Investigating and Validating Spoken Interactional Competence

GOTHENBURG STUDIES IN EDUCATIONAL SCIENCES 424

Investigating and Validating Spoken Interactional Competence:

Rater Perspectives on a Swedish National Test of English

Linda Borger

UNIVERSITY OF GOTHENBURG
ACTA UNIVERSITATIS GOTHOBURGENSIS

ISBN 978-91-7346-981-4 (print)
ISBN 978-91-7346-982-1 (pdf)
ISSN 0436-1121

Doctoral thesis in Education at the Department of Education and Special Education, University of Gothenburg.
The thesis is also available in full text at:
http://hdl.handle.net/2077/57946

Distribution:
Acta Universitatis Gothoburgensis, Box 22, 405 30 Göteborg
acta@ub.gu.se

Photo: Linda Morgan

Print: BrandFactory AB, Kållered, 2018

Abstract

Title:	Investigating and Validating Spoken Interactional Competence: Rater Perspectives on a Swedish National Test of English
Author:	Linda Borger
Language:	English with a Swedish summary
ISBN:	978-91-7346-981-4 (print)
ISBN:	978-91-7346-982-1 (pdf)
ISSN:	0436-1121
Keywords:	interactional competence, Swedish national test of English, paired speaking test, Common European Framework of Reference for Languages (CEFR), socio-cognitive validation framework

This thesis aims to explore different aspects of validity evidence from the raters' perspective in relation to a paired speaking test, part of a high-stakes national test of English as a Foreign Language (EFL) in the Swedish upper secondary school. Three empirical studies were undertaken with the purpose of highlighting (1) the scoring process, (2) the construct underlying the test format, and (3) the setting and test administration.

In Study I and II, 17 teachers of English from Sweden, using national performance standards, and 14 raters from Finland and Spain, using scales from the Common European Framework of Reference for Languages (CEFR), rated six audio-recorded paired performances, and provided written comments to explain their scores and account for salient features. Inter-rater agreement was analysed using descriptive, correlational and reliability statistics, while content analysis was used to explore raters' written comments. In Study III, 267 upper secondary teachers of English participated in a nation-wide online survey and answered questions about their administration and scoring practices as well as their views of practicality. The responses were analysed using descriptive statistics and tests of association.

Study I revealed that raters observed a wide range of students' oral competence, which is in line with the purpose of the test. With regard to inter-rater agreement, the statistics indicated certain degrees of variability. However, in general inter-rater consistency was acceptable, albeit with clear room for improvement. A small-scale, tentative comparison between the national EFL standards and the reference levels in the CEFR was also made.

In Study II, raters' interpretation of the construct of interactional competence was explored. The results showed that raters attended to three main interactional resources: *topic development moves*, *turn-taking management*, and *interactive listening strategies*. As part of the decision-making process, raters also considered the impact of test-takers' interactional roles and how students' performances were interrelated, which caused some challenges for rating.

Study III investigated teachers' implementation practices and views of practicality. The results revealed variations in how the national speaking test was implemented at the local level, which has clear implications for standardisation but must be considered in relation to the decentralised school system that the national tests are embedded in. In light of this, critical aspects of the setting, administration and scoring procedures of the national EFL speaking tests were highlighted and discussed.

In the integrated discussion, the different aspects of validity evidence resulting from the empirical data are analysed in relation to a socio-cognitive framework for validating language tests (O'Sullivan & Weir, 2011; Weir, 2005). It is hoped that the thesis contributes to the field of speaking assessment in two ways: firstly by showing how a theoretical framework can be used to support the validation process, and secondly by providing a concrete example of validation of a high-stakes test, highlighting positive features as well as challenges to be addressed.

Table of Contents

Acknowledgements

List of Figures

List of Tables

Acknowledgements

This thesis was completed in two stages. The first stage, leading to a licentiate degree, ended in 2014. In the licentiate thesis (included here as Study I), I expressed my gratitude to a number of people; I am still indebted to all of you, not least to those involved in the graduate school for language education (FRAM) and to Professor Gudrun Erickson and Professor Liss Kerstin Sylvén, who supervised my licentiate thesis. In 2015, I was given the opportunity to continue my PhD studies and finalise this thesis.

I would like to thank a number of people who have helped and supported me throughout this process. First of all, I would like to express my deepest gratitude to my main supervisor, Professor Gudrun Erickson, for continuously providing me with valuable comments and advice on my research, constantly supporting and believing in me. Your warm encouragement and guidance have been invaluable in completing this thesis. Thank you also for sharing your love for teaching with me.

I would also like to express my gratitude to my co-supervisor, Professor Monica Rosén, for your encouragement, insightful comments and perceptive questions which helped improve this thesis. Thank you also for methodological guidance and for sharing your expertise within the field of quantitative methods in education. I am looking forward to continuing this line of work in the future.

Furthermore, I would like to sincerely thank the participants of this research for their time and valuable contribution to the validation process. Without you this thesis would not have been possible. The thesis also benefited from constructive comments and useful suggestions made by the late Professor Sauli Takala and Professor Dina Tsagari who were discussants at my licentiate and final seminars. I am particularly grateful to Dr Eva Olsson and Dr Henrik Bøhn for their assistance in the development of the coding scheme. Additionally, I am thankful for constructive comments and valuable feedback on initial versions of the questionnaire given by colleagues both within and outside the university. Special thanks go to Marianne Demaret for technical help and advice and to Agneta Edvardsson for kind help in administrative matters. Thanks are also due to Associate Professor Gun-Britt Wärvik, former director of doctoral students, for support and guidance throughout my PhD studies.

I have truly enjoyed being a PhD student at the Department of Education and Special Education and would like to thank my colleagues, including my fellow doctoral students, for providing a friendly and supportive working

environment. I especially value the support and interest shown in my work by colleagues in the National Assessment Project. A warm thanks also goes to my roommates Johanna and Dimitrios.

Thanks to colleagues and friends at Katedralskolan in Linköping, especially Anders, Karl and Maria, for keeping in touch and for showing an interest in my research. I am also very grateful to Håkan, Peter, and Ingela for your friendship. Without our book club meetings, writing this thesis would not have been as stimulating.

Finally, I would like to thank my family. I will always be grateful for the loving support of my mother. You are my greatest inspiration. Gustav and Fredrik, my two boys, thank you for bringing me the greatest joy in life. Last but not least, I would like to express my love to Nils for supporting me throughout this journey. Thank you for your patience, encouragement and for being my best friend.

Linköping and Gothenburg, November, 2018

Linda Borger

Chapter One: Introduction

Practice and research in assessing speaking is regarded as "the youngest subfield of language testing" (Fulcher, 2003, p. 1). The assessment of oral competence has developed over the past few decades, leading to a broadening of the speaking construct to include social dimensions of language use (McNamara & Roever, 2006). More authentic and interactive assessment tasks, such as paired or group orals, are now being incorporated in both large-scale and small-scale assessment contexts. Paired and group formats typically "involve candidates interacting together to perform a task while one or more examiners observe their performance and rate their language proficiency" (Van Moere, 2013, p. 1). Testing in groups can be advantageous in many ways and "it opens up the possibility of enriching our construct definition, and hence the meaning of test scores" (Fulcher, 2003, p. 189-190). However, given the complex interaction patterns and the variability displayed in peer-to-peer interaction, the format has also attracted significant criticism (Foot, 1999; Norton, 2005; O'Sullivan, 2011b). Further research is therefore needed to evaluate the use of this test format in different contexts, including the perspective from different stakeholder groups.

In light of this, the present thesis aims to investigate the assessment of a paired speaking test, part of a high-stakes national test of English in the Swedish upper secondary school, from a rater perspective. In particular, attention is drawn to three areas: (1) the scoring process, (2) the construct underlying the test format, and (3) the setting and test administration.

Background

This chapter serves as an introduction to the thesis in its entirety, starting with a construct definition of speaking. After that, a background to national language testing in Europe is given, including definitions of some central concepts.

Defining the construct of speaking

Referring to the term *construct* in the context of language assessment, Bachman and Palmer (2010) make the following observation: "If we are to make interpretations about language ability on the basis of performance on language assessments, we need to define this ability in sufficiently precise terms to distinguish it from other individual attributes that can affect assessment performance. We also need to define language ability in a way that is appropriate *for each particular assessment situation*" (p. 43). The definition of the construct thus (1) describes the fundamental components or aspects of the ability that a given assessment or assessment task intends to measure and (2) provides the basis for interpreting scores derived from the task. In a similar vein, Fulcher (2003) emphasises that test purpose should "drive the definition of the construct, its range and generalisability (p. 19)".

Speaking is considered to be a complex process. Field (2011) even maintains that speaking is "one of the most complex and demanding of all human operations" (p. 70). Fulcher (2003) points out that any construct definition of speaking must be multi-faceted: "however much we may try to define and classify, the kinds of choices that a second language speaker makes are going to be influenced by the totality of their current understanding, abilities (personal and cognitive), language competence and speech situation" (p. 25). Based on Bachman and Palmer's (1996) framework for describing communicative language ability (see further in Study I), Fulcher (2003) summarised components of oral proficiency that "we might wish to include in a construct definition for a test of second language speaking" (p. 49). According to this inventory, oral proficiency includes knowledge of and ability to use:

- *language competence*
    - *phonology*, relating to pronunciation, stress, and intonation
    - *accuracy* in terms of syntax, vocabulary, and cohesion
    - *fluency*, referring to automaticity and ease of speech, determined by aspects such as hesitations, pausing, repetition, and cohesion
- *strategic capacity*, which includes the cognitive capacity to manage communication and refers to "the relationship between the internal processes and knowledge base of the test taker to the external real-time action of communicating" (Fulcher, 2003, p. 33)

- *textual knowledge*, referring to the structure of talk, e.g. turn taking and openings and closings and adjacency pairs
- *pragmatic and sociolinguistic knowledge*, referring to the rules of speaking and pragmatic appropriacy, as well as situational, topical and cultural aspects of spoken language use

Fulcher (2003) observes, with regard to the elements listed above, that "[n]o attempt has been made to isolate separate categories for interactional competence, for as we have seen it is an approach to understanding the co-construction of speech that focuses on turn taking, or openings and closings, rather than suggesting completely new categories that should be included" (p. 49). However, the ability to interact in a meaningful way with other speakers has received a more pronounced role in the conceptualisation of the second/foreign language (L2)[1] speaking construct during the last two decades, as a result of the communicative approach to language learning and assessment. In connection with this, interactional aspects have also been incorporated to a greater extent in rating criteria. The concept of interactional competence (IC) was first introduced by Kramsch (1986) and has later been developed in slightly different versions in several subsequent publications (Hall, 1993, 1995; A. W. He & Young, 1998; Young, 2000, 2008, 2011). At the heart of the conceptualization of interactional competence lies the notion that communication is co-constructed and shared between interlocutors. Another assumption of the theory is that interactional competence is context-dependent and therefore varies with the interactional practice and with the participants (A. W. He & Young, 1998; Young, 2000). These two characteristic features hold obvious implications and challenges for the testing of interactional skills.

McNamara (1997) defines two main perspectives from which a speaking construct for L2 assessment can be conceptualized: "(1) a loosely psychological one, referring to various kinds of mental activity within a single individual, and (2) a social/behavioural one, where joint behaviour between individuals is the basis for the joint construction (and interpretation) of performance" (p. 447). Several applied linguist scholars (e.g., Chalhoub-Deville, 2003; Johnson, 2001; McNamara, 1997; Young, 2000) have pointed out that approaches to L2

[1] The term L2 is used to refer to both foreign and second language. Traditionally, a distinction has been made between foreign language and second language learning and use. Foreign language is defined as the use or study of a foreign language by non-native speakers in a country where this language is not a local medium of communication. Second language, in comparison, is used as a term for the use or study of a second language by non-native speakers in an environment where this language is the mother tongue or an official language.

assessment based on the theory of communicative competence (Hymes, 1972), most notably Canale and Swain (1980) and Bachman and Palmer (1996) (see further in Study I), represent a primarily cognitive or psychological conceptualization of interaction, which makes them less well-suited as frameworks of interactional competence. Young (2011) maintains that interactional competence adds further linguistic and pragmatic components, such as the ability to manage turn-taking, initiate and develop topics and repair interactional trouble, to the other components of communicative competence. However, the fundamental difference between communicative competence and interactional competence is that "an individual's knowledge and employment of these [interactional] resources is contingent on what other participants do; that is, IC is distributed across participants and varies in different interactional practices" (p. 430). In other words, "IC is not what a person knows, it is what a person does together with others in specific contexts" (Young, 2011, p. 430).

