4, parts of the data which were analysed in Study 2 were actually gathered in Study 1. This data comprised un-analysed responses to the question on native speaker pronunciation, which had been put to the 24 interview informants. The question read: “What about phonology? Some teachers say that a near-native speaker accent is important in order to get a top score? What is your comment on that?” (cf. Article 1, Appendix B).

The actual data collection in Study 2 took as a starting point this question about native-speaker accent, as well as the finding from Study 1 that the teachers were generally very concerned with pronunciation. Based on this evidence, I consulted the pronunciation research literature in order to develop an analytical framework which could be used to investigate teacher orientations towards this construct. The framework centred on the concepts of \textit{nativeness} and \textit{intelligibility} (Levis, 2005) and a related set of four specific pronunciation features found to be important for communication (cf. Article 1). These concepts and features were operationalized in a number of items in a questionnaire, which was then distributed to 46 GSP1/VSP2 English teachers. The purpose of the questionnaire was twofold. Firstly, it aimed to corroborate the findings on native-speaker pronunciation investigated by the interview question. Secondly, it tried to test the extent to which the teachers attended to the four specific pronunciation aspects when assessing student performance. Hence, this specific design may be said to have both validity checking and complementary information seeking purposes (Hammersley, 2008b).

Broadly speaking, the data collection in the second study formed part of the larger, emergent and inductive process. However, the gathering of data in Study 2 itself may be characterized as essentially deductive. The questionnaire, for example, served a hypothesis-testing purpose, checking to what extent the teachers judge pronunciation against a native speaker norm. The research questions that were asked are listed in Table 3, below.

4.3.3 Study 3: Teachers’ understanding of content

In Study 3 data was gathered both inductively and deductively from 10 GSP1/VSP2 teacher participants. Firstly, concurrent verbal protocols were collected \textit{inductively}, using the video-clip as stimulus, in order to produce ‘grounded’ evidence on the teachers’ understanding of the content construct. The verbal protocols were followed up in interview sessions immediately afterwards. In the first part of the interviews I asked the teachers open-ended questions on how they would assess the performance they had just seen. In these two phases
then – the recording of the verbal protocols and the first part of the interviews – no hypotheses, theories or conceptualizations were guiding the information gathering.

However, prior to the data collection phase, I had developed a conceptual framework for describing content, which was intended as an analytical tool in the exploration of the teachers’ statements. This framework, which was built on findings from Study 1, indicated that the teachers were largely assessing content in terms of a Bloom-like taxonomy of analysing and reflecting on subject matter. The framework was further supplemented with theoretical descriptions of content from educational theory and content-based instruction literature (e.g. Anderson & Kratwohl, 2001; Chamot, 2009). At its core, it conceptualized content as a two-dimensional construct, consisting of a subject matter, or what, dimension, and a skills and processes, or how, dimension. The framework was operationalized in a set of interview questions which were put to the teachers in the second phase of the interview (cf. Article 3). The questions tried to tap into the teachers’ notions of subject matter content. Additionally, in order to check the correspondence between the teachers’ understanding of the content construct and the intended construct, I asked the informants about the relevance of a number of content-related issues which were identified in the subject curriculum. Thus, in the second phase of the interviews, the data collection may be regarded as deductive.

The reason for combining inductive and deductive information gathering in this way, was to check whether the verbal protocol data and the teachers’ answers to the open-ended questions could corroborate the responses to the theory-driven questions in the second phase of the interviews (cf. section 4.5, below). In addition, the verbal reports and open-ended interview questions also served a complementary information seeking objective (Hammersley, 2008b), in that they might provide additional evidence which the two-dimensional content model had not managed to capture.

A final word to be made here concerns the validity of verbal protocols. According to Alison Green (1998), it is important that the research participants are given proper instructions, in order to be able to produce rich and dependable reports. It is also essential that they are encouraged to speak as much as possible. I therefore attempted to explain as clearly as possible the purpose of the research and the usefulness of this data collection method, when properly carried out. I also stressed the importance of providing as many comments as possible, encouraging the teachers to verbalize any thought that came to mind when watching the video-clip. In addition, I allowed them five minutes at the beginning of the session to familiarize themselves with the equipment and to trial a commenting sequence. The research question for Study 3 is listed in Table 3, below.
4.4 Participants and procedure

The student participant who was video-taped as she was taking her oral exam was an 18-year-old girl in the Health and Social Care vocational study programme. She consented to participating after having been invited by her English teacher to take part in the study. The recording produced a 22-minute long video-sequence. All of the teacher participants in the three studies (n=80) were fully qualified EFL teachers at the upper secondary level in Norway. They were recruited by means of purposeful sampling (Creswell, 2013), in order to obtain variation in the samples with regard to age, gender, L1, teaching experience, county and study programme affiliation. No economic incentives were provided.

The 24 interview informants in *Study 1* were recruited from 19 different schools in each of the three different counties of Finnmark (n=8), Oslo (n=8) and Østfold (n=8). One of the teachers in Oslo and seven in Østfold were interviewed face-to-face, whereas the rest were interviewed by telephone. They were recruited directly by email or telephone, after I had approached schools and asked them to help me identify informants on the basis of the above mentioned criteria. Three interviews were conducted in English and 21 were carried out in Norwegian. All the informants received a letter in advance informing them of the purpose of the study. The video-clip was distributed to them on a USB memory stick a few days before the interviews were scheduled. Instructions were given to watch the video-clip, score the performance and to justify the decision. Moreover, I asked the teachers to watch the video as near the time of the interview as possible, and to take notes, in order to keep the performance as vividly in their memory as possible. The interviews lasted between 24 and 47 minutes.

The interview informants in *Study 2* were the same as those in Study 1. Of the 46 questionnaire respondents, 14 represented 10 different schools in eight counties (Akershus, Aust-Agder, Nord-Trøndelag, Oppland, Sogn og Fjordane, Sør-Trøndelag, Vestfold and Østfold). They were recruited in the same way as the informants in Study 1. The remaining 32 respondents represented a number of schools from all over the county of Akershus. They were recruited in a teacher seminar via a teacher who invited me to his school to talk about language assessment. At the beginning of the seminar they were informed of the purpose of the study and requested to participate, which all of them agreed to. After my talk, they were shown the video-clip and asked to complete the questionnaire. Being present, I was able to clarify three questions which were asked in relation to the wording of some of the items.

The 10 interview and VPA participants in *Study 3* represented five different schools in Akershus, Oslo and Østfold. They were recruited in the same way as the informants in Study 1. The purpose of the study and the specifics of the research design were explained to the
participants on email in advance. To collect the data I met with each of the teachers individually. The video-clip was shown to them on a laptop-computer, and headphones were provided to let them listen to the student’s speech while producing the verbal protocols. They were also given a digital voice recorder to record their comments. The VPA procedure was then explained, and the teachers were given 5-10 minutes to familiarize themselves with the procedure. I pointed out that they were allowed to stop the video if they needed to, but that they could preferably watch it all in one go without interruption. They were also told to provide as many comments as possible. During the recording of the protocols I left the room in order to minimize reactivity. Immediately afterwards, I interviewed each of them on their understanding of the content construct. Nine protocols were recorded in Norwegian. One, carried out by a native speaker of English, was conducted in English.

4.5 Data analyses
All the data from the interviews and the verbal protocols were explored using the computer programme QSR NVivo10. The questionnaire data was analysed using IBM SPSS Statistics. All the interviews and verbal protocols were transcribed, checked and sent to the teachers for respondent validation (Bryman, 2012). The data obtained in Study 1 was analysed by means of qualitative and quantitative content analysis (Galaczi, 2014; H.-F. Hsieh & Shannon, 2005; Krippendorf, 2013). This analysis may be characterized as essentially inductive, in the sense that I decided to let the analytical categories, as far as possible, develop from the teachers’ statements. However, as I have good knowledge of the English subject curriculum, theories of communicative competence, the Common European Framework of Reference etc., it is of course impossible to claim that the categories I developed were purely data-driven. For example, one category I invented was labelled “compensatory strategies” (cf. Article 1). This category was coded from statements such as: “And if they can’t [find the word they are looking for], they should try to circumvent it, rather than switching into Norwegian”. While analysing such statements, the first concept that came mind was admittedly Canale and Swain’s (1980) term strategic competence, which is conceptually what circumvention rather than switching into your first language is about. Hence, my analysis clearly has elements of deduction, or rather abduction, in it (see e.g. Douven, 2011; Hanson, 1958).

As pointed out in section 4.3.2 the analytical framework which was developed for Study 2 centered on the concepts of nativeness and intelligibility and four related pronunciation features identified as important for intelligibility. Both the interview data and
the questionnaire responses were investigated using this framework. The interview data, which consisted of answers to the question of native speaker pronunciation (cf. section 4.3.2), was analysed both inductively and deductively, using magnitude and provisional coding (Miles, Huberman, & Saldaña, 2014). The magnitude coding was used to answer the question of nativeness, i.e. the extent to which (near-) native speaker pronunciation is a relevant assessment criterion. It involved the assignment of coded statements along a four-point continuum going from “not at all” in agreement with the nativeness perspective to “to a large extent” in agreement. The provisional coding was employed to answer the question of intelligibility, i.e. the degree to which students merely need to make themselves understood. It was carried out by deductively setting up a number of phrases pertaining to intelligibility, such as “comprehensible speech” and “understanding” and then searching the interview transcripts for corresponding phrases. Finally, the questionnaire responses were used to answer the question about the relevance of the four pronunciation features. They were analysed deductively by calculating descriptive statistics, such as means and standard deviations, in order to examine the teachers’ orientations towards these constructs.

Finally, in Study 3 I analysed the verbal protocols and the interview data obtained in the second phase of the project. The analysis was carried out in two cycles using provisional coding (Miles et al., 2014). In the first cycle, teacher statements were coded on the basis of the construct categories developed in Study 1, such as “Vocabulary”, “Grammar” and “Fluency”, as well as the analytical framework developed specifically for Study 3. The analytical framework, based on the before mentioned subject matter and skills and processes dimensions (cf. section 4.3.3), helped identify the teacher statements relating to content. More specifically, these dimensions were identified in terms of noun phrases (subject matter) and verb phrases (skills and processes) (cf. Kratwohl, 2002). For example, the teacher statement She didn’t get the chance to, sort of, talk about the English language as a world language and an international language was analysed in terms of these two phrase types. The verb phrase “talk about” represents the skills and processes dimension, and “English language as a world language and an international language” represents the subject matter dimension.

In the second analysis cycle, I sifted out all the phrases relating to subject matter, which was the main object of study, and re-analysed these statements with a particular focus on noun phrases, as specified by Kratwohl. Finally, I compared these phrases with the corresponding noun phrases representing subject matter in the English subject curriculum.

---

20 These were segmentals (i.e. individual sounds), word stress, sentence stress and intonation (cf. Article 2).
When considered across the research phases, my analysis of the content construct may be said reflect abductive reasoning. For example, in Study 1, I came across statements which indicated that the teachers were thinking in terms of the Bloom-like categories of description, analysis, reflection etc. when assessing content. When I put this to one informant, she confirmed that this was how she thought most teachers reasoned about content. Hence, I inferred that this was a relevant conceptualization of the construct, and I therefore consulted the theory, i.e. Bloom’s revised taxonomy, to further elaborate on this perspective. In Study 3 I returned to the teachers to collect more empirical evidence, which in turn confirmed the theory. Hence, there was a going back and forth between empirical data and theory in a way very similar to the abductive process described by, for example, Hanson (1958, p. 85).