Galaczi and Taylor (2018) adhere to this perspective and characterize interactional competence from a socio-cognitive viewpoint (Weir, 2005), according to which:

> speaking is viewed both as a cognitive and a social interactional trait, with emphasis not just on the knowledge and processing dimension of language use, as seen in the Bachman and Palmer (1996) model, but also on the social, interactional nature of speaking, which has as its primary focus the individual in interaction. As such, the interlocutors and the host of variables they bring to the interactional event become part of the construct of L2 interaction and have implications for the validity considerations supporting the assessment. (p.3)

In accordance with this view, Galaczi and Taylor (2018) define interactional competence as "the ability to co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the speech situation and event" (p. 8). Furthermore, the authors emphasise that interactional ability "is supported by the linguistic and other resources that speakers and listeners leverage at a microlevel of the interaction, namely, aspects of *topic management*, *turn management*, *interactive listening*, *breakdown repair* and *non-verbal or visual behaviours*" (p. 8). This socio-cognitive definition of interactional competence was taken as a basis for the present thesis for understanding how the construct is interpreted by raters and represented in assessment scales.

National testing of foreign languages in Europe

The present thesis is concerned with one form of assessment of student competences, namely national testing[2], and is set within a European context, more specifically in the Swedish educational system. National testing is a relatively new form of assessment which has gained in importance and expanded in Europe since the 1990s (European Commission/EACEA/Eurydice, 2009). This increase also applies to national tests of foreign languages. While national tests in languages have been embedded in national education systems for a long time in some European countries, such as Sweden, most of the current national test systems have been developed relatively recently, many since the 2000s (European Commission/EACEA/Eurydice, 2015). The upsurge of new national assessment systems took place in the wider context of a trend at system level towards decentralisation across Europe. Whereas this process was characterised by increased democratic participation and autonomy for schools, the system also demanded new evidence-based accountability measures for the evaluation of educational outcomes, which was realised in the form of national tests.

In the report "Languages in Secondary Education – An Overview of National Tests in Europe 2014/15" by the European Commission (European Commission/EACEA/Eurydice, 2015), national tests are defined as "standardised tests/examinations set by central/top level public authorities and carried out under their responsibility" (p. 5). Examinees should take the tests under reasonably similar conditions and national tests are to be scored in a consistent way. As pointed out by the authors, national language tests in Europe serve various purposes. However, they can be classified according to their main objective into either a 'high-stakes' category or a 'low-stakes' category. High-stakes tests typically summarise an individual pupil's achievement at the end of a school year or educational level and the results are used to make formal decisions about student' progression and future education. This is the most common type in the European school context. The other category, 'low-stakes' national tests, are used to monitor and evaluate the performance of individual schools and students and/or the education system as a whole, in order to

[2] The terms *assessment* and *testing* are used in accordance with H. D. Brown and Abeywickrama (2010). Assessment is defined as "an ongoing process that encompasses a wide range of methodological techniques" (p. 3). In comparison, a test is a "subset of assessment, a genre of assessment techniques" (p. 3). It is essentially a *method*, or an instrument, through which a test-taker's ability, knowledge, or performance in a given domain is measured and evaluated.

provide information that can help improve teaching and learning, hence they have a more of a formative function. Low-stakes national tests are more common in lower secondary education.

It should be kept in mind, however, that national tests are often intended to accomplish several purposes across the two main categories. This is the case in the Swedish school context, where the national tests are distinctly high-stakes; their main function being to support and advise teachers in their decision-making regarding students' final grades which are also used as a basis for selection to higher education. The main objective of the national assessment system in Sweden is thus to enhance comparability and equity within the school system, as well as stability over time. Traditionally, however, the system has served multiple aims. In addition to providing support for teachers' grading, the tests have also had an implicit function to clarify and communicate subject syllabuses and criteria to teachers, thus potentially having an active, positive impact on teaching and learning. It is also emphasised that national test results can be used for local and national analyses of educational achievement (The Swedish national assessment system will be further described in Chapter 2).

One of the main objectives of foreign language teaching is to develop students' competence in the four main communication skills of reading, listening, writing and speaking. However, the extent to which the four skills are tested in national tests in languages in Europe varies. The results from the above-mentioned European report indicate that reading is the most commonly tested skill, writing and listening are tested to roughly the same extent, while speaking is the least tested skill. (European Commission/EACEA/Eurydice, 2015). In the Swedish context, all four skills are tested and the national assessment materials of foreign languages typically comprise three subtests: a speaking test, a writing test, and a section focusing on reception, i.e. listening and reading comprehension. The present thesis is concerned with one of the subtests, namely the speaking component in the national test of English as a foreign language (EFL) at the upper secondary level. It should be noted that English is the first foreign language in the Swedish school system, and it is a compulsory subject from primary school throughout secondary school.

The fact that speaking is the least tested skill in the European context was rationalised in the following way by the authors of the report: "It is probable that the complexity of testing speaking skills as well as the high costs involved, mean that this skill is either simply not tested, or that the speaking tests are designed at school level instead of centrally" ("Highlights Report: Languages in

Secondary Education," 2015, p. 2). In light of this, the national EFL speaking tests in the Swedish context are especially interesting to investigate from a validation point of view, as they are centrally developed and standardised, but internally marked by teachers at the schools where they are administered. It is generally more common that high-stakes national language tests, as well as low-stakes national tests intended to monitor the education system a whole, are externally marked by teachers or other staff outside the school in question. In contrast, low-stakes national tests used to inform improvements in teaching and learning are more often internally marked (European Commission/EACEA/Eurydice, 2015). The case in Sweden with high-stakes national tests that are internally marked is thus quite unique when considered in a European context. However, the system with teacher markings of national tests is highly debated, both in Sweden and internationally, an aspect that will be explored further in Chapter 2.

Since the establishment of the Common European Framework of Reference for Languages (CEFR) by the Council of Europe in 2001, the document has had a great influence in the development of national language tests in Europe (The CEFR will be further described in Chapter 2). In the majority of European countries, the national language tests are linked to the six common reference levels of language proficiency described in the CEFR (European Commission/EACEA/Eurydice, 2015): A1 and A2 (basic user), B1 and B2 (independent user), C1 and C2 (proficient user). In lower secondary education, A2 and B1 are generally the highest levels tested and at upper secondary level, national tests are generally not set above B2. As regards the national speaking tests investigated in the present thesis, they are conducted at the upper secondary level and are intended to correspond to an entrance level or minimal pass level of a high B1 for the first course (called English 5) and a low B2 for the second course (called English 6) (Swedish National Agency for Education, 2018b).

Another, related aspect of the national EFL speaking tests in the Swedish context, which adds to their interest in terms of research, is the test format. The speaking task consists of a paired or group conversation (two or three students discuss a topic among themselves), with both productive and interactive elements (Council of Europe, 2001). This test format is known as a paired or group speaking test. In a report on the comparability of language testing in Europe, published by the European Commission (2015), 133 national language tests at the lower and upper secondary education levels from 28 EU Member

States were studied. With regard to the speaking tests, a division into three patterns of interaction were made and these were found to vary for the different levels in the CEFR: *interaction with other test-taker* (typically a discussion between test-takers in pairs or groups), *interaction with examiner* (often in the form of an interview) and *monologue* (usually in the form of an oral presentation).

At A2, there was an equal balance between interaction with examiner and interaction with another test-taker. At B1, there was considerably less peer interaction. However, monologue was introduced and the majority of tests included interaction with an examiner, suggesting a stronger emphasis on evaluation of the individual learner's oral proficiency at this level. At B2, there was an equal amount of monologue and examiner interaction, once again stressing a more formalised and possibly rehearsed performance in the case of monologue. Peer interaction was less common. At the highest level, C1, there was an exclusive use of monologue and examiner interaction.

It can thus be seen that paired speaking assessment, which is used in the Swedish school context, is less common among national tests at the B1 level and upwards in a European perspective. It is widely recognised that different test formats assess different aspects of language and there is a solid body of research suggesting that the choice of task, and its corresponding test format, has an impact on test taker performance (see, e.g., Brooks, 2009; ffrench, 2003; Galaczi, 2008; Kormos, 1999; O'Sullivan, Weir, & Saville, 2002). This does not imply, however, that one test format is superior to another; they all have advantages and disadvantages. Testing in pairs or groups can be advantageous in many ways, most notably because the test format has the potential of accessing a fuller range of language functions, especially interactional functions, which are typically suppressed or simply not elicited in more traditional formats, such as the oral proficiency interview with examiner interaction (O'Sullivan et al., 2002). However, there are concerns in terms of assessment, which may discourage from using the format in high-stakes testing contexts. One concern is the effect test-takers may have on each other when interacting, so-called 'interlocutor effects' (O'Sullivan, 2002), and the unpredictability and variability that this brings about. Another issue involves the co-construction of interaction, which makes test-takers' performances interdependent and potentially difficult to separate (May, 2011b). These potential threats to the validity of the paired format will be further developed in Chapter 3.

Research questions and aims

Given the background outlined above (further developed in Chapters 2, 3 and 4), the overarching aim of this thesis is to explore different aspects of validity evidence in relation to a paired speaking assessment, as administered in the context of a high-stakes national test at the upper secondary level of the Swedish educational system. More specifically, three areas were investigated: (1) the scoring process, (2) the construct underlying the test format, and (3) the setting and test administration. The thesis adds to the body of previous research carried out in the context of paired and group oral assessment by investigating both social, contextual parameters and cognitive processes activated by the test task, thus aligning with a socio-cognitive approach to test validation (O'Sullivan & Weir, 2011; Weir, 2005). Accordingly, the aim of the thesis is to contribute knowledge to the validation of paired and group oral assessments in the context of foreign language testing. The following research questions are addressed:

- What degrees of rater variability and consistency of rater behaviour can be observed?
- What features of test-takers' performances are salient to raters?
- How are the national EFL speaking tests administered and scored at the local school level?
- What are teachers' views regarding practicality?

Three empirical studies were conducted with the aim of collecting validity evidence from different perspectives; the common denominator being the point of view of the raters. The three studies are:

Study I Borger, Linda (2014)
Looking Beyond Scores: A Study of Rater Orientations and Ratings of Speaking

Study II Borger, Linda (2018)
Assessing Interactional Skills in a Paired Speaking Test: Raters' Interpretation of the Construct

Study III Borger, Linda (2018)
Evaluating a High-Stakes Speaking Test: Teachers' Practices and Views

Study I used a mixed-methods design to investigate inter-rater agreement and raters' decision-making processes. Thirty-one raters participated in the study and rated six paired performances. In addition to analyses of scores, a qualitative content analysis of raters' written verbal reports was made in order to identify features of the paired performances that contributed to raters' judgement. Study II used the written verbal reports from Study I to investigate raters' perceptions of co-constructed discourse; a qualitative content analysis focusing on raters' interpretation of the construct of interactional competence was conducted. Study III investigated how the national EFL national speaking tests are implemented at the local school level by surveying 267 upper secondary teachers regarding their administration and scoring practices, as well as their views on practicality. The third study thus highlights both contextual and consequential aspects of test use. In each of the three studies, more specific questions are addressed for the purpose of gaining more detailed knowledge contributing to the understanding of the main issues explored in the thesis.

Study I was reported in a licentiate thesis, Studies II and III in research articles; hence, the formats of presentation of the studies differ in scope and size, the licentiate thesis being more comprehensive than the research articles.

Outline of thesis

The thesis consists of an overarching discussion and the three empirical studies. The purpose of the overarching discussion is to account for the contextual background and theoretical framework of the thesis, and to discuss the results of the three empirical studies (I-III) in relation to the main research questions.

In the overarching discussion, the first chapter, 'Contextual background', introduces the Swedish educational system, focusing on two areas: the major reform changes of the last few decades and the great trust placed in teacher assessments and teacher professionalism. Further, the national assessment system is outlined, paying particular attention to the national assessment of English and foreign languages. Also, the national syllabuses for foreign languages and their relation to the CEFR are highlighted. Thereafter, the chapter 'Paired and group speaking assessment' reviews previous research on the paired and group speaking test format. The final part of the background, 'Theoretical Framework' is devoted to validity theory and frameworks of language test validation. A methodology chapter follows the theoretical part, where the methods and material used in the different studies are presented. Next, the 'Results' chapter summarises the results of the thesis, followed by the chapter 'Discussion', in which the validity evidence collected in the three empirical studies are discussed in relation to relevant aspects of validity, following the socio-cognitive framework for test validation (Weir, 2005). Lastly, the chapter 'Conclusions' offers some concluding as well as forward-looking reflections, including implications of the findings for the national assessment system and suggestions for future research into areas and issues treated in the thesis. After this, a Swedish summary is offered and finally the three empirical studies (I-III) are included in full, i.e. the licentiate thesis and the two research articles.

Chapter Two: Contextual background

In the following section, a contextual background to the thesis is given. The Swedish educational system, including the system of national assessment, will first be outlined, focusing on two main areas: the major reform changes of the last few decades and the great trust placed in teacher assessments and teacher professionalism. Then, the Common European Framework of Reference for Languages (Council of Europe, 2001) is briefly introduced, since this document has had considerable influence on the national syllabuses for foreign languages. After this, the national assessment of English is outlined from the late 1960s to the present. Finally, the general principles for test development are briefly described.

The Swedish educational system

To facilitate the understanding of the national assessment system in Sweden, it is first necessary to outline the development of the Swedish educational system from the late 1990s and onwards.

Decentralisation and recentralisation

From being one the most centralised and uniform education systems in Europe (OECD, 1998), a major administrative reform in the early 1990s involved a decentralisation process in which decision-making power and financial responsibility was transferred from the state to the municipalities (Gustafsson, 2013). Parallel to this, two other reforms in the education sector took place, adding to the complexity of local school systems. The first was the introduction of free school option, enabling students to choose and attend schools (public or private) based on preference rather than residential area. The second was the decision that not only municipalities would be allowed to run schools but also independent school providers, i.e. private schools. Independent schools in Sweden are publicly funded but have a high degree of autonomy. Since the introduction of this system, the number of independent schools has successively increased. Today, about 15% of students in compulsory school and

26% of students in upper secondary schools attend independent schools (Holmström, 2018).

In line with the decentralisation of the school system, new, deregulated curricula and syllabi were implemented in 1994 (Swedish National Agency for Education), defining overall learning goals for students but leaving a high degree of autonomy for schools and teachers in deciding on teaching content, methods and materials. In addition, the previous norm-referenced grading system was replaced by a goal- and criterion-referenced grading system, requiring local interpretation and implementation (Tholin, 2006). The criterion-referenced grading system was intended to be used for purposes of monitoring the quality and equality of the school system. While the responsibility for implementing education was decentralised to the municipalities and independent school providers, the central government still kept the overall responsibility for schooling and for establishing national standards and goals, including the development of national tests (Nusche, Halász, Looney, Santiago, & Shewbridge, 2011). This is still the case in the present-day system.

It was believed that more market forces in education would increase efficiency and improve quality, as well as lead to reduced costs. However, the impact of the school decentralisation reforms on student performances and on equity in the school system has been greatly debated in both Sweden and internationally (Nusche et al., 2011). During the 2000s, therefore, a recentralisation of parts of the Swedish educational system was carried out (Rönnberg, 2011), and new means of government control and accountability measures were introduced. This included, for example, the establishment of the national Swedish Schools Inspectorate (henceforth SSI), with the aim of regularly inspecting Swedish schools (Swedish Ministry of Education and Research, 2007a), and the introduction of a new curriculum and syllabi intended to include more concrete goals and criteria and a clearer description of teaching content (Swedish National Agency for Education, 2011).

Evaluation and national assessment system

The Swedish educational system has a long tradition of trust in teacher assessments and teacher professionalism, which is in stark contrast to some other countries in Europe where assessment is seen as a separate activity from teaching and learning, carried out by external psychometric experts (Nusche et al., 2011). In other words, there is a strong focus on classroom-based, continuous assessment, through which teachers evaluate students' progress and provide regular feedback. Teachers are also mandated to assign final grades, which are used for high-stakes purposes such as admission to higher education and evaluation of schools and municipalities. Grades are introduced relatively late, as compared to many other countries, from school year six in the present system.

The system with teachers' continuous assessment is thus a firmly rooted tradition, which, from the early 1950s, was used in combination with a norm-referenced grading system, used for rank-ordering and selection purposes (Swedish Ministry of Education and Research, 1942). The principle behind the norm-referenced system was the assumption of a normal distribution of grades at the national level, which was stable over the years. Standardised national tests, referred to as *centralised tests* in the upper secondary school, were provided to support the equivalence of grading. The main function of the centralised tests was to determine the average level of achievement of the class, while individual grading was mainly based on continuous classroom assessment (Gustafsson & Erickson, 2013).

As mentioned above, there was a shift to a goal-and criterion-based grading system in the mid 1990s, in line with both the decentralisation of the school system, and *the system of management by objectives* ('New Public Management') (Mons, 2009; Nusche et al., 2011), which was being implemented in the public sector. In the new grading system, teachers assessed whether goals and criteria for different levels of the grading scale had been fulfilled or not. To strengthen the comparability of teacher assigned grades, national tests, developed under the responsibility of the Swedish National Agency for Education (henceforth NAE), were provided for some subjects. The subjects have varied, but the common core is Swedish (and Swedish as a second language), English and Mathematics. The national tests were assigned an advisory function and were intended to supplement teachers' continuous assessment. However, it was not regulated to what extent national test results should influence the grading of individual students, or the distribution of grades in an individual class or school. This uncertainty concerning the proportional weight of the national test results in relation to students' final grades has been criticised (Swedish Ministry of Education and Research, 2016), leading to an amendment in the Education Act as from January 2018. This will be further described below.

Another change following the criterion- and norm-referenced system, was the shift to more performance-based tasks in the national tests, requiring complex, qualitative evaluations of oral and written production and interaction. Furthermore, the national tests were assigned multiple aims, in addition to the main purpose of supporting teachers' grading, for example enhancing student learning and implementing the curriculum. Following criticism from different experts concerning the difficulty of catering for a range of different aims in one single national test, also expressed in a government inquiry (Swedish Ministry of Education and Research, 2007b), the aims have been reduced to two at present, namely to:

- enhance equity in assessment and grading and to
- provide empirical data for local and national analyses of educational achievement

Another characteristic of the national assessment system, as mentioned previously, is the internal marking carried out at the schools where the tests are administered, often by the students' own teachers. Co-rating, i.e. a process whereby teachers, within the same school or between schools, collaborate in the assessment process, is highly recommended but not mandatory or regulated. To support teachers' assessment, there are extensive guidelines and test specifications. In addition, commented samples of student performances (benchmarks) are provided for the oral and written performance-based tasks.

The system with internal teacher assessment of the national tests has been widely discussed, both nationally and internationally (Nusche et al., 2011; Swedish Schools Inspectorate, 2013). During the start of the new national assessment system, from 1994 to roughly 2005, there was great autonomy at the local school level. The educational authorities did not interfere, fearing that the national tests would be perceived as school-leaving exams rather than advisory assessment materials (Erickson, 2017a). However, an increasing number of studies indicated that the Swedish education system, with its basis in criterion-referenced grading, was afflicted by problems, such as grade inflation (Cliffordson, 2004) and substantial differences between national test results and teacher assigned final grades, both at school level and across schools (Swedish National Agency for Education, 2007). Concerns regarding teacher bias, fairness and equity were raised, leading the Government to mandate the newly initiated SSI to remark samples of national tests and to compare the external markings with teacher markings.

The results from the annual re-markings carried out by the SSI (2010, 2011, 2012, 2013, 2015, 2016, 2017) point to variability of ratings and considerable differences between the original teacher markings and the external markings for the performance-based parts of the national tests, the general trend being that teacher ratings are more lenient than the external ratings. The SSI (2016, 2017) has also observed that deviations between internal and external markings are smaller when another teacher than the student's own teacher marks the tests. It should be noted that only the written parts of the national tests have been included in the re-markings. Hence, no documentation has been made with regard to the speaking components of the national tests. Furthermore, it should be kept in mind that there are inter-rater studies of the national tests which to some extent contradict the results of the SSI re-markings (Erickson, 2009), as well as raise methodological concerns (Gustafsson & Erickson, 2013).

Two external evaluations of the Swedish education system from the OECD also bear relevance (Nusche et al., 2011; OECD, 2015). In their investigation, Nusche et al. (2011) observed both positive and negative features of the Swedish educational system. The authors concluded that the high trust put in teachers' assessment is positive as it fosters professionalism; however, "[a]s can be expected from a such as decentralised approach, there are large variations in the ways evaluation and assessment are undertaken across the country", leading to "variability in quality assurance practices" (p. 8). Concerns were also raised regarding internal marking of the national tests by teachers, as well as the fact that the national tests include performance-based tasks, which are difficult to assess reliably. Recommendations regarding external moderation and/or rating, as well as professional development for teachers were thus made:

> High quality training and professional development for effective assessment are essential to strengthen teachers' practices. External moderation can further help increase consistency and comparability of national test results. Options for doing this include having a second grader in addition to the students' own teachers, employing professionals for systematic external grading and/or moderation, or introducing a checking procedure by a competent authority or examination board. (p. 7)

The OECD report from 2015 drew similar conclusions regarding the lack of reliability of the national assessment data and "the variable assessment capacity of Swedish teachers" (p. 30). In addition, the report highlights additional aspects. For example, Sweden's performance on international assessments was compared with students' average merit rating in school year nine from 1998-

2012. Whereas the average merit rating has steadily increased during this time, Sweden's performance in international assessments has markedly dropped. In addition, the report draws attention to the fact that grade inflation may be explained by schools' competition for students, a result of the free school choice and independent school reforms in the 1990s. The authors draw the following conclusion:

> Differences in interpretation of assessment criteria, issues of teachers' assessment skills and pressures associated with the high-stakes nature of the results for schools have been identified as partial explanations for a mismatch between higher levels reported internally and evidence of declining performance on international surveys. (p. 156-157)

The current system and on-going activities

In 2011, new curricula and subject syllabuses, including more concrete criteria and a more detailed description of teaching content, were introduced as part of the move towards a somewhat more centralised educational system. The criterion-referenced grading system remained but a new, six-point grading scale (A-F), intended to allow for clearer differentiation among students' performances, replaced the four-point scale from 1994. While the content standards remained more or less unchanged, there were more profound changes in the performance standards, referred to as *knowledge requirements*. The performance standards consist of generic value descriptions (used across subjects) demonstrating progression in relation to the levels in the grading scale. There were strong doubts already from the beginning concerning the degree of support that the knowledge requirements would be able to provide for an equal and fair grading (Gustafsson, Cliffordson, & Erickson, 2014). In a government-initiated study by the National Agency for Education (NAE) (2016b), these concerns were confirmed. Results indicate that more than half of all teachers find the national standards to be unclear and significantly fewer teachers believe they have a clarifying function as compared to before the reform. Furthermore, the non-compensatory rule of the grading system, requiring all aspects of the performance standards to have been reached for a student to be awarded a particular grade, was criticised for affecting fairness in a negative way. Based on the results of the investigation, some changes have been made, including for example, a more liberal use of the compensatory rule. A common framework for all national test has also been developed (Swedish National Agency for Education, 2017b).

In addition, a major, politically initiated inquiry of the national assessment system at large was undertaken, and the results were reported during spring 2016 (Swedish Ministry of Education and Research, 2016). Based on the inquiry, some changes and amendments have been politically proposed and/or decided in order to enhance fairness and equity and increase the validity and reliability of the national assessment materials (Swedish Ministry of Education and Research, 2017b). To start with, it was stated that the aims of the national assessment system had to be clarified, and preferably reduced to only one primary aim, namely to enhance equity and fairness in assessment and grading. This change has, at the time of writing, been carried out. Secondly, the most profound change was the decision to digitalise the assessment system, which is to be completed by 2022. As a first step, the written parts of the national tests, for example the essay in English, should be taken on computer. With regard to the digitalisation of the speaking subtests, no specific information has yet been provided.

Thirdly, the proportional weight of the national test results in relation to teachers' grading has been clarified somewhat in the Education Act. As from 1 January 2018, it is stated that teachers shall 'pay special attention' to the results, however not quantified. National tests still have an advisory function and the test results are to be combined with teachers' continuous observations and assessments. Fourthly, the government has proposed *external rating* of national tests, carried out by a teacher other than the student's own, and *co-rating*, whereby two teachers, one of whom holds the main responsibility, independently mark the test (Swedish Ministry of Education and Research, 2017a). In connection with this, student responses should be anonymised. External rating and co-rating are presently being tried out in a pilot project coordinated by the NAE. In addition to this, it was also decided that the number of mandatory national tests in upper secondary school should be reduced. Taking effect 1 January 2018, only tests in final courses for the different study programs are mandatory and the preceding tests are optional to use.

Common European Framework of Reference for Languages

Since the *Common European Framework of Reference for Languages: Learning, teaching and assessment* (Council of Europe, 2001) has been a major influence in the development of the national syllabuses for foreign languages in Europe, and also for the national assessment of foreign languages in Sweden, the framework will briefly be introduced in this section, before a more detailed account of the national assessment of English in the Swedish school context is provided.

In connection with the shift some fifty years ago from a more structuralist view of language to a functional and interactional/socio-linguistic one, the Council of Europe initiated its work on a common language policy to promote and facilitate co-operation among educational institutions, by providing a metalanguage to describe language proficiency, and to establish international standards for the assessment and certification of language proficiency in different countries. The CEFR was developed as a continuation of the Council of Europe's work in language education during the 1970s and 1980s (see, e.g., van Ek, 1975; Wilkins, 1976), and builds on over twenty years of research. It was published in 2001, and was recently accompanied by a Companion Volume (Council of Europe, 2018), further developing certain aspects of the framework. In the introduction of the CEFR it is stated that the document is intended to provide "a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe" (p. 1).

In addition to being used as a reference instrument by almost all member states of the European Union, the CEFR has also had, and still has, a considerable influence beyond Europe. It is important to emphasize that the CEFR is intended to be "a tool to facilitate educational reform projects, not a standardisation tool" (Council of Europe, 2018, p. 26). Consequently, "there is no body monitoring or even coordinating its use" (p. 26). Also important to stress is the subtitle: *Learning, teaching and assessment.* Although the CEFR is mostly recognized for its use in testing contexts, the framework offers a great deal of information on language in general, both theoretical and practical issues, not least its language education policy, focusing on *plurilingualism* and *pluriculturalism* (Council of Europe, 2001, p. 4–6; 133; 168).