As mentioned above, the interview transcripts were sent to the informants for respondent validation. In order to further validate the analyses, three colleagues were involved in parts of the coding process. In Study 1 two colleagues, who had previously worked as EFL teachers, agreed to code four transcripts (16 per cent of the total). The comparison between their coding and mine resulted in a moderate inter-coder reliability estimate (Cohen’s Kappa = .69). Therefore, we sat down to discuss the coding categories, in order to arrive at a better understanding of how the teachers’ statements could be analysed. After having revised the coding scheme and re-analysed the transcripts, I asked one of the colleagues to code two new transcripts. This inter-coder reliability analysis resulted in a Cohen’s Kappa estimate of .89, which can be regarded as very good (Landis & Koch, 1977). In Study 2, one of these colleagues also agreed to code two transcripts (8 per cent of the total). This coding resulted in a Kappa estimate of .85, which is also very good. In Study 3 a third colleague, who also has previous experience as a teacher in upper secondary school, consented to analyse two transcripts (20 per cent of the total). The inter-coder consistency between my own coding and hers resulted in a Kappa estimate of .83 for the VPA and .78 for the interviews, which may be regarded as substantial. Obviously, it would have been preferable to have more researchers co-code a larger number of transcripts, but time-constraints did not allow for that. Table 3 gives an overview of the central elements in the research process, such as research foci and research questions, data collection, data analyses, main constructs and methods of validation.
Table 3. Overview of the central elements in the research process.

<table>
<thead>
<tr>
<th>Study 1 (Article 1)</th>
<th>Study 2 (Article 2)</th>
<th>Study 3 (Article 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Research focus</strong></td>
<td>Teachers’ understanding of constructs generally</td>
<td>Teachers’ orientations towards aspects of the pronunciation construct</td>
</tr>
<tr>
<td><strong>Research questions</strong></td>
<td>1. How do EFL teachers in Norway understand the constructs and criteria to be tested in an oral English exam at the upper secondary level? 2. What kind of criteria do these teachers see as salient when assessing performance?</td>
<td>1. To what extent do EFL teachers at the upper secondary level in Norway see nativeness as an important criterion in the assessment of pronunciation? 2. To what extent do the teachers see intelligibility as an important criterion in the assessment of pronunciation? 3. To what extent do teachers see segmentals, word stress, sentence stress and intonation as important in the assessment of pronunciation?</td>
</tr>
<tr>
<td><strong>Data collection</strong></td>
<td>Semi-structured interviews</td>
<td>- Semi-structured interviews - Questionnaires</td>
</tr>
<tr>
<td><strong>Number of participants</strong></td>
<td>24 interviewees</td>
<td>- 24 interviewees - 46 questionnaire respondents</td>
</tr>
<tr>
<td><strong>Data analysis</strong></td>
<td>Qualitative and quantitative content analysis (Galaczi, 2014)</td>
<td>Qualitative, using magnitude and provisional coding (Miles et al., 2014), and quantitative, calculating descriptive statistics</td>
</tr>
<tr>
<td><strong>Main constructs</strong></td>
<td>- Communication - Content</td>
<td>- Nativeness - Intelligibility</td>
</tr>
</tbody>
</table>
4.6 Research validity

In section 1.1 test validity was described as the extent to which inferences drawn from assessment scores are ‘sound’. In a similar vein, research validity may be described as the trustworthiness of the interpretations that are made from research results (Hesse-Biber & Leavy, 2011; Kleven, 2008; Yin, 2016). A number of qualitative researchers avoid using the term ‘validity’, associating it with the measurement paradigm and quantitative research methods. This is worth noting, as the current investigation utilizes a qualitative priority (cf. section 4.2.1). However, as Kleven (2008) points out, regardless of scientific outlook, the essence of a research endeavour is the dependability of the knowledge claims that are ultimately being put forward. In this perspective, it is of minor importance whether one chooses to use the labels validity, trustworthiness, credibility or dependability to describe the quality of the research outcomes. Rather, what is important is to clarify issues that may undermine the soundness of the conclusions drawn. No universal, standardized list of criteria for the clarification of these issues exist, but in the following I will use Patton’s (2015) 10 general criteria for assessing the quality of qualitative research in order to evaluate the validity of the current study. The 10 criteria are:

1. Clarity of purpose
2. Epistemological clarity
3. Questions and hypotheses flow from and are consistent with purpose and epistemology
4. Methods, design and data collection procedures are appropriate for the nature of the inquiry
5. Data collection procedures are systematic and carefully documented
6. Data analysis is appropriate for the kind of data collected
7. Strengths and weaknesses are acknowledged and discussed
8. Findings should flow from the data and analysis
9. Research should be presented for review
10. Ethical reflection and disclosure

Point no. 8 on this list will be considered in Chapter 5, and point no. 10 will be treated in section 4.8. In addition to these ten aspects of research quality, I will consider the question of generalizability of the results, as this point is commonly regarded as vital in social science research. I turn to this question in section 4.9.
As for the first point on Patton’s (2015) list, the purpose of the present investigation was pointed out in section 1.4. By studying Norwegian EFL teacher raters’ understanding of constructs, rating behaviour, and correspondence between aspects of teachers’ understanding of construct with what the curriculum and other defining documents identify as construct relevant and irrelevant, I provide empirical evidence of these issues in a domain which is under-researched. Not only is there little evidence of oral EFL assessment in curriculum-based educational settings generally; such research has hardly been undertaken in Norway and has been explicitly requested by the research community (cf. section 1.3.1).

In terms of epistemological clarity, I established in section 2.1 that I adhere to a pragmatist position in the sense that I see the terminology employed in the thesis as tools for understanding the phenomena under investigation. The same is true of the methods used. In this thesis I have utilized different introspective methods because I have found them useful for studying the teachers’ understanding of what should be tested. Hence, this “what works” approach draws attention to the research problem rather than to the methods used (Creswell, 2013, p. 28). Obviously, this implies some assumptions about what kind of inferences that can be drawn from the research results. In section 4.3.1 above, I pointed out that introspective research assumes that it is possible to get access to individuals’ cognition through the use of these particular methods. I recognize of course that individuals will not always be able to verbalize their thoughts (Freeman, 1996; Polanyi, 1966) and that this is a potential weakness in such research designs. However, all designs involve assumptions about some aspect of the world, and what is important is to make these assumptions explicit. Beyond this, a core question in the pragmatic paradigm, according to Patton (2015), is what practical consequences can be drawn from a given study (p. 152). In the current study this question relates first of all to the problem of not having a common rating scale on the national level. I will return to this problem in Chapter 5.

The third and the fourth points on Patton’s list are interrelated and have been demonstrated in this chapter. I have argued in section 4.3.1 that the use of introspective research instruments is a highly rational choice in a study of teachers’ cognitive processes. Moreover, the use of multiple methods, as employed here, is quite consistent with a pragmatist epistemology (Creswell, 2009, p. 9). In addition, I have argued that the chosen methods – interviews, questionnaires and verbal protocols – in order to answer the six research questions outlined in Table 3 have been appropriate for providing sound evidence consistent with the purpose of the study, i.e. to investigate the rating process in the GSP1/VSP2 oral English exam. That being said, I recognize that I would have added
descriptive power to the results by, for example, combining verbal protocols and interviews in the Study 1.

In terms of points five and six on the list, I have in the present chapter and in the appendices attached also documented the systematicity of the data collection procedure and shown how the different phases of the research process followed logically from each other. Similarly, I have shown how the data analyses have been undertaken and discussed the appropriateness of these analyses with regard to the data collection and the nature of the study (cf. sections 4.3 and 4.5, above).

When it comes to point seven on Patton’s list, I would argue that one of the strengths of this investigation is its use of methods which are highly relevant for the nature of the study. Also, the use of multiple methods, particularly in Study 3, both strengthened the validity of the results and broadened the perspective on the nature of that particular inquiry, i.e. the content construct. Moreover, the involvement of as many as 24 interview informants in the first study produced rich descriptions of the teachers’ understanding of the constructs.21 This claim is supported by the results in Study 3, which showed that the informants there had a similar understanding of the content construct as the participants in Study 1.

However, there are also weaknesses in the investigation which need to be addressed. The problem of reactivity has already been mentioned, cf. sections 4.3.1 and 4.4. Similarly, I noted in section 4.5 that a larger number of coders to validate the interview and verbal protocol analyses would have strengthened the descriptive power of the conclusions arrived at. This is also a relevant point in the sense that the second inter-coder reliability measure, which was obtained in Study 1 (Cohen’s Kappa .89), may have been influenced by the fact that I discussed the coding categories with the colleague who was carrying out the second co-coding. Hence, if I had involved other coders, who had not been involved in the discussion of the categories, the results may have been different. An additional perspective relating to the previous point is the general problem of representing an emic, or insider, perspective of research participants through analyses made by a researcher, as it is well-postulated that observation will always be theory-laden (Hanson, 1958). This, by the way, was touched upon in my discussion of the abductive nature of some of my analyses (cf. section 4.5).

In addition, the use of only one video-clip as stimulus in the three studies may have affected the teachers’ focus unduly and influenced their responses in the verbal protocols, interviews and questionnaire. As I argue in all of the articles, had there been another student,

21 Compare the interview study by Guest, Bunce & Johnson (2006, quoted in Bryman, 2012, p. 426), where the authors found that around 12 interviews (out of 60) were sufficient for achieving data saturation.
their attention might have shifted to different performance aspects. Against this claim, however, it could be said that the teachers are generally well accustomed to thinking in terms of, and verbalizing, criteria, particularly for formative purposes in the classroom. Hence, the effect of only one video-clip is uncertain. However, what they are not accustomed to is producing concurrent verbal protocols. It could therefore be held that the use of such protocols is an unnatural way for teachers to talk about criteria and that it may have impacted on the data in unforeseen ways.

With regard to points eight and nine on Patton’s list of criteria, I will return to a summary and discussion of the findings in Chapter 5. As for the presentation of the findings for review, Article 1 has been published in Sage Open, Article 2 has been submitted to Language Assessment Quarterly, whereas Article 3 has been submitted to Assessment in Education: Principles, Policy and Practice (cf. section 5.2)

4.7 Ethical considerations
Criterion 10 on Patton’s (2015) list concerns ethical considerations and needs to be addressed here in some more detail. Drawing on Diener and Crandall (1978), Bryman (2012, p. 135) lists four main ethical issues which need to be considered in social science research. These relate to:

(i) whether there is harm to participants;
(ii) whether there is lack of informed consent;
(iii) whether there is an invasion of privacy;
(iv) whether deception is concerned.

All of these questions seem to depend on how the specific notions of harm, consent, invasion and deception are interpreted. However, in terms of the first point, Bryman (2012) argues that harm can be related to the confidentiality of records. In the present investigation this aspect is relevant in relation to the video-recording of the student taking the oral exam, the audio-recordings of interview informants and VPA participants, as well as the interview and VPA transcripts. According to the Norwegian National Research Ethics Committee (2006), records must be protected whenever participants can be directly or indirectly identified through, for example, the combined variables of place of work, age, gender, occupation etc. In my project, the most obviously identifiable participant was the student recorded in the video-clip.
Although neither her name nor her school is recognizable in the video-sequence, it is evident that she is easy to identify. Hence, the video-clip was stored on my computer, which is only accessible through a username and a password, as well as on USB memory sticks, which were kept in a locked cabinet. The teachers who received them were explicitly instructed to either send them back to me immediately after assessment, or to delete their contents. Beyond the student who was video-recorded, all the teachers were represented only by a code in the audio-recordings, transcripts and articles. Personal details, including names and email-addresses, which could be linked to the codes, were kept in a notebook in a locked cabinet.