The CEFR comprises a descriptive scheme of language proficiency involving language learners' **general competence** (e.g. knowledge of the world, socio-cultural and intercultural knowledge and professional experience, if any; CEFR Section 5.1) as well as their **communicative language competence** (linguistic, pragmatic, and socio-linguistic; CEFR Section 5.2) and **strategies** (both general and communicative language strategies). Furthermore, the framework distinguishes four categories of **communicative language activities** (reception, production, interaction and mediation), four **domains of language use** (the educational, occupational, public and personal), and three types of parameters that shape language use (situational context, text type or theme, and task-related conditions and constraints) (Council of Europe, 2001; Little, 2007). This overall approach is summarised in Chapter 2 of the CEFR (p. 9).

The CEFR is based on an *action-oriented approach*, according to which language users are viewed as 'social agents'. Language is consequently seen as a tool for communication rather than as a subject to study *per se*: "The methodological message of the CEFR is that language learning should be directed towards enabling learners to act in real-life situations, expressing themselves and accomplishing tasks of different natures" (Council of Europe, 2018, p. 27). In line with this, illustrative descriptor scales of language proficiency for different communicative language activities are provided in the framework. The illustrative scales are summarised in a global scale, which describes foreign language proficiency at six levels: A1 and A2, B1 and B2, C1 and C2. It also defines three 'plus' levels (A2+, B1+, B2+). Level A is defined as 'basic user', level B 'independent user' and level C 'proficient user'. In addition to the global scale, there is a self-assessment scale "intended to help learners to profile their main language skills, and decide at which level they might look at a checklist of more detailed descriptors in order to self-assess their level of proficiency" (p. 25). The self-assessment grid is further used in the European Language Portfolio, developed for pedagogical purposes (Little, 2009).

While the CEFR has had and still has a significantly positive impact in both testing and teaching contexts in Europe and beyond, the framework has also met with substantial criticism, concerning e.g. theoretical underpinning, methodology, and issues related to normativity and culture. In this, the use of the document in a wide sense is very often the focal point of concern (see, e.g., Erickson & Pakula, 2017; Fulcher, 2004; Hulstijn, 2007; McNamara, 2010; O'Sullivan & Weir, 2011).

National assessment of English

In this section, the national testing of English, and to some extent other foreign languages, is briefly outlined, focusing on the development of the speaking component. Since the design of the national tests is closely linked to curricula and syllabus reforms, this relation is also highlighted.

1969-1994

In the 1940s and 1950s, Sweden had a system of school-leaving examinations, which included both written tests and an oral exam. These exams disappeared in 1968 and were replaced by so called 'standard tests' in lower secondary school and 'centralised tests' in the upper secondary school. These tests were related to the then existing norm-referenced grading system and were developed by the National Board of Education (Marklund, 1987). In line with the dominating test theories of the period (see, e.g, Lado, 1961), the centralised tests included predominantly closed-ended items of the multiple-choice type, giving high priority to aspects of reliability. However, following the shift from the 'psychometric-structuralist era' to the 'psycholinguistic-sociolinguistic era' (Spolsky, 1976), the centralised language tests were successively revised and more open-ended items were included (Erickson, 1999).

In 1972 and in 1980, new foreign langue syllabuses for upper secondary school and compulsory school respectively were implemented. The revised language syllabuses from this time clearly expressed a functional and communicative view of language in which oral and written communication were given more emphasis than before. Two influential products from this time were Wilkin's (1976) *functional-notional* approach to syllabus design and *The Threshold Level* by van Ek (1975). Wilkins proposed that communicative needs should be taken as a starting point for syllabus design, instead of grammatical structures, which was traditionally the case. Grammatical structures were still important but could be seen as tools to realise these meanings. Furthermore, in *The Threshold Level*, the lowest level of foreign-language ability was specified by describing what a learner should be able to *do* when using the language to communicate in a foreign environment. This work was later continued in the development of the CEFR.

In the early 1980s, the national test development for foreign languages was commissioned to the University of Gothenburg, where it is still located today. During this period, in the beginning of the 1980s, a ten-year project to investigate and develop more integrative, authentic and direct methods of testing oral and written communication in the national tests, in line with the communicative movement which was gaining in popularity ay this time, was initiated (See Lindblad, 1992, for a detailed account). Since the national syllabuses for foreign languages, following the reforms of 1972 and 1980s, increasingly emphasized interaction and communicative competence, the need for a national test of speaking was strengthened. Lindblad (1992) also referred to 'backwash effects' and 'sign-posting functions' as important reasons why an oral component should be included in the Swedish national tests:

> […] the best way for a teacher to indicate that a certain part of a subject is important is probably to test it. Conversely, by not testing it the teacher sends a signal that it is less important. […]
>
> The reasons for establishing national models for the systematic testing of oral performance can thus be summarized in the well-known concept of "backwash effect". These influence students and teachers alike. Such tests also serve the purpose of defining what the term "oral proficiency" as used in the national syllabuses stands for.
>
> (p. 280)

In addition, there were indications that teachers were positive towards an oral subtest as part of the national test battery. In compulsory school, a survey that included questions on the assessment of oral proficiency was conducted with teachers in connection with the national tests in 1990. The results showed that more than 80% of the teachers who responded to the survey believed there was a certain or a great need of a national speaking test in English, although many were concerned about practical issues (Erickson, 1999).

In compulsory school, the first oral national test was administered on a large scale in connection with the national test in 1991. However, it was still optional and teachers could decide whether they wanted to conduct the test with their students or not. About 30% of schools ordered the oral national test. Teachers' reactions were mainly positive, but the concerns about practical issues remained. In connection with the national test administration in 1994, a peer interaction format, involving a conversation between students, was offered for the first time. As part of in-service teacher training, a videotape was provided containing samples of student conversations, in which the teacher had a minimal role. In addition, there were conversations between students, and between students and teachers, about oral language proficiency and assessment. These videotapes were intended to be used at in-service seminars when groups of teachers, e.g. a group of teachers at a particular school, could watch the performances and discuss them. The videotaped material was met with great interest and was ordered by a large number of schools (Erickson, 1999).

After the pilot period with optional oral tests, the speaking component finally became a mandatory part of the national test battery in 1998 for compulsory school and in 2000 for the upper secondary school. Even though both individual and paired/group formats were tried out, the paired or group test format was chosen for the mandatory test. There were several reasons for this. First, the paired and group format reflected the focus of the foreign language syllabuses on interaction, thus having an implementing function, which could lead to positive washback effects (Taylor, 2005). Secondly, as explained above, many teachers expressed concerns about practical issues and the time-consuming nature of test administration in connection with the oral tests. Conducting the oral tests in pairs was therefore seen as a more feasible alternative to conducting individual interviews. Finally, continuous studies of attitudes during the pilot period showed that the acceptance among teachers for using paired models was satisfactory, and successively increasing.

1994-present day

New national tests were implemented in connection with the introduction of the goal-related grading system in 1994, and the revised curriculum and syllabuses. The influence of the functional and communicative view of language was further strengthened in the foreign language syllabuses. The national tests of English from this time included tasks aimed at testing *receptive* competence and oral as well as written *production* and *interaction*. As can be seen, the terminology used in the CEFR had been adopted, instead of the 'four skills' used earlier.

Already in 2000, the next revision of the foreign language syllabuses took place, in which the link to the CEFR was made more explicit, for example by the emphasis placed on *interaction* and *intercultural competence* (Erickson & Pakula, 2017). Furthermore, the progression between compulsory and upper secondary school was made more direct in the revised system by subsuming English and the foreign languages in one model consisting of seven levels, referred to as 'steps'. As pointed out in Erickson and Pakula (2017), having six steps, in alignment with the six common reference levels in the CEFR, was discussed. However, this was decided against for various reasons (see Erickson & Pakula, 2017).

In 2011, the latest revision of the curriculum and syllabuses was made. A new six-point grading scale replaced the previous four-point scale. This reform further strengthened the relationship between the Swedish syllabuses for foreign languages and the CEFR, by making an explicit link between the entrance level or pass level (the grade E) of the seven steps in foreign languages and the common reference levels in the CEFR (see model in Swedish National Agency for Education, 2018b).

Development of national tests of English

The construct

As mentioned above, the Swedish national syllabuses for foreign languages are to a considerable extent similar in approach, and tentatively related to the reference levels of the CEFR. Communicative language activities focused upon in the national tests are reception (listening and reading), and oral and written production and interaction. Furthermore, strategic competence and adaptation to purpose, recipient and situation are explicitly defined as learning outcomes. Subsystems like vocabulary, grammar and pronunciation are viewed as important fundamentals but not as goals *per se*. It should be noted that different aspects of language proficiency are integrated in the subtests. For example, there may be a prompt for the writing and speaking assignment, in the form of a text to read. In the speaking test, both oral production and interaction are tested, which means students both need to speak English and understand what their partner is saying. Furthermore, aspects of intercultural competence are incorporated in the tests, mainly reflected in the choice of texts and topics for the oral and written parts (Erickson, 2017b)

A typical national test consists of three subtests: a speaking test, in which pairs, or groups of tree students, participate in a discussion about a given theme; a test focusing on the receptive skills listening and reading, with a variety of texts and tasks combined into a single score; and a writing test, in which students are sometimes offered a choice between two different subjects.

Test construction and guiding principles

As mentioned in Erickson and Åberg-Bengtsson (2012), in order to cope with the complex task of developing tests taken by a national cohort of students (*N* ≈ 120,000), marked internally by teachers, fundamental principles and guidelines, common to all materials, have been established. These are publicly available on the national assessment project website[3], together with sample tests

[3] https://nafs.gu.se/english/information

and scoring guides, and include, among others, the following aspects, as outlined in Erickson and Åberg-Bengtsson (2012):

> To make what is most essential assessable – not making what is easily measurable the most important;
>
> To give students the chance to show what they actually *know* and *can do*, not primarily trying to detect/focus on what they do not know/cannot do, e.g. by providing broad, multi-faceted, varied, monolingual tests, with – to as large an extent as possible – progression of difficulty, within and between tasks;
>
> To enhance validity and reliability and avoid bias, for example by developing theoretically and empirically well founded tests in collaboration with a wide group of stakeholders, by pre-testing all materials in large, randomly selected groups across the country, and by following closely the sue of the test;
>
> To present individual results in profiles;
>
> To comment on strengths before weaknesses; when analysing weaknesses, distinguish between errors that [might] *disturb* and errors that *destroy* communication, i.e. between errors representing different degrees of gravity (Erickson, 2006)
>
> (p. 3)

As mentioned in one of the points above, the national tests of English are developed in a distinctly collaborative process with different groups of stakeholders, e.g. students, practicing teachers, teacher educators and researchers. All tasks are piloted and pre-tested in large, randomly selected groups of students across the country, and in this process, participating teachers and students are asked to comment on different aspects of the materials, thus contributing to the development of the tests. In addition, standard setting procedures follow established routines (e.g., Angoff, 1971) and are carried out with experienced teachers and teacher educators (For further information see Erickson & Åberg-Bengtsson, 2012). It should also be mentioned that the national tests of English are so-called *proficiency tests*, which means they aim to test test-takers' global competence or overall ability, without focussing on any specific course content (H. D. Brown & Abeywickrama, 2010).

Test results and reactions

Results and reactions to national tests are monitored and made publicly available in reports on the websites of the Swedish National Agency for Education and the test development project. In general, students at all levels perform well on the national tests of English in relation to the national standards. Around five percent of students do not reach the pass level at the end of compulsory school and about 20% are awarded the highest grade (Erickson & Åberg-Bengtsson, 2012).

In terms of teachers' reactions to the tests of English, they are generally very positive, both to the principle of national testing as such and to the different assessment materials. The test developers at the University of Gothenburg conduct annual questionnaires with teachers who administer and mark the tests. During the past 15 years, the large majority of teachers who answer the questionnaires have expressed positive opinions, "often concerning the breadth and variation of the tasks, the close connection between the materials and the syllabuses, the profiled presentations of the results, and the support for scoring and grading provided in the guidelines" (Erickson & Åberg-Bengtsson, 2012, p. 9). Regarding the speaking component, teachers are positive towards the paired test format in terms of students' opportunities to display their oral abilities and its close alignment with the foreign language syllabuses. The criticism given concerns mainly workload, as well as aspects of feasibility in connection with the administration of the speaking tests. Teachers' are also asked to report on their students' reactions to the tests. In general, around five percent are considered negative and the rest either neutral or positive.

Chapter Three: Paired and group speaking assessment

In the following section, previous research on paired and group speaking tests, is briefly outlined, focusing on strengths and weaknesses of the test format.