The second principle Bryman lists concerns lack of informed consent. In the current investigation I sent written letters (consent forms) to all prospective participants informing them of the purpose of the study, the research questions, the use of data, and an assurance that all the participants would be guaranteed anonymity (cf. Appendix 7). In addition, the consent forms going out to the students informed them that they could withdraw from the project at any time (cf. Appendix 7). According to the Norwegian National Research Ethics Committee (2006) informed consent also means that the information that is given to participants should be given in a “neutral manner” and that prospective participants should be informed that participation is voluntary (p. 12, my translation). The reason for this is not to put pressure on anyone to sign up for something that they may later regret. In my letters to the teachers and students I took care to use a neutral language, and one of the executive officers at the Local Educational Authority (LEA) in the county of Østfold read through the consent forms and made suggestions for improvements.

The third principle, invasion of privacy, can be linked to issues of anonymity and confidentiality and is particularly relevant in relation to the student who was being filmed. The question of anonymity relating to the confidentiality of the recording has already been mentioned. As for the dissemination of the video-clip to the teachers, one might consider it a potential problem that I was not able to guarantee that the teachers would not show the video-sequence to third parties. However, as teachers are bound by professional secrecy and are trained to handle such issues, I regard this matter as relatively unproblematic.

Similarly, the principle relating to deception is not particularly relevant in this thesis. Apart from the fact that I did not disclose my research question in Study 3 to the teachers until in the second half of the interview, which I do not consider as essentially unethical, I have at no point tried to keep any information relevant to the participants away from them. On the whole, I consider the research topic to be fairly uncontroversial, and it did not involve minors or “other vulnerable populations” (Miles et al., 2014, p. 60). All required authorizations were
obtained from the Norwegian Data Protection Official for Research (cf. Appendix 8), the
LEAs, and the different school principals, and data was stored and treated in accordance with
the guidelines of the National Research Ethics Committee.

4.8 Generalizability
The final point I would like to make in this chapter concerns the generalizability of the
results. Generalizability concerns the representativeness of research findings and is a standard
aim in quantitative research, where it is commonly regarded as an integrated aspect of
research validity (Bryman, 2012). In quantitative research generalizability typically involves
the use of statistical sampling techniques in order to generalize findings from a randomized,
representative sample to the population from which the sample was drawn (Onwuegbuzie &
Leech, 2010). As Bryman (2012) points out, it is impossible to generalize statistically from a
small, non-randomized sample to a population. The findings from the three purposefully
collected teacher samples gathered in the present investigation can therefore not be
statistically generalized to the population of EFL upper secondary school teachers in Norway.

However, other forms of generalization are possible. As Gobo (2008) has argued, in a
number of research fields, such as geology, genetics, history, linguistics and paleontology,
researchers work with non-probability samples which are “regarded as being just as
representative of their relative populations and therefore as producing generalizable results”
(p. 200). Two types of generalizations that are relevant for the present investigation are
analytical, or theoretical, generalization (Hammersley, 1992; Mitchell, 1983; Yin, 2016), and
generalizable patterns (Larsson, 2009; Patton, 2015, p. 107). The former implies that it is the
“cogency of the theoretical reasoning” rather than statistical criteria that are decisive for
judging the extrapolation of the results (Mitchell, 1983, p. 207). The latter can be understood
as “configurations, which can be recognized in the empirical world” (Larsson, 2009, p. 33).

In this study the notion of theoretical generalization is particularly relevant in relation
to the content construct. In Article 1 I found evidence that the teachers were largely thinking
of the content construct in terms the ability to explain, analyse and evaluate whatever is up for
discussion. From this empirical finding I developed a theoretical model of content based on
Bloom’s revised taxonomy, which involved a subject matter dimension and a skills and
processes dimension (cf. section 4.3.3). This theoretical model was further empirically
supported by the data gathered in Study 3. Hence, the conclusion that the content construct
can be understood in this way, is not a matter of statistical generalization, but rather
“theoretical representativeness” (Gomm, 2008, p. 235). In other words, it suggests that such a model is a relevant way of describing theoretically how teachers think about content when assessing performance in the oral exam.

The idea of generalizable patterns, on the other hand, pertains to a number of other findings which are not so much a matter of theoretical representativeness, but rather of empirical resemblance. Thus, the findings including the general agreement among the teachers on the main constructs to be tested, the variations in orientations towards native speaker pronunciation, the indications of construct irrelevant variance and construct underrepresentation, may be seen as patterns that are transferable from the specific research situation – which resembles authentic oral exam situations – to real-life, GSP1/VSP2 oral English exams (cf. Yin, 2016, p. 106).

4.9 Short summary
In this chapter I have shown how the research design fits the nature of the study and critically evaluated aspects of the design which may compromise the inferences that can be drawn from the results. I have also given examples of such inferences. In the next and final chapter I elaborate on these inferences, summing up the main findings from the three articles and point to some possible implications and avenues for further research.
Chapter 5: Summary and discussion

5.1 Introduction to the chapter
As stated in section 1.4, the main aim of this investigation is to explore EFL teacher raters’ understanding of the constructs to be tested in an oral English exam in Norway. In addition, the study examines aspects of rater behaviour and the correspondence between the teachers’ understanding of constructs and the intended constructs as specified by the English subject curriculum and accompanying government documents. The purpose of the inquiry is to provide empirical evidence of these aspects of the rating process in an assessment context which is poorly researched. In this final chapter I present and discuss the main findings of the study in light of the purpose and the research focus of the investigation. I start by providing a summary of the three articles and continue with a discussion of the contribution of the study to the body of research literature on oral L2 assessment generally and the Norwegian context in particular, as well as some potential implications for EFL teaching and assessment in Norway. Finally, I suggest some avenues for further research.

5.2 Summary of the articles
5.2.1 Article 1
Article 1, entitled “Assessing spoken EFL without a common rating scale: Norwegian EFL teachers’ conceptions of construct”, was published in Sage Open, October-December, 2015. The aim of the article was to explore Norwegian EFL teachers’ general assessment orientations in the GSP1/VSP2 oral English exam. Two research questions (RQs) were addressed:

RQ1. How do EFL teachers in Norway understand the constructs and criteria to be tested in an oral exam at the upper secondary level?

RQ2. What kind of criteria do these teachers see as salient when assessing performance?

The data was gathered from semi-structured interviews with 24 teachers who had been asked to share their assessment orientations after having seen a video-taped performance of a student taking the GSP1/VSP2 oral English exam (cf. section 4.2.3). The analyses of the interview transcripts showed that the teachers paid attention to a large number of different
performance aspects. In fact, the number of categories developed was so substantial \(n=56\) that I was unable to report on all of them. Hence, I concentrated on the 38 categories that were common to all the three exam tasks (cf. section 4.2.2).

As for the teachers’ overall understanding of what was to be tested, the results showed that the informants focused on two main constructs, namely *Communication* and *Content*. These two constructs comprised a number of sub-categories, where *Linguistic competence* (belonging to Communication) and *Application, analysis, reflection* (belonging to Content) turned out to be the most important. Linguistic competence, in turn, consisted of three large sub-categories, namely Grammar, Vocabulary and Phonology, Phonology being the most substantial of the three.\(^{22}\) In general, there was evidence that the teachers largely understood the main constructs in the same way, but that there was some variation with regard to the more finely grained performance aspects, such as Phonology. This finding corresponds with those of Brown et al. (2005) and Borger (2014), which indicated overall agreement on the main aspects to be tested, but some disagreement on the more narrowly defined performance features.

One notable difference in teacher orientations, however, was the importance attributed to Content. Although the informants largely understood Content in the same way – conceptualized as a Bloom-like taxonomy of analysing and reflecting on subject matter, as well as the ability to address the task or topic question – the analysis showed that the teachers weighted this construct differently. More specifically, there was evidence that the general studies programme (GSP) teachers were more concerned with Content than the vocational studies programmes (VSP) teachers. For example, the GSP teachers clearly penalized the student in the video-clip for not answering the task question properly. It was hypothesized that this could be explained in terms of the GSP teachers being used to working with students who are generally more proficient in English. Such an interpretation is congruent with findings that raters are more likely to pay attention to linguistic features at the lower levels of proficiency, whereas they have a stronger focus on content at the higher levels (Brown et al., 2005; Pollitt & Murray, 1996; Sato, 2012).

Beyond the differences in orientations relating to the Content construct, I also found that some teachers openly admitted that they would assess VSP students more leniently than GSP students, thus explicitly bringing to the fore the debate about the fairness of making the VSP students take the same course and the same exam as the GSP students (cf. section 1.3.3).

\(^{22}\) “Phonology” and “pronunciation” are here being used interchangeably (cf. Chapter 4, footnote 17.)
This point relates to the question of the social consequences of test use (Messick, 1989), as mentioned in section 2.3.3.

The analysis of scoring behaviour, based on the grades awarded, showed that three teachers gave the student in the video-clip a 2, 15 teachers awarded her a 3 and six teachers awarded her a 4 (mean value $M=3.13$; standard deviation $St. dev. = .61$). However, as only one student performance was scored, it was not possible to calculate an inter-rater reliability measure, such as a Cronbach’s alpha, for the grades awarded. No analyses of relationships between scoring behaviour and teacher background variables were reported in Article 1, but statistically significant variation was found in terms of study programme affiliation, as the GSP teachers scored more harshly than the VSP teachers ($r = .50$, $p < .05$). Moreover, the fact that some teachers explicitly stated that they would assess VSP students more leniently shows variation in rater severity which is a threat to the validity of the scoring outcomes.

As for the evaluation of the scoring inference, it must be pointed out that the current investigation is not a validation study. I have not made any comprehensive, systematic inventory of the constructs to be tested as specified by the subject curriculum and defining documents. Yet, when comparing the teachers’ statements with the subject curriculum and the circular UDIR-1-2010, there is clear evidence that some teachers were attending to construct-irrelevant performance features. Four informants explicitly referred to the student’s level of preparedness as a relevant assessment criterion, and five teachers mentioned effort as an aspect to be tested. According to the above-mentioned circular, preparation and effort are not to be assessed (UDIR, 2010). Such a focus on construct-irrelevant features is not uncommon in the assessment or spoken proficiency (Ang-Aw & Goh, 2011; Borger, 2014; Brown, 1995; C.-N. Hsieh, 2011; Lyn A. May, 2006; Orr, 2002)

On the basis of these analyses I argued that the findings point to the problem of not having a common rating scale on the national level. For lack of data, I cannot compare the results in this study to teacher rater cognition and behaviour in a similar setting with a common rating scale. However, there is evidence that such a scale may enhance the validity and reliability of the scores (Fulcher, 2012; Ginther, 2013). In addition, I maintained that the inclusion of a comprehensive content construct – as evidenced by the many references to content in the English subject curriculum – is problematic in an assessment context where the proficiency levels of the students seem to vary a lot. Judging from the teachers’ accounts there are indications that the VSP students are, on average, at a lower proficiency level in English.

---

23 The grades given range from 1 (= ‘fail’) to 6 (= ‘excellent’).
As a consequence, the VSP teachers seem to prioritize language aspects over content features. However, given the design and administration of the oral English exam as it is currently framed, I also argued that teachers should make students explicitly aware of the importance of answering task questions properly and to reflect on the questions that they are given.

5.2.2 Article 2

Article 2 was co-written with Thomas Hansen and entitled “Assessing pronunciation in an EFL context: Teachers’ orientations towards nativeness and intelligibility”. It is currently under review for Language Assessment Quarterly. The article took as a starting point the strong focus on phonology reported in Study 1 (Article 1), and the divergent views on the question of native speaker pronunciation, which were found in the first pilot study (cf. section 4.2.3). These are interesting findings in view of the apparent fact that pronunciation, until recently, has been neglected in ESL/EFL pedagogy and research (Baker, 2013; Talia Isaacs, 2014; Moyer, 2013). Also, it is worth noting that pronunciation instruction and assessment may be said to have been guided by two contradictory principles, the nativeness principle and the intelligibility principle (Levis, 2005). The former holds that the goal of pronunciation pedagogy is to make students achieve native-like pronunciation; the latter implies that the goal is for students to make themselves understood. And, although the intelligibility principle, at least in theory, seems to have been embraced by a large number of teaching practitioners in recent years (Brown et al., 2005; Hansen, 2011; Timmis, 2002), there are indications that teaching and assessment practices are still largely guided by the nativeness principle (Deterding, 2010; Jenkins, 2007; Seidlhofer, 2011). Against this backdrop, we explored Norwegian EFL teachers’ attitudes towards nativeness and intelligibility. In addition, we measured the teachers’ orientations towards the assessment of four specific pronunciation features which have been found to be important for intelligibility, namely segmentals (i.e. individual sounds), word stress, sentence stress and intonation. Study 2 (Article 2) therefore addressed the following three RQs:

RQ1. To what extent do EFL teachers at the upper secondary level in Norway see nativeness as an important criterion in the assessment of pronunciation?

RQ2. To what extent do the teachers see intelligibility as an important criterion in the assessment of pronunciation?

RQ3. To what extent do the teachers see segmentals, word stress, sentence stress and intonation as important in the assessment of pronunciation?
RQ1 and RQ2 were answered with data from both the 24 informants interviewed in Study 1 and from the 46 questionnaire respondents participating in Study 2, totalling 70 teacher participants. RQ3, on the other hand, was only answered with data from the 46 questionnaire respondents. The video-clip used as prompt in Study 1, was also used in this study (cf. section 4.2.4).

As for RQ1, the results showed that the teachers had conflicting views on the issue of nativeness. For example, 11 teachers strongly disagreed and 13 strongly agreed that students should be assessed against a native-speaker norm. However, we also found evidence of ambivalent attitudes towards nativeness. Some of the interview informants advocated acceptance of non-native speaker pronunciation at the same time as they would prefer top-scoring students to approximate a native-speaker accent.

Regarding RQ2, however, there was much more agreement among the teachers. On average they strongly agreed that intelligibility was important. For instance, 37 of the 46 questionnaire respondents strongly or completely agreed that they would automatically mark a student down from the grade 6 if it was difficult to understand what he or she said.

Finally, in terms of RQ3, the analysis showed that the teachers moderately to strongly agreed that segmentals, word stress and sentence stress were important. For example, on a set of questionnaire items tapping into the respondents’ orientations towards the importance of the four pronunciation features for a top score, segmentals turned out to be the most important (median value \( Md=5 \); mean value \( M=4.46 \)), followed by word stress (\( Md=4; M=4.02 \)) and sentence stress (\( Md=4; M=3.67 \)). However, when it comes to the assessment of intonation, the teachers were more cautious. The average response values of \( Md=3 \) and \( M=3.07 \), including 32 respondents who chose the mid-range option (i.e. neither agree nor disagree), indicate that they were either less concerned with, or uncertain of, how to relate to intonation (cf. Dubois & Burns, 1975). An additional analysis of possible relationships between rater background variables and interview and questionnaire responses found significant positive relationships between teacher experience and all the four pronunciation features. Differently put, the more experienced teachers were slightly more concerned with these features than the less experienced teachers.

All in all, the findings in Article 2 indicated that there was agreement among the teachers about intelligibility. Nativeness, on the other hand, was a contentious issue. Some teachers were strongly in favour of assessing performance against a native speaker norm,

\[24 \text{ A five-point Likert scale going from } 1 = \text{«completely disagree» to } 5 = \text{«completely agree» was used to elicit responses.} \]
whereas others were strongly against. In this respect, the results are in accord with other studies showing divergent orientations in this area (Coskun, 2011; Jenkins, 2007; Timmis, 2002). Additionally, there were teachers who displayed ambivalent attitudes to this issue. However, it seems clear that the teachers overall are concerned with a number of pronunciation issues, such as segmentals, word stress and sentence stress. As for intonation, they are more reluctant, either being unsure of its relevance or finding it less important. If the latter is the case, they are in agreement with proponents of the intelligibility principle within the English as a Lingua Franca paradigm, who point to studies showing that intonation is not important for making oneself understood (Deterding, 2010; Jenkins, 2000).

On the basis of these findings, we also argued that the pronunciation construct itself needs to better defined. This includes a clarification of the relationship between nativeness and intelligibility, which we consider to be interrelated, rather than contradictory (cf. Levis, 2005). Moreover, as the reference to pronunciation in the English subject curriculum is rather vague, 25 we suggested that the introduction of a common rating scale with a more clearly operationalized pronunciation construct is considered. At any rate, it is vital that those pronunciation aspects that are important for intelligibility are made clear to raters. Until empirical studies have proven otherwise, we suggested that the features discussed here, such as certain segmentals, word stress and sentence stress, are singled out as salient criteria.

5.2.3 Article 3

Article 3, entitled “Assessing content in a curriculum-based EFL oral exam: The importance of higher-order thinking skills”, has been submitted to Assessment in Education: Principles, Policy and Practice. The starting point of the article was the variability in teacher perceptions on the issue of content, which was reported in Article 1, as well as claims that the content construct is complex and not well understood in all contexts (Frost, Elder, & Wigglesworth, 2012). Based on the finding from Article 1 that content was largely understood in terms of a Bloom-like taxonomy of analysing and reflecting on subject matter, I developed an analytical framework mainly drawing conceptually on Bloom’s revised taxonomy of educational objectives (Anderson & Kratwohl, 2001). This framework divided content into a subject matter dimension and a skills and abilities dimension (cf sections 4.3.3 and 4.5). As the teachers in Article 1 were mostly concerned with the skills and abilities dimension when 25 The 2006 curriculum did not mention it at all, and the 2013 version states that “[The student shall be able to] use patterns of pronunciation, intonation […]” (cf. appendices 1 and 2)
assessing performance, I decided to let the focus of Article 3 be what kind of subject matter aspects they were paying attention to in the oral English exam. Hence, the following RQ was formulated:

What do EFL teachers at the upper secondary school level in Norway perceive as relevant subject matter content to be assessed in the GSP1/VSP2 oral English exam?

As part of the investigation of the validity of the scoring inference (cf. section 1.4), I also decided to use the analytical framework to identify aspects of subject matter in the competence aims of the subject curriculum. This allowed me to compare the intended construct to be tested with the teachers’ own understanding of the construct. Data was collected from verbal protocols and semi-structured interviews involving 10 teachers.

The initial analysis of subject matter elements in the English curriculum showed that several competence aims involve a considerable number of such elements. Examples of these are “Cultural and societal conditions in English-speaking countries” and “English-speaking cultural forms of expression” (cf. Article 3, Table 2). In total, ten subject matter categories were developed from the analysis of the subject curriculum.

The exploration of the teacher statements in the interviews and verbal protocols revealed that the teachers also understood subject matter in very general terms. The analyses yielded 13 subject matter classifications, including categories as different as Personalized knowledge and Indigenous peoples. The former implies that some teachers would accept, at the lower proficiency levels at least, that students describe topics of personal interest. The latter means that they would expect students to display specific knowledge of indigenous peoples (cf. Article 3 – Table 4). However, most of the teachers were quick to point out that, because of the very general nature of the subject matter construct in the curriculum, they do not expect students to remember specific details, such as particular historical events or geographical locations. Rather, the teachers are largely open to assessing whatever is up for discussion. Some of them also report that this relates to the fact that the Norwegian Directorate for Education and Training has specified that the examiners are supposed to help the students to “display their competence”, not to seek out their “lack of competence” (UDIR, 2014a, p. 2, my translation). This position has two consequences. Firstly, it means that subject matter largely becomes an issue of students having general world knowledge. Secondly, and more importantly, the crucial element in the content construct is not the subject matter dimension, but rather the skills and abilities dimension. What is important is that the students are able to analyse and reflect on whichever topic they are presented with. And, the higher up
The investigation of the validity of the scoring inference showed that the teachers’ descriptions of content was generally highly congruent with the subject matter aspects identified in the curriculum. This, however, was not unexpected, given the very general nature of subject matter described in the competence aims. The only discrepancy between the teachers and the curriculum was found in the area of metacognitive strategies, where the teachers appeared to be unsure of their relevance, despite references to this component in one of the competence aims. One teacher, in fact, persistently denied that metacognitive strategies were to be tested in the oral exam. Hence, the intended construct to be tested was underrepresented with regard to this element. Such construct underrepresentation is not uncommon in ESL/EFL speaking tests (Ang-Aw & Goh, 2011; Cai, 2015).

A final analysis, inspired by the finding in Article 1 that the GSP teachers gave more weight to the content construct, showed that the GSP and VSP teachers in Article 3 were equally concerned with the relevance of this construct. The three VSP teachers in the current sample (n=10) all acknowledge the importance of analysing and reflecting on subject matter, as well as answering the task or topic question properly. In this respect, the findings in Article 3 do not support the findings in Article 1. However, the VSP informants disagreed with some of the GSP teachers on what kind of performance is indicative of achievement with regard to a given grade level. For example, one of the GSP teachers insisted that the student in the video-clip had failed to answer the task question, which she regarded as a major flaw in her performance. When I quoted this statement to one of the VSP teachers, she agreed that answering the task is important, but strongly objected to the claim that the student had not answered the task properly. In her view, the student had responded well to the question. In this sense, the results in Article 3 corroborate the findings in Article 1 that VSP instructors are more lenient in their assessment of content than GSP teachers. Moreover, this result is consistent with research findings showing that raters give different scores for the same performance (Douglas, 1994; Eckes, 2009; Orr, 2002).
metacognitive strategies components, as well as the disagreement on what kind of performance is characteristic of a given level in relation to a criterion. Thus, in order to further strengthen the validity of the scoring inference, I suggested that the educational authorities consider the introduction of a common rating scale, or at least common scoring guidelines, as this may improve the validity of the scores (Fulcher, 2012). A related point is the introduction of more rater training, which seems to be arbitrarily organized in the Norwegian educational system. Some of the teachers in the studies comprising this thesis report that they have never had any rater training at all. Such training may have a positive effect on the consistency of the scores (Taylor & Galaczi, 2011).

5.3 Research contribution

The main research contribution of the present investigation is increased knowledge of how EFL teacher raters assess performance in an English L2 assessment context at the upper-intermediate proficiency level (CEFR level B1/B2). The three studies point towards ways in which Norwegian EFL teachers understand the constructs to be assessed. Additionally, they provide empirical evidence of aspects of scoring behaviour, as well consistency between the teachers’ understanding of constructs and the intended constructs to be tested.

Regarding empirical contributions, three main findings related to the teachers’ understanding of constructs must be mentioned. First of all, this thesis has shown that the teachers, overall, have the same understanding of the main constructs to be tested, but that they differ with regard to more finely grained assessment criteria, such as phonology. Lack of research findings in similar contexts, i.e. curriculum-based school settings without a common rating scale, makes direct comparisons with other studies difficult. The only enquiry examining exactly the same context, albeit with a smaller teacher sample and a more general research focus (i.e. Yildiz, 2011), found that the teachers heeded similar constructs as the teachers in the present investigation, but that there was some more variation in the teachers’ orientations. Apart from Yildiz’ study, Brown et al. (2005), investigating an EAP context without a rating scale, and Borger (2014), exploring a Scandinavian upper secondary school context with a common rating scale, are the studies identified in the literature review which most strongly resemble the current one (cf. Chapter 3). Both corroborate the conclusions arrived at here.