The paired and group speaking format is considered to have many advantages (see, e.g., Ducasse & Brown, 2009; Van Moere, 2013). For example, as regards administration, testing speaking in groups is more practical in terms of both time and cost, compared to an individual test format. Furthermore, the format is less cognitively burdensome to examiners, as they can focus on the rating process instead of having an active role in the conversation as an interviewer. It has also been indicated that test-takers are more positive towards a paired or group speaking test format and view peer interaction as less intimidating than interaction with an examiner (Egyud & Glover, 2001; Fulcher, 1996; L. He & Dai, 2006; Ockey, 2001; Van Moere, 2006). In addition, there is the potential of a positive washback effect, as the test format may encourage communicative and interactional speaking tasks in the language classroom. Although not a sufficient argument on its own, a close link between testing and teaching is positive in terms of construct representativeness.

Also, one of the main arguments made in favour of the paired and group speaking test format is the potential of eliciting a wider range of language functions than is generally possible in traditional speaking test formats (Johnson, 2001). A series of discourse-based studies have been undertaken to examine discourse functions and use of language functions in paired and group oral tasks, often in contrast with the individual speaking test format (see, e.g., Brooks, 2009; ffrench, 2003; Galaczi, 2008; Kormos, 1999; O'Sullivan et al., 2002). For example, discourse in paired and group tasks has been found to be more 'authentic' and conversation-like, with test-takers having more equal status than in interviews where the examiner is leading the conversation (van Lier, 1989; Young, 1995). Further, it is indicated, both through discourse-based and rater cognition studies, that the paired and group format activates strategic processes that are likely to be used in real life conversations, such as negotiation of meaning, clarifications, confirmations, interactive listening, rephrasings and scaffolding (Brooks, 2009; Ducasse & Brown, 2009; Galaczi, 2014; Gan, Davison, & Hamp-Lyons, 2009; May, 2011b). In addition, paired and group formats have been demonstrated to generate a more comprehensive range of language functions than singleton formats (ffrench, 1999; O'Sullivan et al., 2002).

Nevertheless, the paired and group format has also received significant criticism. To start with, challenges in terms of eliciting a richer and more authentic speech sample with more varied interactional functions have been identified. Although supposedly having equal social status, there is evidence that test-takers do not always work cooperatively and mutually to produce co-constructed discourse in paired/group speaking test. For example, one of the speakers may take on a dominant role in the conversation, test-takers' may produce parallel speech by not engaging with or extending each other's ideas, or one of the speakers may choose not to ratify their partner's topic (Galaczi, 2014; May, 2009; Van Moere, 2007).

He and Dai (2006) examined the frequency of interactional language functions (ILF) present in group oral test performance in the College English test (CET-SET) in China. The authors found that test-takers produced very few of the expected ILF:s and they rarely negotiated meaning, which could be explained by the fact that test-takers framed "the discussion task as an assessment event rather than communicative interaction with other members" (p. 392). For example, students were concerned with expressing their own ideas and focused less on responding to what other test-takers said. In addition, it was evident that candidates considered the teacher examiners and not primarily the fellow test-takers in the group as their target audience. Adding to these results, Luk (2010) and Lam (2015), who both investigated a school-based group speaking test in Hong Kong using conversation analytic methodology, found that test interactions were characterized by "institutionalized and ritualized talk" (Luk, 2010, p. 47) and that test-takers "oriented to the teacher-raters as a 'privileged overhearer'" (Lam, 2015, p. 344).

In response to this, Van Moere (2013) rightfully remarks that "peer test tasks must be properly framed in order to maximize the strategies or functions to be assessed" (p. 2). He further points out that "[p]aired and group talk may be considered as valid not because of similarity to ordinary conversation, but because, if properly set up, it enables language testers to observe a wider variety of cognitive and strategic processes than might be gained from other oral assessment formats alone" (p. 2).

In terms of reliability and fairness, there are concerns about the 'instability' of the paired and group test format as variations in a test-taker's performance may be related to group dynamics and the impact that the test-takers have on each other when co-constructing speech. In interview tests, a similar problem concerns the impact of interviewer variation on scores (Brown, 2003). However, in the interview, a greater degree of standardization can be achieved by using interlocuter frames or scripts (Taylor, 2003), whereas interlocutor variables in group tests are difficult to control. Van Moere (2013) notes that "[l]anguage testers are intuitively wary of assessing students in interactive groups because of the sheer number of uncontrollable variables and unknown effects associated with these variables; i.e. gender, age, status, friendship, shyness, talkativeness, opportunities for taking the floor, willingness of individuals to participate, different proficiency levels in the group" (p. 414).

Consequently, there are a range of studies that have investigated interlocutor characteristics that may impact test-taker performance, for example *gender* (O'Loughlin, 2002; O'Sullivan, 2000); the *varying proficiency levels* of the interlocutors (Csépes, 2009; Davis, 2009; Iwashita, 2001; Nakatsuhara, 2006; Norton, 2005); *personality* (Berry, 1993, 2007; Nakatsuhara, 2009; Ockey, 2009); and *acquaintanceship among interlocutors* (O'Sullivan, 2002). Overall, the results of these studies are mixed, and sometimes inconsistent, indicating that (a) interlocutor effects are highly context-dependent, and that (b) there is no linear relationship between interlocutor characteristics and discourse outcomes and scores. Nevertheless, despite somewhat contradictory findings regarding interlocutor effects and their impact on scores, it is generally agreed that the matching of students in paired/group speaking tests is an issue that needs to be carefully considered.

Another, related challenge of paired/group oral testing relates to how raters assign individual scores based on co-constructed performances, and how such scores are to be interpreted in terms of validity claims (Fulcher & Davidson, 2007; Taylor & Wigglesworth, 2009). May (2009) argued that awarding shared scores for IC may be "one way of acknowledging the inherently co-constructed nature of interaction in a paired speaking test" (p. 397). However, this issue is debated and Nakatsuhara (2013), for example, maintained that joint scores would be unfair in cases of asymmetric interaction where one test-taker tries hard to invite and involve more quiet partners but fails to do so.

Finally, when looking at inter-rater reliability and consistency of scores in the context of paired and group speaking assessments, the number of research studies conducted seem more limited. One example is Van Moere (2006; 2007), who examined a group oral test administered at a Japanese University with 113 participants sitting the exam twice in groups composed of different interlocutors. The results indicate that differing rater severity was not the main cause of the variation in test-takers' scores. Instead, variations in performance from one test occasion to another was attributed to social factors related to interlocutors or group dynamics, which could "be due to observable factors such as personality or talkativeness, or more intangible interpersonal factors in the way group members react to each other which affect content and delivery of speech acts" (p. 435).

Somewhat contradictory results were observed in a study by Bonk and Ockey (2003) who examined the same group oral test as Van Moere (2006; 2007) in two consecutive administrations (more than 1000 examinees), using manyfacet Rasch analysis. The findings showed that there were large rater differences in terms of severity in both administration of the test. However, this characteristic was not stable over time; returning raters moved towards greater severity and consistency, suggesting that rigorous rater training could overcome difficulties of rater inconsistency.

To sum up, the evidence presented in previous research shows both strengths and weaknesses of the paired and group speaking test format. A solution to mitigate the validity threats associated with the variability of the test format has therefore been to include paired or group tasks as one of several speaking components in a test battery, which is done in the Cambridge English speaking tests (Galaczi, 2014). This ensures that oral test scores are not based solely on one decision. In addition, task design and the framing of paired peaking tasks seem to be essential in order to elicit the interactional functions that the test format intends to do. It has also been suggested that rater training could be an important tool to improve rater agreement (Davis, 2016; Graham, Milanowski, & Miller, 2012). Although some research has found that variability can persist even after extensive rater training (Hoyt & Kerns, 1999; Lumley & McNamara, 1995), Graham et al. (2012) maintain that "correctly designed training can improve agreement" (p. 15). The issue of rater training as a way of improving rater agreement in paired speaking tests will be explored further in Chapter 7 Discussion.

Chapter Four: Theoretical framework

Validity has gained in importance since the middle of the 20th century and is now seen as a most central concept in the development and evaluation of language tests. The notion of validity is traditionally associated with the question of whether a test "measures accurately what it is intended to measure" (Hughes, 1989, p. 22). This view presupposes that validity is an attribute or characteristic of the test itself. However, this notion has undergone changes over the past half century. In this chapter, the concept of validity is first introduced before examples of validation frameworks for language testing are exemplified.

The concept of validity

In the early time of validity investigation, three 'types' of validity were predominant; content, criterion and construct (Cronbach & Meehl, 1955). These were viewed as more or less separable. Content validity is concerned with the degree to which the test content is representative of the domain to be tested. Criterion validity involves a comparison "between a particular test and a criterion to which we wish to make predictions" (Fulcher & Davidson, 2007, p. 5). When test results are used to predict achievement on a future criterion, e.g. academic success, it is referred to as *predictive validity*. If test results are used "to predict a criterion at the same time the test is given" (ibid., p. 5), e.g. by comparing a new test to an established one, or by comparing two test groups at the same time, it is referred to as *concurrent validity*. Finally, construct validity, often regarded as encompassing the other two, involves demonstrating that a test is actually measuring the construct it claims to be measuring. In line with this conceptualisation, validation research in the 1950s through the 1970s typically involved correlational and content analyses, as well as factor analytic techniques.

Messick (1989b) changed the way validity was viewed, by arguing that it was a *unitary* concept:

> Traditional ways of cutting and combining evidence of validity, as we have seen, have led to three major categories of evidence: content-related, criterion-related, and construct-related. However, because content- and

> criterion-related evidence contribute to score meaning, they have come to be recognized as aspects of construct validity. In a sense, then, this leaves only one category, namely, construct-related evidence (p. 20)

Furthermore, Messick (1995) described construct validity as a superordinate category, in which other sources of validity evidence, previously regarded as separable, were integrated in a 'comprehensive view of validity':

> Validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual – as well as the potential – consequences of score interpretation and use (i.e., construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility. (p. 742)

In his definition of validity, Messick highlighted the fact that validity is *a matter of degree* as well as *a multi-faceted concept*, requiring different types of evidence to support any claims for the validity of a particular test use, which are "not alternatives but rather supplements to one another". Messick (1989b) consequently defines validity as:

> an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (p. 13)

According to this definition, validity is not a characteristic of the test itself, but is associated with the interpretation and use of test scores.

To illustrate the concept further, Messick (1989b) presented a matrix showing different facets of validity, shown in Table 1 below:

Table 1. Facets of validity as a progressive matrix (adapted from Messick,1989b, p. 20)

	Function of testing	
Source of justification	*Test interpretation*	*Test use*
Evidential basis	Construct validity (CV)	CV + Relevance/utility (R/U)
Consequential basis	CV + Value Implications (VI)	CV + R/U + VI + Social consequences

In the left column, the 'source of justification' is seen, which can take the form of either evidence or consequences of testing. In the first row, the 'function of testing' is found and this includes two components: interpretation and use. The evidential basis for test interpretation is construct validity. Furthermore, as can be seen, the evidential basis for test use is also construct validity, with the addition of the context for which the test is designed or used. The consequential basis of test interpretation is associated with value implications, which in Fulcher and Davidson (2007) are described as: "the theory and philosophy underlying the test, and what labels the test designer gives to the constructs. Labels send out messages about what is important or 'valued' in performance on the test, and this is part of the intended meaning of the score" (p. 13). Finally, the matrix shows that the consequential basis of test use is linked to social consequences of using the test – on an individual and/or society. This is commonly referred to as consequential validity. It is emphasised that the categories of the matrix are not watertight, but rather 'fuzzy', which adds to the complexity of the model.

To sum up, with the unitary definition of validity, the focus of validation shifted from the test itself to test score interpretation and use. Ideally, this could be accomplished through the creation of a validation argument by gathering different sources of validity evidence to support a particular test use. In language testing, such an argument-based approach, building on scholars such as Kane (1992) and his associates, is evident in for example Bachman (2005), Bachman and Palmer (2010) and Chapelle, Enright, and Jamieson (2008).

Messick (1989b) also contributed to an expansion of the concept of validity to include social values and consequences by maintaining that evaluation of social consequences of test use, as well as the value implications of test interpretation, both 'presume' and 'contribute to' construct validity (p. 21). The role of consequential aspects of validity in validation frameworks, including e.g. washback and social responsibility, is a controversial issue, frequently discussed in the language testing literature as well as in educational measurement in general (Davies, 1997; McNamara, 2006; McNamara & Roever, 2006; Mehrens, 1997; Popham, 1997; Shohamy, 2001). There is an on-going debate regarding whether test consequences should actually be evaluated as part of validity, or under other conceptualizations of test quality (Xi & Davis, 2016).