Secondly, the results of the present thesis support findings that teachers tend to heed linguistic performance aspects at the lower proficiency levels, whereas they put more
emphasis on content at the higher levels of proficiency (Brown et al., 2005; Pollitt & Murray, 1996; Sato, 2012). For example, some of the teachers working exclusively with VSP students, who are reportedly at a lower proficiency level in English, stress that they prioritize language-related aspects when assessing performance.

Thirdly, the investigation shows that the pronunciation construct is a complex one, in the sense that teachers disagree on the relevance of applying a native speaker standard when judging phonology. This is consistent with conclusions drawn in similar studies (Brown et al., 2005; Deterding, 2010; Jenkins, 2007; Timmis, 2002). What was particularly interesting here was that 19 per cent of the teachers were strongly in favour of using a native speaker standard, whereas 16 per cent were strongly against. This suggests that nativeness is a contentious issue among many English teachers in Norway. Such conflicting views have also been reported elsewhere (Jenkins, 2007; Jenkins et al., 2011). Above and beyond this, the findings in Article 2 support the contention that pronunciation has been neglected in language teaching pedagogy (Baker, 2013; Derwing & Munro, 2009; Isaacs, 2014) in the sense that the subject curriculum is surprisingly vague in its reference to this construct.

As for scoring behaviour, no measure of inter-rater reliability could be given due to the fact that only one video-taped student performance was used. However, the analysis showed that the GSP teachers were significantly harsher in their grading than the VSP teachers, which means a threat to the reliability of the scores. When correcting for this difference, the results indicate acceptable levels of rater consistency, especially when considering that the scoring of performance always involves two teacher raters (Henning, 1996). This is in line with studies on inter-rater reliability, which generally show high levels of consistency (McNamara, 1996; Eckes, 2011; see also Fulcher, 2003), but the lack of a common rating scale in the present study, as well as the general lack of rater training reported by the teachers, makes this finding original.

The comparison between aspects of the teachers’ understanding of constructs and the intended constructs as specified by the English subject curriculum and other government documents showed generally good correspondence, considering the lack of rater training and rating scale. This corresponds to the findings made by Brown et al. (2005) and Borger (2014). Construct irrelevant assessment criteria included preparation and effort. Construct underrepresentation was identified in the teachers’ reluctance to include metacognitive strategies when assessing performance. Failure to heed the constructs properly in this way is quite common in oral ESL/EFL assessment (Ang-Aw & Goh, 2011; Brown, 1995; Cai, 2015;
C.-N. Hsieh, 2011; Orr, 2002), but there is little evidence on this issue in curricular-based contexts where no scoring guidelines are provided.

One *theoretical contribution* provided in this thesis concerns the conceptualization of the content construct. Applying Bloom’s revised taxonomy of educational objectives (Anderson & Kratwohl, 2001), I argued in Article 3 that the teachers to a large extent understand content as consisting of a subject matter (or *what*) dimension and a skills or abilities (or *how*) dimension. Furthermore, given the very general nature of the subject matter dimension in the curriculum, I maintained that the teachers largely understand subject matter in terms of *general world knowledge*. Moreover, they tend to assess the skills or abilities dimension in terms of whether students are able to recount or describe subject matter – at the lower levels of achievement – or whether they are able to analyse and reflect on subject matter, at the higher levels of achievement. Of the two dimensions, the skills and abilities dimension is the more important one, as the teachers primarily tend to focus on the extent to which students possess higher-order thinking skills. This is a perspective which may theoretically generalize to other contexts (Mitchell, 1983) in the sense that it provides a highly relevant conceptual tool for assessing the content construct.

Another theoretical contribution of this thesis relates to the claim put forward in Article 2 that nativeness and intelligibility are interrelated rather than contradictory principles (cf. Levis, 2005). As was argued in Article 2, alternative approaches to native speaker pronunciation pedagogy tend to take as a starting point native speaker features when attempting to define what is important for intelligibility and what is not. The suggestions made within the English as a Lingua Franca paradigm, which stresses intelligibility over nativeness, is a case in point. These suggestions are based on a number of features taken from Received Pronunciation and General American. Hence, the notion that the nativeness principle and the intelligibility principle are in incongruous, appears paradoxical. Rather, the two principles seem interrelated, although their relationship may be somewhat blurred. I therefore suggest that more research be undertaken to clarify this association (cf. Article 2).

### 5.4 Implications for the Norwegian educational context

Given the fact that there was no common rating scale in the current study, as well as a reported lack of training by many teachers, one may conclude that the teachers’ agreement on the main constructs to be tested is acceptable. However, there are instances of construct underrepresentation and construct-irrelevant variance which do pose a threat to the validity of
the scores, such as the attention paid to effort and the failure to heed metacognitive strategies. In addition, the fact that some teachers differ in what kind of performance they consider as indicative of achievement on the different levels of proficiency is problematic for score consistency. Examples of such differences were found, for example, in the assessment of content. The analysis of scoring behaviour also shows significant differences between the GSP teachers and VSP teachers, the VSP teachers being more lenient. However, when controlling for the variable of study programme affiliation, the score consistency may be regarded as satisfactory, especially when considering that two teacher raters will always be involved in the scoring of performance in the oral exam.

Against this background, the question of standardization, which has been a recurring issue throughout this thesis, needs to be considered. This question first of all concerns the lack of a common rating scale (cf. sections 1.1, 1.3.3 and 2.4). As touched upon in section 2.4, the question of standardization may be regarded as a continuum, depending on the purpose of the test, the constructs to be assessed and the extent to which validity or reliability is emphasized (AERA, APA, & NCME, 2014; Harlen, 2012). It also depends on paradigm perspective (cf. section 1.2). From a measurement paradigm viewpoint dispensing with rating scales may seem rather peculiar, since they are widely considered to contribute to improving the reliability of scores (Brown, 2012; Davis, 2015; Fulcher, 2012; Ginther, 2013; Taylor & Galaczi, 2011; Weigle, 1998). Also, it may be seen as somewhat puzzling that rating scales and rater training are not provided for the oral exam, since they are provided in the corresponding written English exam. In fact, in language performance testing they are hardly ever dispensed with (Fulcher, 2012; Ginther, 2013).

However, from an assessment paradigm perspective, rating scales are not necessarily the universal solution to increased quality in educational contexts. According to Baird et al. (2014), for example, assessment is intrinsically linked to teaching and learning and should not be treated in isolation (p. 82). This relates to the notion that rating scales may not be able to capture the complexities of what is to be tested, especially when the frame of reference for the definition of the constructs is a comprehensive subject curriculum (cf. sections 2.2 and 2.4). A similar issue concerns the objection that the introduction of a rating scale may blur the distinctive character of the course of study, causing the criteria to “create a new structure in the subject” (Thronsden et al., 2009, p. 109, my translation). Relatedly, a less formalized structure gives teachers the opportunity to integrate valuable learning outcomes which may not be specified in the English subject curriculum, but which are included in the Core Curriculum (R. Jensen, personal communication, April 7, 2016). Finally, there is the
interactionist perspective (cf. section 2.5.1), which holds that communicative competence is co-constructed in local contexts, which makes the application of universal criteria problematic (Chalhoub-Deville, 2003; Chapelle, 1998; He & Young, 1998, see also Bachman, 2007).

On the basis of this consideration, and the threats to validity and reliability identified in the current investigation, I side with Thronsden et al. (2009), who suggest that national rating scale guidelines, which may be locally adapted, are introduced to support the scoring of performance in the GSP1/VSP2 exam. Such guidelines have, in fact, been implemented in Norwegian lower secondary schools and provide teachers with common operationalizations of the main constructs at the same time as broader assessment concerns may be included in cases where this is deemed relevant. In addition, I suggest that the Norwegian educational authorities offer rater training on a more systematic basis. Given the positive effects of such training (Barnwell, 1989; Brown, 2012; Taylor & Galaczi, 2011), it may strengthen the already established focus on a “shared assessment culture”.

Along with the question of standardization, another recurrent issue in this thesis is the consequences of letting GSP and VSP students take the same exam. As mentioned in section 1.3.3, this has been a controversial topic for a number of years. The analyses conducted in the current investigation show clear indications of rater variability between GSP and VSP teachers. Not only are VSP teachers more lenient in their scoring, some of them also openly admit that they will apply non-relevant criterion information, such as effort, in the assessment of performance in order to let VSP students pass the exam. In addition, there are teachers working with both student groups who indicate that they score GSP students more harshly than VSP students because they find the system unfair on the VSP students. However, considering the fact that measures to make English more relevant for vocational students were introduced after the data for the present investigation was collected (cf. section 1.3.3), it may be that the teachers’ attitudes towards this issue have changed. Yet, the matter should be further looked into, as this type of rater variability poses one of the most serious threats to the validity of the scores in this thesis. This relates to the social consequences of test use (cf. section 2.3.3).

26 The Norwegian educational authorities’ focus on a «shared assessment culture» (Norw. ‘tolkingsfellesskap’) entails the voluntary development, by the teachers, of “a common understanding of the subject curriculum, and the assessment of competence, through dialogue and discussion” (UDIR, 2014c, p. 1, my translation).
5.5 Concluding remarks

As pointed out by Alderson and Bachman in the quote given in the introduction to this thesis (p. 1) speaking is probably the most difficult skill to assess reliably. Against this fact, and the fact that there is neither a common rating scale, nor systematic rater training in the assessment context explored in this thesis, I conclude that the teachers involved in this study had surprisingly similar views on how to assess performance in the GSP1/VSP2 oral English exam. One may speculate that this is caused by competent teachers who have been working consciously to develop a “shared assessment culture” (cf. footnote 26). Nevertheless, there is clear evidence of rater variability, and I therefore suggest that the educational authorities consider introducing common rating scale guidelines on the national level, as well as more systematic rater training.

In addition, more research is needed. One area in which more research should be conducted is pronunciation, particularly in terms of the relationship between nativeness and intelligibility and the issues of ‘correctness’ and ‘error’. As has been pointed out in this thesis, the operationalization of the pronunciation construct is problematic, and more studies are required in order to clarify which phonological features that should be prioritized in teaching and assessment. Another underresearched area concerns the question of how to assess content, specifically regarding the interface between language and content (cf. Snow & Katz, 2014). Finally, as regards the Norwegian context, it would be relevant to investigate how Norwegian EFL teachers assess oral English when producing oral achievement marks, as these make up a substantial proportion of the students’ final English grade at the upper secondary school level.
References


77


Appendix 1

Utdanningsdirektoratet

English subject curriculum – 2013 (abridged)

Dette er en oversettelse av den fastsatte læreplanteksten. Læreplanen er fastsatt på bokmål.

Established as a Regulation by the Ministry of Education and Research on 21 June 2013

Valid from 01.08.2013

Purpose

English is a universal language. When we meet people from other countries, at home or abroad, we need English for communication. English is used in films, literature, songs, sports, trade, products, science and technology, and through these areas many English words and expressions have found their way into our own languages. When we want information on something of private or professional interest, we often search for it in English. In addition, English is increasingly used in education and as a working language in many companies.

To succeed in a world where English is used for international communication, it is necessary to be able to use the English language and to have knowledge of how it is used in different contexts. Thus, we need to develop a vocabulary and skills in using the systems of the English language, its phonology orthography, grammar and principles for sentence and text construction and to be able to adapt the language to different topics and communication situations. This involves being able to distinguish between oral (spoken) and textual (written) styles and formal and informal styles. Moreover, when using the language for communication we must also be able to take cultural norms and conventions into consideration.