In addition to reconceptualising the notion of validity, Messick (1989b) described two major threats to construct validity: *construct underrepresentation* and *construct irrelevant variance*, which are useful concepts both at a theoretical and an operational level. *Construct underrepresentation* means that "the test is too narrow and fails to include important dimensions or facets of the construct" (p. 34). For example, a test for the purpose of placing students in a writing course, which only measures their vocabulary knowledge, is not a valid indicator of students' writing ability. In comparison, *construct irrelevant variance* means that "the test contains excess reliable variance that is irrelevant to the interpreted construct" (p. 34). An example of this could be variation in test scores that are attributed to interlocutors' personal characteristics in the context of paired/group speaking assessment. Both types exist in all assessments. Consequently, in all test validation, convincing arguments need to be presented in order to refute these threats.

Despite its enormous influence, the operationalisation of Messick's model has proven difficult to achieve, due to the complexity of the model and lack of clarity with regard to practical guidance in the process of test validation (McNamara & Roever, 2006; Shepard, 1993). With this in mind, the presentation now continues with further developments of frameworks of test validation within the field of language assessment.

Frameworks of test validation in language assessment

Newton and Shaw (2014) explain the difference between validity and validation in the following way: "Validity and validation are two sides of the same coin. Validation is an investigation into validity, so validity is the property that is to be investigated; and validation is the process by which it is investigated. *Validity theory* provides a conceptual framework to guide *validation practice*" (p. 2). To explain the concept of validation further, and relating it to the context of language assessment, two more definitions will be given. First, Chapelle and Voss (2013), aligning with an argument-based approach to language test validation, define the concept in the following way:

> Validation is defined as the justification of the interpretations and uses of testing outcomes. In this sense validation appears at first to be a one-sided evaluation, if the aim is solely to produce justifications; but the idea is that, in the process of attempting to justify something, one confronts both sides of an argument. Despite the intended aim of justification, validation is supposed to entail inquiry into the meaning of test scores, their use, and their consequences.

In addition, perhaps a more accessible definition of language test validation is given in "The Guide to submitting validity evidence" produced by ICAO's

Aviation English Language Test Service (ICAO AELTS is a language proficiency test for aviation English).

> Validity is a multifaceted concept, and different types of evidence are required to support claims made regarding the validity of test scores. Both quantitative and qualitative data and research methods can be used in the validation process. All evidence should be methodically collected, analyzed and reported. Some aspects of the validation process occur before the test event (i.e., in the design and development phase) and other aspects of the validation process occur after the test event (i.e., based on data obtained in the trialing and live testing phases). Validation should be considered an ongoing process. For example, testers are often required to return to the design and development phase based on the results of ongoing validation studies conducted after trailing and live testing.

To put it briefly, validation entails "a systematic gathering of empirical evidence that provides insights into the extent to which a test measures what it is supposed to measure, relative to its purpose and use" (Timpe-Laughlin & Choi, 2017, p. 21). A validation study may thus shed light on both strengths and weaknesses of an exam. In the following, four current validation frameworks from the language testing literature will be presented, which all, directly or indirectly, bear relevance for the analysis and interpretation and use of the results in the present thesis.

Test usefulness

A notable example of an attempt to simplify Messick's (1989) work and make it more operationalizable is presented in Bachman and Palmer's (1996) model of test usefulness, which is intended to be used for "quality control throughout the entire test development process". The model is consistent with Messick's perspective of validation which advocates gathering different sources of validity evidence to support score interpretation and test use. The authors use the term 'usefulness' as an overarching concept in place of construct validity, to include five 'test qualities': reliability, construct validity, authenticity, interactiveness, impact, as well as practicality, which fills the function of prioritizing the investigations of the five qualities (See Figure 1 below). Four of the test qualities - *reliability*, *construct validity*, *authenticity* and *interactiveness* – address test score interpretation, whereas the remaining two – *impact* and *practicality* – attend to consequential aspects of test use. The authors argue that it is the overall usefulness of a test that should be maximized, rather than the individual 'test qualities'. In achieving this, the combined effect of the test qualities on the overall usefulness of a test needs to be evaluated. Furthermore, the importance of each test quality is context-dependent and therefore must be determined for each unique testing situation.

Reliability is defined by the authors as “a function of the consistency of scores from one set of tests and tasks to another” (p. 21). *Construct validity* refers to “the extent to which we can interpret a certain test score as an indicator of the ability(ies), or construct(s), we want to measure” (p. 21). *Authenticity* concerns the degree to which the test task characteristics are relevant to the features of tasks in the real world, referred to as target language use domain (TLU), and is thus related to the traditional concept of content validity. *Interactiveness* has to do with the extent to which the test tasks involve the individual test taker’s characteristics (language ability, background knowledge and motivations) in accomplishing a test task. *Impact* refers to the consequences of test use for individuals (e.g. test takers and teachers), educational systems and society at large, including effects on teaching (washback). *Practicality*, meanwhile, pertains to the implementation of tests and is concerned with the relationship between “the resources required in the design, development, and use of the test and the resources that will be available for those activities” (p. 36). Resources are further divided into three types: (a) human resources, (b) material resources, and (c) time.

Usefulness = Reliability + Construct validity +
Authenticity + Interactiveness + Impact + Practicality

Figure 1. Qualities of test usefulness (Bachman & Palmer, 1996, p. 18)

Bachman and Palmer’s model shifts the emphasis from validity to test usefulness, thus providing an alternative view of the concept. Although Xi and Sawaki (2017) argue that “[b]ecause of its value in guiding practical work, this framework quickly came to dominate empirical validation research and became the cornerstone for language test development and evaluation” (p. 195), others, such as Fulcher and Davidson (2007), claim that “it has not been extensively used in the language testing literature” (p. 15). Fulcher and Davidson think this “may be because downgrading construct validity to a component of ‘usefulness’ has not challenged mainstream thinking since Messick” (p. 15). Nevertheless, there are both more recent and earlier examples of research where the model has been applied (e.g., Chapelle, Jamieson, & Hegelheimer, 2003; East, 2015; Spence-Brown, 2001).

Argument-based approaches

Over the last few decades, an argument-based approach to validation has grown in popularity and use. According to this approach, "validation is seen as a process of developing and appraising the strength of an argument concerning the interpretation and uses of test scores" (Newton & Shaw, 2014, p. 3). Argument-based approaches to test validation in educational measurement (Kane, 1992; Kane, Crooks, & Cohen, 1999) have inspired parallel developments in language assessment. For example, Bachman (2005) and Bachman and Palmer (2010) have built on Kane's work (more information provided below) to develop an Assessment Use Argument (AUA), intended to guide both test development and use. Another example is the work of Chapelle et al. (2008), who have adopted Kane's framework for language testing. In the following, the model outlined in Chapelle et. al (2008) will be focused on. First, however, a brief introduction to the argument-approach is given.

Kane and his associates have used practical argumentation theories in their argument-approach to test validation (Toulmin, 1958). According to them, validation is seen as a process consisting of two stages: The first stage is a specification of an *interpretive argument*[4], which is simply "an overall structure including essential inferences, assumptions and warrants, but excluding much of the backing from empirical evidence and logical analysis that would be required in order to judge its strength" (Newton & Shaw, 2014, p. 140). Once the overall structure of the interpretative argument is complete, the evaluator can move on to the second stage, which entails constructing a *validity argument*, in which theoretical and empirical evidence from validation studies are used to evaluate the strength of the overall argument. When the validity argument is deemed to be adequately strong, validation stops.

Chapelle et al. (2008) used an argument-based approach in their validation of the Test of English as a Foreign Language (TOEFL). The examined language domain is academic English use. The validity argument was presented by first articulating an interpretive argument that included the following claims or assumptions (see Figure 2 below):

[4] In 2013, Kane changed the label 'interpretative argument' to 'interpretation and use argument' (IUA) since the earlier formulation had given insignificant weigh to uses.

> (1) that tasks on the test were appropriate for providing relevant observations of performance from the examinees on relevant tasks; (2) that the evaluation of examinees' performance resulted in accurate and relevant summaries (test score) of the important characteristics of the performance; (3) that the observed scores were sufficiently consistent to generalize to a universe of expected scores; (4) that the consistency of the expected scores can be explained by the construct of academic language proficiency; (5) that the construct of academic language ability predicts a target score indicating performance in the academic context; and (6) that the meaning of the scores is interpretable by test users, who therefore use it appropriately (Chapelle & Voss, 2013, p. 7)

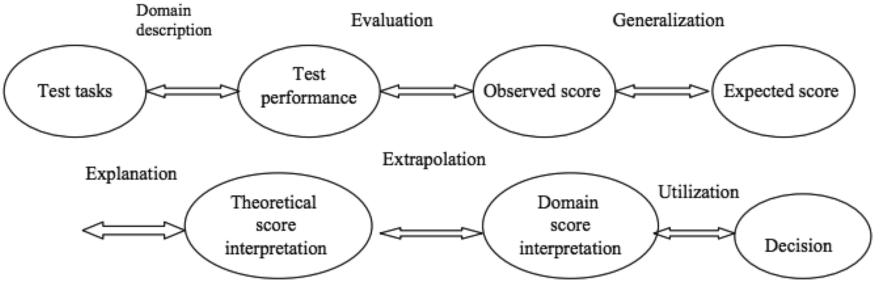

Figure 2. Links in an interpretative argument (Adapted based on Chapelle et al., 2008) (Xi & Sawaki, 2017, p. 197)

The inferences are indicated by the nominalisations with "-tion" suffixes. The validation process was further illustrated by examples of research used to support the six claims. A few examples will be given to illustrate. For the first claim, support was gathered by examining tasks that students typically perform in English-medium universities. The second claim was supported by research that involved studying scoring rubrics. To support the third claim, analyses of generalizability were undertaken and student performances were examined. The fourth claim was backed by the use of several studies, among others a factor analysis that showed that the test data corresponded to the hypothesized component structure.

Chapelle (2008, p. 349) describes the validity argument using a staircase metaphor. Only when an inference is supported by the backing of the appropriate research, can the next step be taken: "In this way the argument can be seen as incremental and additive. A gap in the support for any one of the steps reveals a weak stair, which may preclude a continuation to the final intended conclusion" (Chapelle & Voss, 2013, p. 8). In other fields, the argument-approach has been likened to a bridge (Kane et al., 1999) or a chain (Crooks, Kane, & Cohen, 1996). Furthermore, Chapelle, Enright, and Jamieson (2010), as summarized in Knoch and Chapelle (2017), listed the following four advantages attained from the use of argument-based validity over alternatives:

> First, they found it more productive to state the intended score interpretations and uses by specifying multiple inferences with their supporting warrants rather than relying solely on the construct definition of the abilities to be measured. Second, they found that when they followed the procedures for developing the validity argument, the types of validation research that would be required became apparent. In fact, the assumptions were specific enough to prompt particular research questions, thereby providing links between the validity argument and the validation research. Third, the logic among parts of the validity argument depicts the rationale that connects a test taker's performance on the test to the use of the test scores by showing how each inference builds upon the conclusion from the previous one, and how the research supports each inference. The logic that was built upon connections was therefore preferred over a listing of types of validity evidence because a list of evidence does not show how the validation research supports the score use. Fourth, a clear validity argument that includes specific assumptions underlying inferences presents the opportunity to challenge the validity argument by questioning its logical development or the support for any of its inferences. (p. 2)

There are many examples of argument-based approaches to investigating the validity of test score interpretations, uses and consequences in language testing contexts (Brooks & Swain, 2014; Liying Cheng & Sun, 2015; Enright & Quinlan, 2010; Frost, Elder, & Wigglesworth, 2012; Youn, 2015). However, both within the field of language testing (Davies, 2012) and within educational measurement more broadly (Newton & Shaw, 2014, p. 134-145), concerns have been raised regarding the challenges posed by the method. Moss (2003, 2013), for example, has questioned the utility of the argument-based for classroom-based assessment, as well as for local use of standardised assessments by teachers and other education professionals. Moss and her colleagues thus argue for a shift of focus to local contexts.

Similarly, within the context of language testing, Xi and Sawaki (2017) point out that "the level of complexity and sophistication required for constructing tailored arguments for specific uses may still discourage use among teachers and practitioners despite attempts to make it more accessible (Bachman & Palmer, 2010)" (p. 205). Xi and Davis (2016) also refer to the absence of a "common yardstick against which arguments can be judged", making it difficult "to evaluate the completeness, coherence, and plausibility of each argument" (p. 77). On a related note, Knoch and Chapelle (2017), state that:

> [e]ach validity argument example introduces a slightly different conceptualization for framing the concerns of language testers. Language testers tend to think of analyses investigating construct validity, reliability, authenticity, and rating, for example. A validity argument can include results from these types of analyses, but each analysis needs to be motivated by the role that its results play in the validity argument. (p. 3).

A way forward is outlined in Xi and Davis (2016) where developing use-specific argument structures, or templates, for different test uses (e.g., admissions, licensure, placement) is proposed (see also Chapelle & Voss, 2013).

Construct validity approaches

Chapelle (1998) characterized three approaches to defining a construct, which can be used for framing validation studies. Referring to Messick (1981), Chapelle (1998) argues that 'performance or response consistency' is a central term in relation to construct validity. However, the problematic aspect about construct definition is "to hypothesize the source of performance consistency" (p. 34). In light of this, different theoretical perspectives of construct definition "can be understood by identifying how they explain response consistency (Messick, 1981)" (p. 34). According to Chapelle (1998), there are three main approaches to construct validity: a *trait* perspective, a *behaviour* perspective and an *interactionalist* perspective.

In a *trait definition* of a construct, performance consistency is related to the characteristics of the test-taker, e.g. the person's knowledge and processes (e.g. speech processes in the context of a speaking test). This means that if a person's performance on a test is consistent, it is attributed to the knowledge and skills of the test-taker, or put differently the "correspondence between the score and the actuality of the construct in the test taker" (Fulcher & Davidson, 2007, p. 16).

In contrast, in a *behaviourist definition* of a construct, performance consistency is attributed to contextual factors (e.g. the relationships between test-takers in a group conversation test). That is to say, consistent test performance is assumed to say something about the context, for example the setting, topic and participants. As pointed out in Fulcher and Davidson (2007):

> In 'real world' communication there is always a context – a place where the communication typically takes place, a subject, and people who talk. For example, these could be a restaurant, ordering food and the customer and waiter. According to this view, if we wish to make an inference about a learner's ability to order food, the 'real world' facets should be replicated in the test as closely as possible, or we are not able to infer meaning from the test score to the real world criterion" (p. 16).

This approach is exemplified in e.g. the work of Tarone (1998), in which it is argued that performance on test tasks varies (within individuals) in response to contextual factors. Contextual variables could be different test tasks or facets of test tasks, such as whether the test-taker speaks to an interviewer or a peer/classmate in a speaking test. Tarone (1988) argues that the there is no 'stable, or homogeneous competence' underlying performance but a 'variable, or heterogenous capability' which changes according to situational factors. As summarised by Fulcher and Davidson (2007): "In other words, there are no constructs that really exist within individuals. Rather, our abilities are variable, and change from one situation to another" (p. 16).

A third stance is the interactionalist understanding of score meaning which sees consistent performance as a result of "traits, contextual factors, and their interaction" (Chapelle, 1998, p. 34). However, the interactionalist approach cannot be achieved by simply combining the trait and behaviourist approaches. This is because "when trait and context dimensions are included in one definition, the quality of each changes. Trait components can no longer be defined in context-independent, absolute terms, and contextual features cannot be defined without reference to their impact on underlying characteristics" (p. 43). Furthermore, the interactionalist perspective posits that "*performance is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts*" (Chapelle, 1998, p. 43).

Also, Chapelle points to the fact that there is a need for a component that controls the interaction between trait ad context, namely metacognitive strategies, or strategic competence (Bachman, 1990). As pointed out in Fulcher and Davidson (2012): "In this approach we acknowledge that the test contains only a sample of the situation or situations to which we wish to generalize. Part of investigating the validity of score meaning is therefore collecting evidence to show that the sample is domain-relevant, and predictive of the wider range of abilities or performances that we wish to say something about" (p. 17).

As noted by Chapelle (1998), the interactionalist approach poses difficulties in terms of assessment since it combines contrasting perspectives. Therefore, the interactionalist requires both context-specific considerations, as well as considerations that are person-specific. This makes the interactionalist approach ideal for tests requiring language use in discourse contexts, such as speaking tests. However, "unlike the behaviourist, who simply attempts to mirror the context of future language use to improve prediction, the interactionalist attempts to use discourse to elicit the defined linguistic knowledge, processes, and metacognitive strategies during test performance" (p. 48).

The interactionalist approach to construct validity has been articulated by several researchers (Chalhoub-Deville, 2003; Chalhoub-Deville & Deville, 2005; A. W. He & Young, 1998; Kramsch, 1986; McNamara, 2001; Young, 2000), who focus mainly on the context of the assessment of interactive speaking. These researchers often draw on literature outside of language assessment, e.g. sociolinguistics, ethnography and speech act theory. One example is Young (2011), who applies an interactionalist approach to both theoretical and empirical work on interactional competence (see further in Study II).

Bachman (2007) emphasised that the theoretical issues raised by the three approaches to construct validation, described by him as (1) trait/ability-focused, (2) task/context-focused, and (3) interaction-focused, have important implications and also present challenges "for both empirical research in language testing and for practical test design, development, and use" (p. 70). In addition, he maintains that all three approaches are valuable and need to be addressed in the design, development and use of language tests:

> These theoretical issues also provide valuable insights into how we can enrich the ways in which we conceptualize what we assess and how we go about assessing it. For research, they imply the need for a much broader, more catholic methodological approach, involving both so-called quantitative and qualitative perspectives and methodologies. For practice, they imply that *exclusive* focus on any one of these approaches (ability, task, interaction), to the exclusion of the others, will lead to potential weaknesses in the assessment itself, or to limitations on the uses for which the assessment is

> appropriate. This means that we need to address all three in the design, development, and use of language assessments. (p. 70-71)

Socio-cognitive framework

Weir (2005) proposed a framework of test validation in which test developers should generate evidence of the validity of a test from different perspectives. The framework is 'socio-cognitive' in that it considers both aspects of cognition, i.e. test-takers' cognitive abilities and processes, as well as the social context in which the task is performed. As explained by Shaw and Weir (2007):

> The framework is socio-cognitive in that the abilities to be tested are demonstrated by the mental processing of the candidate (the cognitive dimension); equally, the use of language in performing tasks is viewed as a social rather than a purely linguistic phenomenon. (p. 3)

A third, important dimension of the framework is the process of scoring. According to Shaw and Weir (2007), construct validity "is the result of the constructed triangle of trait, context and score (including its interpretation)" (p. 2-3). The approach is thus "effectively an *interactionalist* position which sees the construct as residing in the interactions between the underlying cognitive ability and the context of use – hence the socio-cognitive model" (p. 3). In differentiating it from earlier validation frameworks within language testing, Shaw and Weir (2007) emphasise that the socio-cognitive framework "seeks to marry the individual psycholinguistic perspective with the individual and group sociolinguistic perspective" (p. xi). The authors also argue that "the socio-cognitive approach helps promote a more 'person-oriented' than 'instrument-oriented' view of the testing/assessment process than earlier models/frameworks; it implies a strong focus on the language learner or test taker, rather than the test or measurement instrument, as being at the centre of the assessment process, and it acknowledges the extent to which the assessment process is itself part of a larger social endeavour" (p. xi).

Five main components of validity are described in the framework, which also includes an account of how the various validity elements fit together and interact, both *temporally* and *conceptually*. According to Weir (2005), the key types of validity evidence that a test developer needs to address to ensure fairness are:

- Context validity

- Theory-based validity (referred to as cognitive validity in subsequent publications)
- Scoring validity
- Criterion-related validity
- Consequential validity

Weir (2005) emphasizes that the types of validity evidence are not 'alternatives, but complementary' (p. 13). This implies that "[l]anguage testers need to give both the socio and the cognitive elements an appropriate place and emphasis within the whole, and avoid privileging one over another. The framework reminds us that language use – and also language assessment – is both a socially situated and a cognitively processed phenomenon" (Shaw & Weir, 2007, p. xi). A unified approach to establishing the overall validity of a test is thus adopted. Furthermore, the model comprises both *a priori* (before-the-test event) validation components, mainly represented by context and cognitive validity (theory-based) and *a posteriori* (after-the-test event) validation components, mainly represented by scoring validity, consequential validity and criterion-related validity. The various elements of the model are presented as being independent of each other for descriptive purposes. However, according to Weir (2005), "[t]here is a symbiotic relationship between context- and theory-based validity and both are influenced by, and in turn influence, the criteria used for marking which are dealt with as part of scoring validity" (p. 20). In other words, context, cognitive and scoring validity interact with each other. According to O'Sullivan and Weir (2011), the relationship between the elements in the model can be looked at in different ways. One way is to look at the 'core' elements of construct validity (cognitive, context and scoring validity) as "essentially inward-looking, in that they are focused on aspects of the test itself" (p. 24), whereas the consequence and criterion-related elements can be seen as primarily outward-focused.

Regarding the first component, context validity, the term, as used in Weir (2005), is equivalent with the traditional concept of content validity, or coverage of tasks. However, Weir (2005) argues that the term context is better to use to refer to 'the social dimensions of language use'. Context validity thus pertains to the representativeness, authenticity or coverage of test tasks in relation to 'the larger universe of tasks' from which the test is intended to be sampled: "This coverage relates to linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed arising from both the task itself and its administrative setting" (p. 19).

The second type of validity, theory-based validity, later referred to as cognitive validity by Khalifa and Weir (2009), refers to the cognitive processes underlying language use, in the form of both test-takers' cognitive processes in performing the task and the resources they bring to the test situation (e.g. knowledge of content and language ability).

Scoring validity is the third type of validity in the framework. It is related to both context and theory-based validity and refers to the consistency of scores: "In other words, it accounts for the degree to which examination marks are free from errors of measurement and therefore the extent to which they can be depended on for making decisions about the candidate" (Weir, 2005, p. 23). This can typically be investigated trough different types of reliability estimates. However, it should be noted that scoring validity is a concept that comprises an investigation of *all aspect*s of the scoring process, from rater recruitment and training to the scoring rubrics and the assignment of final grades.

The fourth type of validity evidence is criterion-related validity, which is synonymous with the traditional definitions of predictive and concurrent validity (see above). This component refers to the extent to which test scores reflect or correlate with "a suitable external criterion or performance" (Weir, 2005, p. 35). Predictive validity thus involves comparing test scores with some other measure of the same ability.

Finally, the last component of the framework is consequential validity, building on Messick's validity theory of the social consequences of testing. Consequential validity includes aspects such as washback, social impact and test bias, often examined *a posteriori*. However, in their reconceptualization of the framework, O'Sullivan and Weir (2011) (see below) instead view consequences as an *a priori* aspect of test validation, "seeing all decision taken in the development process form the perspective of their impact on the test taker" (p. 3).

As mentioned above, the socio-cognitive framework was updated in O'Sullivan and Weir (2011), see Figure 3 below. The reconceptualization of the model involves a reduction to three basic elements: the test system, the test-taker and the scoring system, in an attempt to make the model even more manageable.

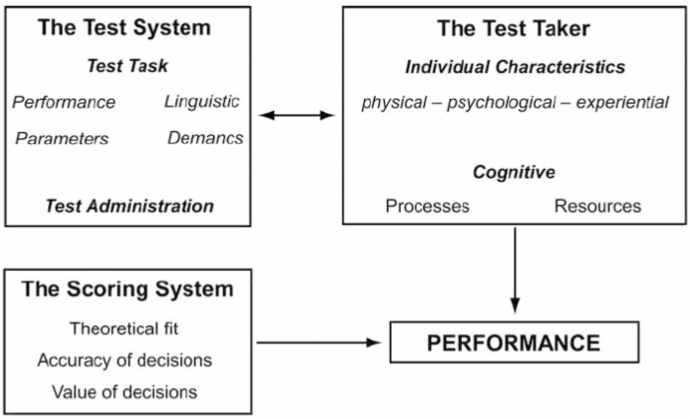

Figure 3. A reconceptualization of Weir's socio-cognitive framework (from O'Sullivan, 2011b, p. 261) © Routledge

In Weir (2005), the framework is illustrated and exemplified in relation to actual test examples and practice by applying it to the 'four skills', reading, listening, writing and speaking. This shows that the framework has great potential for practical operationalisation by language testers and teachers. The framework has been used in a range of test validation and development projects. For example, the examination board of the Cambridge ESOL has applied it to its examinations (Galaczi & Vidakovic, 2010; Khalifa & Weir, 2009; Shaw & Weir, 2007; Taylor, 2011). The framework has also been used in multiple international contexts (e.g., O'Sullivan, 2005). It has also provided a theoretical basis for CEFR linking projects (Kantarcioglu, 2012; O'Sullivan, 2011a; Wu, 2011).

Some weaknesses of the socio-cognitive framework have also been articulated. Xi and Davis (2016) state that, in comparison to argument-based frameworks (Bachman & Palmer, 2010; Kane, 2013), "the socio-cognitive framework does not explicitly include a formal argument structure for organizing validity claims and guiding research activities" (p. 75). On the other hand, Weir and Shaw (2005) explain that the socio-cognitive framework "is ostensibly concerned with specifying and inter-relating focus areas for the validation process rather than with how the validation case should be argued per se" (p. 10). Additionally, Xi and Davis (2016) claim that another weakness of the framework is that "overall, there is relatively little guidance regarding how questions should be prioritized when collecting evidence to support inferences" (p. 75).

By way of summarising, we can see that the language test validation frameworks presented in this chapter are all based on Messick's explication of validity (1989), according to which different types of validity evidence should be collected to support a unitary concept of construct validity. The frameworks are similar in this way. However, they conceptualize the different aspects or components of validity in slightly different ways. Also, they are more or less formally structured when it comes to organizing validity claims and guiding validation research. For the purpose of the present thesis, it was found that the socio-cognitive framework, which builds on the interactionalist approach to construct validity, was the most practical to use (see further in Chapters 5 and 7).