Language learning occurs while encountering a diversity of texts, where the concept of text is used in the broadest sense of the word. It involves oral and written representations in different combinations and a range of oral and written texts from digital media. When we are aware of the strategies that are used to learn a language, and strategies that help us to understand and to be understood, the acquisition of knowledge and skills becomes easier and more meaningful. It is also important to establish our own goals for learning, to determine how these can be reached and to assess the way we use the language. Learning English will contribute to multilingualism and can be an important part of our personal development.

In addition to language learning, the subject of English shall contribute to providing insight into the way people live and different cultures where English is the primary or the official language. The subject of English shall provide insight into how English is used as an international means of communication. Learning about the English-speaking world and the increasing use of English in different international contexts will provide a good basis for understanding the world around us and how English developed into a world language. Literary texts in English can instil a lifelong joy of reading and a deeper understanding of others and of oneself. Oral, written and digital texts, films, music and other cultural forms of expression can further inspire personal expressions and creativity.
Thus, English as a school subject is both a tool and a way of gaining knowledge and personal insight. It will enable the pupils to communicate with others on personal, social, literary and interdisciplinary topics. The subject shall help build up general language proficiency through listening, speaking, reading and writing, and provide the opportunity to acquire information and specialised knowledge through the English language. Development of communicative language skills and cultural insight can promote greater interaction, understanding and respect between persons with different cultural backgrounds. Thus, language and cultural competence promote the general education perspective and strengthen democratic involvement and co-citizenship.

**Main subject areas**

The subject of English is structured into main subject areas with competence aims. The main subject areas supplement each other and must be considered together.

The subject of English is a common core subject for all the upper secondary education programmes. Learning in this subject shall therefore be made as relevant as possible for pupils by adapting each subject to the different education programmes.

English has competence aims after the second, fourth, seventh and tenth years in primary and lower secondary school and after the first year in the programmes for general studies (Vg1) or after the second year of a vocational education programmes (Vg2).

**Overview of main subject areas:**

<table>
<thead>
<tr>
<th>Year</th>
<th>Main subject areas</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–10</td>
<td>Language learning</td>
</tr>
<tr>
<td>Vg1</td>
<td>Oral communication</td>
</tr>
<tr>
<td>Vg2</td>
<td>Written communication</td>
</tr>
<tr>
<td>(vocational education programme)</td>
<td>Culture, society and literature</td>
</tr>
</tbody>
</table>

**Language learning**

The main subject area Language learning focuses on what is involved in learning a new language and seeing relationships between English, one's native language and other languages. It covers knowledge about the language, language usage and insight into one's own language learning. The ability to evaluate own language usage and learning needs and to select suitable strategies and working methods is useful when learning and using the English language.

**Oral communication**

The main subject area Oral communication deals with understanding and using the English language by listening, speaking, conversing and applying suitable communication strategies. The main subject area involves developing a vocabulary and using idiomatic structures and grammatical patterns when speaking and conversing. It also covers learning to speak clearly and to use the correct intonation.
The main subject area involves listening to, understand and using English in different situations where communication needs to be done orally. General politeness and awareness of social norms in different situations are also an important element. This also involves adapting the language to purposeful objectives and adapting the language to the recipient, i.e. by distinguishing between formal and informal spoken language.

The use of different media and resources and the development of a linguistic repertoire across subjects and topics are also key elements of the main subject area.

**Written communication**

The main subject area Written communication deals with understanding and using English language through reading, writing and using suitable reading and writing strategies.

The main subject area includes reading a variety of different texts in English to stimulate the joy of reading, to experience greater understanding and to acquire knowledge. This involves reading a large quantity of literature to promote language understanding and competence in the use of text. Reading different types of texts can lay the foundation for personal growth, maturation and creativity and provide the inspiration necessary to create texts.

The main subject area includes writing different texts in English in different situations where written communication is necessary to stimulate the joy of writing, to experience greater understanding and to acquire knowledge. This also involves adapting the language to purposeful objectives and to the recipient, i.e. by distinguishing between formal and informal written language. The main subject area involves developing a vocabulary and using orthography, idiomatic structures and grammatical patterns when writing. It also covers creating structure, coherence and concise meaning in texts.

The use of different media and resources and the development of a linguistic repertoire across subjects and topics are also key elements of the main subject area.

**Culture, society and literature**

The main subject area Culture, society and literature focuses on cultural understanding in a broad sense. It is based on the English-speaking countries and covers key topics connected to social issues, literature and other cultural expressions. This main area also involves developing knowledge about English as a world language with many areas of use.

The main subject area involves working with and discussing expository texts, literary texts and cultural forms of expression from different media. This is essential to develop knowledge about, understanding of and respect for the lives and cultures of other people.

**Basic skills**

Basic skills are integrated in the competence aims where they contribute to the development of competence in the subject, while also being part of this competence. In the subject of English, the basic skills are understood as follows:

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

Being able to express oneself in writing in English means being able to express ideas and opinions in an understandable and purposeful manner using written English. It means planning, formulating and working with texts that communicates and that are well structured and coherent. Writing is also a tool for language learning. The development of writing proficiency in English involves learning orthography and developing a more extensive repertoire of English words and linguistic structures. Furthermore, it involves developing versatile competence in writing different kinds of generalised, literary and technical texts in English using informal and formal language that is suited to the objective and recipient.

Being able to read in English means the ability to create meaning by reading different types of text. It means reading English language texts to understand, reflect on and acquire insight and knowledge across cultural borders and within specific fields of study. This further involves preparing and working with reading English texts for different reasons and of varying lengths and complexities. The development of reading proficiency in English implies using reading strategies that are suited to the objective by reading texts that are advancingly more demanding. Furthermore, it involves reading English texts fluently and to understand, explore, discuss, learn from and to reflect upon different types of information.

Numeracy in English means being able to use relevant mathematical concepts in English in different situations. This involves familiarity with units of measure used in English-speaking countries and to understand and to communicate in figures, graphic representations, tables and statistics in English. The development of numeracy in English involves using figures and calculations to develop a repertoire of mathematical terms in English related to daily life and general and technical fields.

Digital skills in English means being able to use a varied selection of digital tools, media and resources to assist in language learning, to communicate in English and to acquire relevant knowledge in the subject of English. The use of digital resources provides opportunities to experience English texts in authentic situations, meaning natural and unadapted situations. The development of digital skills involves gathering and processing information to create different kinds of text. Formal requirements in digital texts means that effects, images, tables, headlines and bullet points are compiled to emphasise and communicate a message. This further involves using digital sources in written texts and oral communication and having a critical and independent attitude to the use of sources. Digital skills involve developing knowledge about copyright and protection of personal privacy through verifiable references to sources.

**Competence aims**

**Competence aims after Vg1 – programmes for general studies and Vg2 – vocational education programmes**

**Language learning**

*The aims of the studies are to enable pupils to*
• evaluate and use different situations, working methods and learning strategies to further develop one’s English-language skills
• evaluate own progress in learning English
• evaluate different digital resources and other aids critically and independently, and use them in own language learning

**Oral communication**

*The aims of the studies are to enable pupils to*

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

**Written communication**

*The aims of the studies are to enable pupils to*

• evaluate and use suitable reading and writing strategies adapted for the purpose and type of text
• understand and use an extensive general vocabulary and an academic vocabulary related to one’s education programme
• understand the main content and details in texts of varying length about different topics
• read to acquire knowledge in a particular subject from one’s education programme
• use own notes to write texts related to one’s education programme
• write different types of texts with structure and coherence suited to the purpose and situation
• use patterns for orthography, word inflection and varied sentence and text construction to produce texts
• produce different kinds of texts suited to formal digital requirements for different digital media
• evaluate different sources and use contents from sources in an independent, critical and verifiable manner

**Culture, society and literature**

*The aims of the studies are to enable pupils to*

• discuss and elaborate on culture and social conditions in several English-speaking countries
• present and discuss current news items from English language sources
• discuss and elaborate on the growth of English as a universal language
• discuss and elaborate on different types of English language literary texts from different parts of the world
• discuss and elaborate on English language films and other forms of cultural expressions from different media
• discuss and elaborate on texts by and about indigenous peoples in English-speaking countries
• select an in-depth study topic within one’s education programme and present this

Assessment

Provisions for final assessment:

Overall achievement assessment

<table>
<thead>
<tr>
<th>Year</th>
<th>Provision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 10</td>
<td>The pupils shall have one overall achievement grade for written work and one overall achievement grade for oral performance.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vg1 Programme for General Studies</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vg2 Vocational Education Programme</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Provision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 10</td>
<td>The pupils may be selected for a written examination. The written examination is prepared and graded centrally. The pupils may also be selected for an oral examination. The oral examination is prepared and graded locally.</td>
</tr>
</tbody>
</table>

| Vg1 programme for general studies |                                                                                           |
| Vg2 vocational education programme |                                                                                           |

Examinations for pupils

<table>
<thead>
<tr>
<th>Year</th>
<th>Provision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 10</td>
<td>The pupils may be selected for a written examination. The written examination is prepared and graded centrally. The pupils may also be selected for an oral examination. The oral examination is prepared and graded locally.</td>
</tr>
</tbody>
</table>

| Vg1 programme for general studies |                                                                                           |
| Vg2 vocational education programme |                                                                                           |
Appendix 2

English subject curriculum – 2006 (abridged)

Competence aims after Vg1 – programmes for general studies and Vg2 – vocational education programmes

Language learning

The aims are that the pupil shall be able to

- exploit and assess various situations, working methods and strategies for learning English
- describe and evaluate the effects of different forms of verbal expression
- assess and comment on his/her progress in learning English
- use a wide selection of digital and other aids independently, including monolingual dictionaries

Communication

The aims are that the pupil 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 specialized themes related to his/her own education programme
- express himself/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
- read texts from different genres and with different objectives
- write formal and informal texts with good writing structure and coherence based on themes that interest him/her and are important for society
- read and write texts related to his/her own education programme
- select and use content from different sources independently, critically and responsibly
- use technical and mathematical information in communication
- produce composite texts using digital media
- select an in-depth study topic within his/her own education programme and present this to the other pupils
Culture, society and literature

The aims are that the pupil 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
Appendix 3

English subject curriculum (2006 – Norwegian version)27

Etter Vg1 – studieforberedende utdanningsprogram og Vg2 – yrkesfaglige utdanningsprogram

Språklæring

Mål for opplæringen er at eleven skal kunne
- utnytte og vurdere ulike situasjoner, arbeidsmåter og strategier for å lære seg engelsk
- drøfte likheter og forskjeller mellom engelsk og andre fremmedspråk og utnytte dette i egen språklæring
- bruke relevant og presis terminologi for å beskrive språkets formverk og strukturer
- beskrive og vurdere egen framgang i arbeidet med å lære engelsk
- bruke et bredt utvalg digitale og andre hjelpemidler, inkludert ettspråklige ordbøker, på en selvstendig måte

Kommunikasjon

Mål for opplæringen er at eleven skal kunne
- beherske et bredt ordforråd
- bruke språkets formverk og tekststrukturer i skriftlige og muntlige framstillinger
- forstå lengre framstillinger i skrift og tale om ulike personlige, litterære, tverrfaglige og samfunnsmessige emner
- trekke ut vesentlige opplysninger fra muntlige og skriftlige tekster og drøfte forfatterens synspunkt og holdninger
- uttrykke seg skriftlig og muntlig på en nyansert, situasjonstilpasset måte, med flyt, presisjon og sammenheng
- velge hensiktsmessige lytte-, tale-, lese- og skrivestrategier tilpasset formål, situasjon og sjanger
- ta initiativ til å begynne, avslutte og holde en samtale i gang
- lese formelle og uformelle tekster i ulike sjangere og med ulike formål
- skrive formelle og uformelle tekster med god struktur og sammenheng om personlige, tverrfaglige og samfunnsmessige temaer
- velge og bruke innhold fra ulike kilder på en selvstendig, kritisk og ansvarlig måte
- bruke teknisk og matematisk informasjon i kommunikasjon
- produsere tekster med sammensatt innhold i digitale medier
- velge et tverrfaglig fordypningsområde innenfor eget programområde og presentere dette

http://www.udir.no/kl06/eng1-02

---

27 The full version of this subject curriculum is retrievable from http://www.udir.no/kl06/eng1-02.
Kultur, samfunn og litteratur

Mål for opplæringen er at eleven skal kunne

- drøfte sosiale forhold, samfunnsforhold og verdier i ulike kulturer i flere engelskspråklige land
- presentere og diskutere internasjonale nyheter og aktuelle hendelser
- gjøre rede for hovedtrekk i utvikling av engelsk fra et anglosaksisk språk til et internasjonalt verdensspråk
- analysere og drøfte en film og et representativt utvalg engelskspråklige litterære tekster fra sjangrene dikt, noveller, roman og skuespill
- drøfte et utvalg engelskspråklige litterære tekster fra ulike deler av verden og ulike tidsepoker, fra 1500-tallet til moderne tid
- drøfte litteratur av og om urfolk i den engelskspråklige verden
- lage og vurdere egne muntlige eller skriftlige tekster inspirert av litteratur og kunst
### Appendix 4

**Literature search for the extended abstract**

<table>
<thead>
<tr>
<th>Database</th>
<th>Search term(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web of science</strong></td>
<td>- Oral language assessment rating √</td>
</tr>
<tr>
<td></td>
<td>- Oral language assessment rater √</td>
</tr>
<tr>
<td></td>
<td>- Oral language test rating √</td>
</tr>
<tr>
<td></td>
<td>- Oral language test rater √</td>
</tr>
<tr>
<td></td>
<td>- Spoken language assessment rating √</td>
</tr>
<tr>
<td></td>
<td>- Spoken language assessment rater √</td>
</tr>
<tr>
<td></td>
<td>- Spoken L2 test rating √</td>
</tr>
<tr>
<td></td>
<td>- Spoken L2 test rater √</td>
</tr>
<tr>
<td></td>
<td>- L2 language assessment rating √</td>
</tr>
<tr>
<td></td>
<td>- L2 language assessment rater √</td>
</tr>
<tr>
<td></td>
<td>- Oral language rater effects √</td>
</tr>
<tr>
<td></td>
<td>- Oral language assessment rater effects √</td>
</tr>
<tr>
<td></td>
<td>- Oral language rater orientations √</td>
</tr>
<tr>
<td></td>
<td>- Rater cognition language assessment √</td>
</tr>
<tr>
<td></td>
<td>- Rater cognition language test √</td>
</tr>
<tr>
<td></td>
<td>- Rater orientation language assessment √</td>
</tr>
<tr>
<td></td>
<td>- English exam teacher assessment √</td>
</tr>
<tr>
<td></td>
<td>- Teacher summative language assessment √</td>
</tr>
<tr>
<td></td>
<td>- Classroom-based language assessment √</td>
</tr>
<tr>
<td></td>
<td>- Teacher-based language assessment √</td>
</tr>
<tr>
<td></td>
<td>- Language assessment speaking rater √</td>
</tr>
<tr>
<td></td>
<td>- Oral language proficiency rater √</td>
</tr>
<tr>
<td></td>
<td>- Rater variability √</td>
</tr>
<tr>
<td></td>
<td>- content-based assessment √</td>
</tr>
<tr>
<td></td>
<td>- curriculum-based assessment √</td>
</tr>
<tr>
<td></td>
<td>- assessment AND “topical knowledge” √</td>
</tr>
</tbody>
</table>

<p>| <strong>Academic search premier</strong> | - “language assessment” AND rating √                                           |
|                           | - “language assessment” AND rater √                                             |
|                           | - “language test” AND rating √                                                  |
|                           | - “language testing” AND rating √                                               |
|                           | - “language testing” AND rater √                                                |
|                           | - “language assessment” AND spoken AND rating √                                 |
|                           | - “language assessment” AND oral AND rating √                                   |
|                           | - “language assessment” AND “rater cognition” √                                 |
|                           | - “language assessment” AND “rater orientation” √                              |</p>
<table>
<thead>
<tr>
<th>Source</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERIC</td>
<td>“language assessment” AND oral AND rating  √</td>
</tr>
<tr>
<td></td>
<td>“language test” AND oral AND rating  √</td>
</tr>
<tr>
<td></td>
<td>“language testing” AND oral AND rater  √</td>
</tr>
<tr>
<td></td>
<td>“language assessment” AND spoken AND rater  √</td>
</tr>
<tr>
<td></td>
<td>“rater cognition” AND “language assessment” √</td>
</tr>
<tr>
<td></td>
<td>“English exam” AND teacher AND assessment  √</td>
</tr>
<tr>
<td></td>
<td>“language assessment” AND teacher AND summative  √</td>
</tr>
<tr>
<td></td>
<td>“language assessment” AND “rater effects”  √</td>
</tr>
<tr>
<td></td>
<td>“language testing” AND “rater effects”  √</td>
</tr>
<tr>
<td>Language Testing</td>
<td>oral AND rating  √</td>
</tr>
<tr>
<td></td>
<td>“rater cognition” AND oral  √</td>
</tr>
<tr>
<td></td>
<td>“rating process”  √</td>
</tr>
<tr>
<td></td>
<td>summative  √</td>
</tr>
<tr>
<td></td>
<td>“curriculum-based”  √</td>
</tr>
<tr>
<td></td>
<td>“rater variability”  √</td>
</tr>
<tr>
<td>Language Assessment Quarterly</td>
<td>oral AND rating  √</td>
</tr>
<tr>
<td></td>
<td>“rater cognition” AND oral  √</td>
</tr>
<tr>
<td></td>
<td>“rating process”  √</td>
</tr>
<tr>
<td></td>
<td>summative  √</td>
</tr>
<tr>
<td></td>
<td>“curriculum-based”  √</td>
</tr>
<tr>
<td></td>
<td>“rater variability”  √</td>
</tr>
<tr>
<td>Language Teaching</td>
<td>“language assessment” AND “subject knowledge”  √</td>
</tr>
<tr>
<td></td>
<td>“language assessment” AND “subject content”  √</td>
</tr>
<tr>
<td></td>
<td>“curriculum-based assessment”  √</td>
</tr>
<tr>
<td></td>
<td>“assessing topical knowledge”  √</td>
</tr>
<tr>
<td>Idunn</td>
<td>muntlig AND vurdering  √</td>
</tr>
<tr>
<td></td>
<td>“vurdering av muntlige”  √</td>
</tr>
<tr>
<td></td>
<td>“muntlig eksamen”  √</td>
</tr>
<tr>
<td></td>
<td>“summative vurdering” AND muntlig  √</td>
</tr>
<tr>
<td></td>
<td>vurdering AND muntlig AND ferdigheter  √</td>
</tr>
</tbody>
</table>
| Google scholar | “language assessment” AND oral AND rating √  
|               | - “language assessment” AND Oral AND rater √  
|               | - “language test” AND oral AND rating √  
|               | - “language test” AND oral AND rater √  
|               | - “language testing” AND oral AND rater √  
|               | - “language assessment” AND spoken AND rating √  
|               | - “language assessment” AND spoken AND rater √  
|               | - “rater cognition” AND “language assessment” √  
|               | - “rater cognition” AND “language assessment” √  
|               | - “English exam” AND teacher AND assessment √  
|               | - “language assessment” AND teacher AND summative √  
|               | - “language assessment” AND “rater effects” √  
|               | - “language testing” AND “rater effects” √  
| The Directorate for Education and Training (www.udir.no) | - muntlig AND vurdering √  
|               | - “vurdering av muntlige” √  
|               | - “muntlig eksamen” √  
|               | - “summative vurdering” AND muntlig √  
|               | - vurdering AND muntlig AND ferdigheter √  

| Other sources: | - Companion to Language Assessment, 2014  
|               | - Dobson, Eggen, and Smith (2009)  
|               | - Fulcher (2003)  
|               | - Fulcher (2010)  
|               | - Fulcher and Davidson (2007)  
|               | - Anthony Green (2014)  
|               | - T. McNamara (1996)  
|               | - Thronsen et al. (2009)  
|               | - Master & PhD Theses  

Note: The list of sources is not exhaustive. In the earlier stages of the PhD-project research studies were also located elsewhere, such as in the web pages of the Norwegian National Centre for Foreign Languages in Education (www.fremmedspraksenteret.no), in the public media, and in other digital and printed literature.
References


Appendix 5

Pilot questionnaire

Assessing oral English – GSP1/VSP2 level

This questionnaire is designed to gather data for a master’s degree project and a PhD-project on the assessment of oral English at the GSP1/VSP2 level. It will give valuable information on how teacher raters assess oral English performance. You have just seen a video-clip of two students in a mock oral English exam situation. How would you score their performances?

Grade: _____

Please answer the questionnaire by indicating to what extent you agree with the different statements listed. Note that the statements concern the assessment of oral English performance generally, and not the specific performances of the two students you have just seen. The questionnaire has three sub-sections: one concerning the oral English exam generally, one concerning the presentation part of the exam and one concerning the discussion part (between the candidate and the interlocutor). We would also be very pleased if you could comment on the appropriateness of the different statements in the questionnaire.

1. Background questions

1.1 Have you ever examined your own students in an oral English exam?  Yes [ ]  No [ ]
1.2 Have you ever been interlocutor at the GSP1/VSP2 level?  Yes [ ]  No [ ]
1.3 Have you ever been external examiner at the GSP1/VSP2 level?  Yes [ ]  No [ ]
1.4 Are you only employed in the VSP studies programme?  Yes [ ]  No [ ]
1.5 Are you only employed in the GSP study programme?  Yes [ ]  No [ ]
1.6 Are you employed both in the GSP and the VSP studies programmes?  Yes [ ]  No [ ]
1.7 Have you ever had any formal rater training?  Yes [ ]  No [ ]

---

28 The questionnaire has been translated from Norwegian.
1.8 How often have you been interlocutor / external examiner in the Oral English exam at the GSP1/VSP2 level?

- Never
- 1-2 times
- 3-5 times
- 6 times or more

2. Assessing performance in the oral English exam generally

2.1 Language aspects

<table>
<thead>
<tr>
<th></th>
<th>Completely disagree 1</th>
<th>2</th>
<th>3</th>
<th>Completely agree 4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.1 Good vocabulary is important in order to achieve a top grade.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.2 Clear pronunciation is important in order to achieve a top grade.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.3 Good “native speaker” pronunciation is important in order to achieve a top grade.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.4 God fluency is important in order to achieve a top grade (i.e. coherent speech without too much hesitation and unnatural pauses).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.5 The student must adapt his/her language to the situation in order to achieve a top grade (i.e. language use which corresponds to the genre conventions for a formal setting).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.6 Grammatically correct language use is important in order to achieve a top grade.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.7 The use of cohesive ties is important in order to achieve a top grade.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.1.8 COMMENTS / OTHER IMPORTANT ASPECTS:

_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________

2.1.9 Would you say that any of these aspects (criteria) are more important in the scoring of performance than others?

- Yes
- No
- Don’t know
If **yes**, which three criteria (vocabulary, clear pronunciation, fluency etc.) in section 2.1 do you consider the most important (1 meaning most important and 3 meaning least important)?

1. ______________________
2. ______________________
3. ______________________

3. The presentation part of the exam

3.1 Content

<table>
<thead>
<tr>
<th>Completely disagree</th>
<th>Completely agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

**3.1.1** Relevance and range of content is important in order to achieve a top grade.

**3.1.2** A creative presentation design is important in order to achieve a top grade.

**3.1.3** The ability to reflect on the material presented is important in order to achieve a top grade.

**3.1.4** Good use of sources is important in order to achieve a top grade.

**3.1.5** Candidates who have obviously prepared well (without this being necessarily reflected in the performance) should be credited for this.

**3.1.6 COMMENTS / OTHER IMPORTANT ASPECTS:**

_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________
### 3.2 Visual aids

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3.2.1</strong> Good use of visual aids is important in order to achieve a top grade.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.2.2</strong> The use of technically advanced aids (e.g., using a number of advanced features of the PowerPoint program) will affect the score in a positive way.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.2.3</strong> Visual aids (PowerPoint etc.) containing a number of language errors will mark the student down considerably.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.2.4</strong> The ability to paraphrase and to elaborate on, for example, the bullet points in the PowerPoint slides is important for a top score.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.2.5</strong> Variation in presentation modes (i.e., swapping between different media, such as text, pictures etc.) is important for a top score.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**COMMENTS / OTHER IMPORTANT ASPECTS**

________________________________________________________________________________
________________________________________________________________________________
________________________________________________________________________________

### 3.3 Presentation skills

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3.3.1</strong> Good fluency, pace and rhythm is important for a top grade.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.3.2</strong> The ability to free oneself from the manuscript is important for a top grade.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.3.3</strong> To perform in a creative way is important in order for a top grade.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.3.4</strong> To speak sufficiently loudly and clearly is important for a top grade.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.3.5</strong> Appropriate body language (including good eye contact) is important for a top grade.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.3.6</strong> Good presentation structure is important for a top grade.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.3.7</strong> Candidates who do not show signs of nervousness during the presentation should be credited for this.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.3.8</strong> Candidates who display a high level of engagement during the presentation should be credited for this.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.3.9</strong> Candidates who have good digital skills should be credited for this.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.3.10 COMMENTS / OTHER IMPORTANT ASPECTS:
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________

3.1.11 Would you say that any of these aspects (criteria) are more important in the scoring of performance than others?

Yes ☐ No ☐ Don’t know ☐

If yes, which three criteria (vocabulary, clear pronunciation, fluency etc.) in section 2.1 do you consider the most important (1 meaning most important and 3 meaning least important)?

1. ______________________
2. ______________________
3. ______________________

4. Discussion part

4.1 Content

<table>
<thead>
<tr>
<th></th>
<th>Completely disagree</th>
<th>Completely agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1.1 The ability to retell facts is important for a top grade.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1.2 The ability to reflect on facts is important for a top grade.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1.3 The ability to draw on syllabus texts is important for a top grade.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1.4 Relevant subject matter content is important for a top grade.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.1.5 COMMENTS / OTHER IMPORTANT ASPECTS:
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________
4.2 Communicative interaction

<table>
<thead>
<tr>
<th></th>
<th>Completely disagree</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Completely agree</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4.2.1</strong> The ability to listen well is important in order to achieve a top grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>4.2.2</strong> Candidates who take the initiative to talk and who show willingness to participate in the conversation should be credited for this.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>4.2.3</strong> If a candidate gives very little response, this will negatively affect the score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**4.3.4 COMMENTS / OTHER IMPORTANT ASPECTS:**
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________

**4.2.5** Would you say that any of these aspects (criteria) are more important in the scoring of performance than others?

- Yes ☐
- No ☐
- Don’t know ☐

If yes, which three criteria (vocabulary, clear pronunciation, fluency etc.) in section 2.1 do you consider the most important (1 meaning most important and 3 meaning least important)?

1. ______________________
2. ______________________
3. ______________________

Do you have other comments on the questionnaire?
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________
_______________________________________________________________________________________________________________

Thanks for your time!
Appendix 6

Pilot interview guide

The assessment of oral EFL at GSP1/VSP2 oral English Exam

1. BACKGROUND:
   1.1 Age:
   1.2 L1:
   1.3 Education:
   1.4 Experience (number of years at the upper secondary level):
   1.5 Experience as rater at the oral exam (fra vg1/YF2):
   1.6 Study programme affiliation: General studies ___ Vocational studies ___
   1.7 Have you worked outside of your county?
   1.8 Have you attended rater training?

2. How well would you say that your school works with the identification of criteria for assessment in English (on the basis of the subject curriculum)?

3. Do you have a common written rating scale for the assessment of oral English (developed either at your school or in your county)?

4. Which criteria do you apply generally when rating performance in the oral English exam?

5. Which of these criteria are the most important, would you say?

6. To what extent is correctness in pronunciation important?

7. Do you assess holistically or analytically?

8. Some teachers say that they are more lenient when scoring students in the vocational studies programmes compared with students in the general studies programmes. What is your comment on that?

---

29 The questions have been translated from Norwegian.
Forespørsel til elever om deltagelse i forskningsprosjekt vedrørende muntlig eksamen i engelsk

Undertegnede jobber for tiden med et forskningsprosjekt om vurdering av muntlig engelsk i videregående skole. Prosjektets formål er å se på hvordan muntlig eksamen i engelskfaget gjennomføres på forskjellige skoler og hvordan sensorer vurderer de enkelte kandidatene.

For å kunne dokumentere dette nærmere ønsker jeg å komme i kontakt med elever som er villige til å la seg filme under muntlig eksamen – dersom de skulle bli trukket ut i engelsk. Det legges opp til at filmingen gjøres så diskret som mulig slik at ikke elevens prestasjon under eksamen skal bli påvirket. Undertegnede vil ikke være til stede; kun et kamera vil bli satt opp. Alle opplysninger vil bli behandlet konfidensielt og personlige opplysninger vil ikke bli lagret i noe register. Videomaterialet skal brukes i forskningsøyemed, dvs. som utgangspunkt for forskning på sensorers karaktersetting, og til sensorskoloring, dvs. for at sensorer skal kunne trenes i å gi best mulig vurdering. Du vil ved å si ja til dette, kunne gi et viktig bidrag til vurderingsforskningen i Norge.

Dersom du skulle si ja til dette, gjøres det oppmerksom på at du når som helst – både før, under og etter eksamen – kan trekke deg fra deltagelse i prosjektet.

Dersom du er under 18 år, er det også nødvendig med bekreftelse fra en av dine foresatte.

Med vennlig hilsen
Henrik Bøhn

☐ Ja, jeg gir tillatelse til filming under muntlig eksamen i engelsk dersom jeg skulle bli trukket ut.

_________________________  _______________________
Navn                                                                                                      Skole
_________________________  _______________________
Dato                                                                                                      Underskrift
_________________________  _______________________
Dato                                                                                                      Foresattes underskrift (hvis under 18 år)
Forespørsel til lærere om deltakelse i forskningsprosjekt vedrørende muntlig eksamen i engelsk

Undertegnede jobber for tiden med et forskningsprosjekt om vurdering av muntlig engelsk i videregående skole. Prosjektets formål er å se på hvordan muntlig eksamen i engelskfaget gjennomføres på forskjellige skoler og hvordan sensorer vurderer de enkelte kandidatene.


I tillegg til selve filmingen, ønsker jeg å komme i kontakt med eksaminatorer/sensorer som er villige til å la seg intervjuve – alternativt svare på et spørreskjema – om hvordan karakterene har blitt satt.

Dersom du skulle si ja til dette, gjøres det oppmerksom på at du når som helst – både før, under og etter eksamen – vil kunne trekke deg fra deltakelse i prosjektet.

Din deltakelse i et slikt prosjekt vil kunne være et viktig bidrag til lærings- og vurderingsforskningen.

Med vennlig hilsen
Henrik Bøhn

____________________________________________________________________________
Ja, jeg gir tillatelse til filming under muntlig eksamen i engelsk dersom elever fra min klasse skulle bli trukket ut og jeg er eksaminator.

Ja, jeg gir tillatelse til filming under muntlig eksamen dersom jeg er oppnevnt som ekstern sensor.

Ja, jeg sier meg villig til å delta i intervju.

Ja, jeg sier meg villig til å svare på et spørreskjema.

__________________________________________________________________________
Navn                                                                     Dato og underskrift
Appendix 8

Norsk samfunnsvitenskapelig datatjeneste AS
NORWEGAN SOCIAL SCIENCE DATA SERVICES

Henrik Bohn
Avdeling for økonomi, språk og samfunnslag
Høgskolen i Østfold
Remmen
1757 HALDEN

Vær dato: 13.06.2011
Vær ref: 26813 / 773S
Deps dato: Deps ref:

KVITTERING PÅ MELDING OM BEHANDLING AV PERSONOPPLYSNINGER

Vi viser til melding om behandling av personopplysninger, mottatt 23.03.2011. All nødvendig informasjon om prosjektet forelå i sin helhet 12.05.2011. Meldingen gjelder prosjektet:

26813

Sammenligning Assessment of Students' Oral EFL Ability: Validity and Reliability
Considerations in Oral Exams at the Upper Secondary Level

Behandlingsansvarlig
Fylkeshaven i Østfold, ved institusjonens øverste leder

Daglig avfalls
Henrik Bohn

Personvernombudet har vurdert prosjektet og finner at behandlingen av personopplysninger er meldeplyktig i henhold til personopplylingsloven § 31. Behandlingen tilfredsstiller kravene i personopplylingsloven.

Personvernombudets vurdering forutsetter at prosjektet gjennomføres i tråd med opplysningene gitt i meldeskjemaet, korrespondanse med ombudet, eventuelle kommentarer samt personopplylingsloven/-
helseregisterloven med forskrifter. Behandlingen av personopplysninger kan settes i gang.


Personvernombudet vil ved prosjektets avslutning, 01.07.2015, rette en henvendelse angående status for behandlingen av personopplysninger.

Vennlig hilsen,

Bjørn Henrichsen

Tone Njoelstad Slotvik
Part II

Articles
Decloration

Describing the independent research contribution of the candidate

In addition to the dissertation, there should be enclosed a declaration describing the independent research contribution of the candidate for each paper constituting the dissertation.

The declaration should be filled in and signed by candidate and co-authors. Use the back of the page if necessary.

The declaration will show the contribution to conception and design, or development and analysis of a theoretical model, or acquisition of data, or analysis and interpretation of data, contribution to drafting the article or revising it critically for important intellectual content etc.

Article no. 1

Title:
Assessing spoken EFL without a common rating scale: Norwegian EFL teachers’ conceptions of construct

The independent contribution of the candidate:
The article was researched and written solely by Henrik Bohn.

Signature of candidate

Signature of co-authors

Article no. 2

Title:
Assessing pronunciation in an EFL context: Teachers’ orientations towards nativeness and intelligibility.

The independent contribution of the candidate:
Data collection was solely carried out by Henrik Bohn. He also conducted the main analysis and wrote approximately 80 percent of the article. Thomas Hansen carried out minor parts of the analysis and wrote approximately 20 percent of the article.

Signature of candidate

Signature of co-authors
Article no. 3

Title:
The importance of higher-order thinking skills: The assessment of content in a curriculum-based EFL oral exam.

The independent contribution of the candidate:
The article was researched and written solely by Henrik Bøhm.

[Signature]
Signature of candidate

[Signature]
Signature of co-authors