Chapter Five: Method and material

Three empirical studies were undertaken for the purpose of exploring the main research questions in this thesis, all relating to the speaking component of the national test of English in the Swedish upper secondary school, with a focus on the raters' perspective:

- What degrees of rater variability and consistency of rater behaviour can be observed? (Study I)
- What features of test-takers' performances are salient to raters? (Study I and II)
- How are the national EFL speaking tests administered and scored at the local school level? (Study III)
- What are teachers' views regarding practicality? (Study III)

Figure 4 offers an overview of the three studies with regard to participants and research focus. The design is further described in this chapter. As can be seen, Study I and II include the same sample of raters, 17 raters from the Swedish context and 14 from a CEFR-related context (see more information under Participants below). The first study investigates rater variability and raters' decision-making processes, while the second focuses on raters' interpretation of the construct of interactional competence. The third study includes 267 Swedish teachers of English who responded to a survey about their administration and scoring practices, as well as their perceptions of practicality.

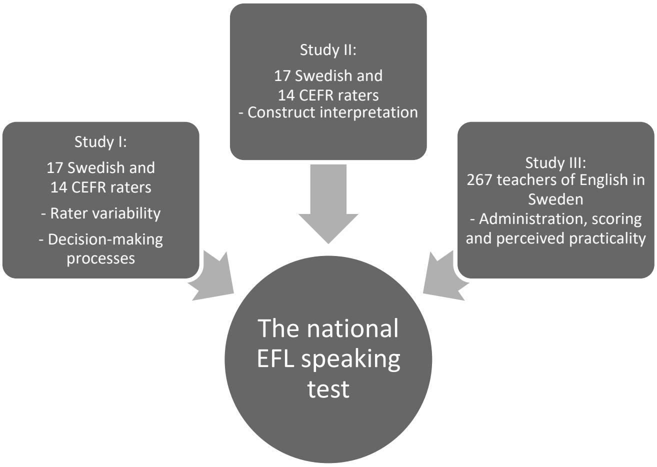

Figure 4. Overview of the three studies: participants and research focus

In terms of validation, it is possible to link the three studies to different kinds of validity evidence, as outlined in Chapter 4. A validation framework which is increasingly used within the context of language testing validation, and which offers a manageable structure for the present thesis, is the socio-cognitive framework for language testing (Weir, 2005). Based on this framework, Figure 5 provides an overview of the main kinds of validity evidence addressed in the three studies, although it must be remembered that the different validity elements, especially context, cognitive and scoring validity, interact (Weir, 2005). It should also be emphasised that the validity evidence gathered in this thesis is limited to the raters' perspective; hence, a comprehensive investigation of construct validity cannot be made. The relationship in Figure 4 will be further explored in Chapter 7 Discussion.

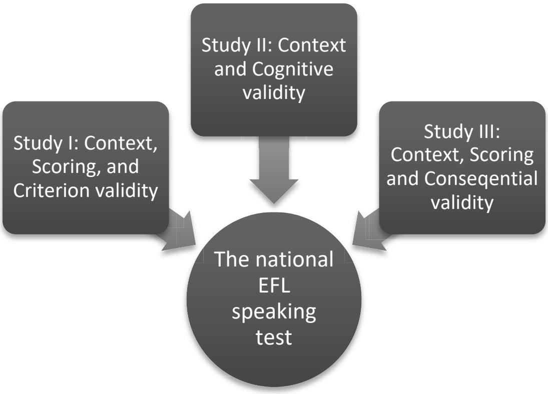
Study II: Context and Cognitive validity
Study I: Context, Scoring, and Criterion validity
Study III: Context, Scoring and Conseqential validity
The national EFL speaking test

Figure 5. Overview of the three studies: validation focus (Weir, 2005)

In the following sections, the speaking test, the participants, the material and the methods of analysis will be described.

The speaking test

The speaking test focused on in the present thesis is the oral component of the national test of English as a foreign language (EFL) in the upper secondary school. The upper secondary school in Sweden is based on courses. For the subject English there are three courses: *English 5*, *English 6* and *English 7*. *English 7*, however, is an elective course and is not included in the present investigation. The grading scale has the classifications A–F, with A being the highest grade and E the lowest passing grade. F means not passed. As mentioned in Chapter 2 Contextual background, the courses for foreign languages are related to, and partly comparable to, the common reference levels in the CEFR. For the two courses in question, the approximate, minimal pass level (grade E) is comparable to a high B1 (B1.2) for English 5 and a low B2 (B2.1) for English 6. This also applies to the national tests.

The speaking test task consists of a paired or group conversation in which students should express, develop, and discuss a given topic/theme on their own and in interaction with others. The task is divided into two parts; the first focusing on oral production and interaction, the second on oral interaction. Test instructions stipulate that two students, or possibly three, should take the test together. The teacher is responsible for composing groups. The students have about 15 minutes preparation time before the test, individually and privately, and the total time allowed for the speaking test is about 15 minutes.

The tests are assessed in relation to the national performance standards for oral production and interaction. To help teachers in making their holistic grading decision, analytic assessment factors describing qualitative aspects of spoken proficiency, are also provided. In addition, for each test administration, audio-recorded, commented sample performances, illustrating different grade levels, are included in the assessment materials. Also, detailed instructions and guidelines for the whole national EFL test, with a specific section about the administration and scoring of the speaking test, are given in a booklet. The test instructions strongly recommend that teachers record the oral tests, which is important for documentation. It also makes re-listening and co-rating possible. Co-rating, i.e. a process whereby teachers collaborate in the grading process, is strongly recommended; however not regulated.

During the test, the teacher is normally present but should keep in the background and let the students control the conversation. Teachers are instructed to point out to students that the responsibility for keeping the conversation going is a joint effort and that they should give each other equal speaking opportunities. Students are also encouraged to use communicative strategies, e.g. questions and comments that help bring the conversation forward, and production strategies, e.g. paraphrasing.

In Study I and II, six audio-recorded student conversations, amounting to twelve individual performances, from a pre-testing round of the tests for course English 6 were used (See Erickson & Åberg-Bengtsson, 2012, for a detailed account of the test construction). For Study III, which included a survey, the questions were asked with reference to the speaking tests in both courses (English 5 and English 6), which are based on the same model, consisting of two parts: Part 1 focuses on oral production and interaction and Part 2 focuses on interaction and discussion. In Part 1, students present something to their partner, in English 6 usually a short text they have read during the preparation time, followed by a discussion in the pair. Part 2 consists of a more general discussion about aspects of the theme. On the National Assessment Project webpage, sample tests are provided for reference[5].

[5] https://nafs.gu.se/prov_engelska/engelska_gymn/exempel

Participants

Test-takers

Study I and II used authentic data in the form of six audio-recorded student conversations, amounting to twelve individual performances, from a pretesting round of the national speaking test for course *English 6* at the upper secondary school. The pairs consisted of a female and a male student, to make it easier for raters to distinguish between candidates when listening to the recordings. The conversations were chosen to be representative of different proficiency levels. The pairs of test-takers were around 17 years old and came from different parts of Sweden. No other background information on the test-takers was collected.

Raters

The raters in Study I and II came from different contexts, the common denominator being the relation to the CEFR. The first group consisted of 17 formally qualified upper secondary teachers of English in Sweden, from eleven different schools in two national regions. Convenience sampling was used. The researcher contacted several upper secondary schools in two regions of Sweden and provided information about the study which was forwarded to teachers. The teachers were invited to a one-day seminar, when the data was generated. The time required to participate in the research project was thus one working day. Of the 17 teachers who volunteered to participate, there were four men and 13 women. Three of them were native speakers of English, and the others had Swedish as their first language (L1). As regards teaching experience, this ranged from 1-29 years (teaching experience was not a requirement). Four participants had little teaching experience (< 5 years) and the rest had worked for a longer time (6-29 years).

The second group consisted of 14 raters from two European countries; Spain (n = 7) and Finland (n = 7). They rated the same twelve performances as the Swedish raters; however, using the common reference levels of the CEFR. The methodological choice to include 'external' raters was motivated by the opportunity this provided to make a small-scale, tentative comparison between the national EFL standards in the Swedish school context and the reference levels in the CEFR. As mentioned in Chapter 2, the national syllabuses for foreign languages, including the performance standards, are tentatively related to the CEFR-levels. However, the seven steps of language proficiency defined in the Swedish system have not yet been fully empirically aligned to the six levels of the CEFR. For this reason, it was useful to include CEFR raters. In Study II, which investigated the construct of interactional competence from the raters' perspective, the inclusion of the CEFR raters (see below) also enabled a tentative analysis of the nature of, and relationship between, the descriptors for IC in the Swedish EFL performance standards and CEFR scales.

The raters from Finland and Spain were selected through purposeful sampling. There were twelve females and two males in this group. The raters were all EFL education professionals (working at schools/universities and/or ministries) with a high level of familiarity with the CEFR, as well as previous experience using CEFR scales. However, it should be mentioned that they had limited knowledge of the speaking test in the Swedish national testing context before taking part in the study. As with the Swedish raters, the Finnish and Spanish raters attended a one-day seminar (one for each group) led by the researcher, in connection with the data generation. Hence, the required time to participate was one working-day.

Study III did not include any scoring data. In light of this, the participants are categorised as 'teachers'. This study was survey-based, and the participants were 267 teachers of English in Sweden who responded to an online survey developed by the researcher. Of the respondents, 75% were female. The average age was 47, ranging from 26 to 68 years. The participating teachers had taught for an average of 16 years (range 1-42, *SD* = 10). As regards teacher certification, a majority of the respondents reported being certified EFL teachers (96%).

Rating scales

In Study I and II, two sets of rating scales were used, one from the Swedish context and one from the CEFR (see below). In Study III, the rating scale in the Swedish context was indirectly addressed, by survey-items asking about the perceived support of the different assessment materials available in conjunction with the national tests. The following account will therefore focus on the rating scales used in Study I and II.

In Study I and II, twelve students' performances were scored holistically by both rater groups; however, using different rating scales. The Swedish raters used the rating scale from the Swedish EFL syllabus, which included the national performance standards for oral production and interaction in English

(See Appendix 2 in Study I), graded on a ten-point scale. They were also provided with analytic assessment factors intended to be a support in arriving at the holistic rating decision, which are included in the assessment materials for the national tests. The analytic assessment factors describe qualitative aspects of oral production and interaction and are divided into two main areas: *content* and *language and ability to express oneself* (See Appendix 3 in Study I).

The CEFR raters used two complementary scales from *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment – A Manual* (Council of Europe, p. 184-186), covering the full range of CEFR levels (A1-C2) including 'plus levels': a global scale and an analytic scale. The analytic scale included five aspects: range, accuracy, fluency, interaction and coherence. Like the Swedish raters, the CEFR raters assigned holistic scores, based on a nine-point global scale (See Study I). Both rater groups were thus guided by analytic criteria to help them in their decision-making processes but they assigned holistic scores.

The fact that the rater groups in Study I and II used different rating scales, has implications for comparability. However, it should be kept in mind the two rating scales served different purposes. In the case of the Swedish raters, the scale served the purpose of examining inter-rater agreement. In the case of the CEFR raters, the scale was used to compare their judgements with the intended CEFR levels of the speaking test.

Data collection procedure

For Study I and II, data were collected during one-day seminars held with the participants in June, September and November of 2013. The structure and content of the seminars were identical for both the Swedish and the CEFR rater groups. After a general introduction including some information about the study as well as some basic training and familiarisation, the raters independently listened to the six conversations using headphones, with stops and repetitions where needed. Raters were asked to make notes while listening, which is typically part of the normal rating procedure. After listening to each conversation, raters wrote a summarising comment for each test-taker's performance, highlighting features that, in their opinion, contributed to the grade they had assigned. This was done on computer in a Word document. The comments varied in length with a mean value of 72 words, range 9-230 words.

As shown in other rater report studies (e.g., Ducasse & Brown, 2009), raters differed in quantity and type of comments made.

In Study III, an online survey was administered to teachers of English at upper secondary schools in Sweden during spring 2017. Simple random sampling was used to select 150 schools from a database, compiled by *Statistics Sweden*, including all upper secondary schools in Sweden (excluding adult education). The invitation of the survey was sent via email to the administration and head teacher of the 150 schools in the sample with a request to forward it to all English teacher at their school. Two reminders were sent out. This resulted in 267 individual responses. The response rate was relatively high (79% at the school level). Furthermore, the achieved sample was representative in terms of distribution between independent and public schools as well as geographic spread. No obvious non-response bias was found.

The questionnaire was constructed by the researcher and built on two sources: (1) test specifications and guidelines for the national speaking tests (Swedish National Agency for Education, 2016a) and (2) the framework of test usefulness outlined in Bachman and Palmer (1996). The questionnaire was pre-tested and piloted. The final survey included 60 items divided into four parts: implementation practices, (2) assessment in relation to policy documents and purposes of the test, (3) perceptions of test content and format, and (4) demographic information. A subset of items, focusing on teachers' implementation practices and views of practicality, were included in Study III.

In addition, three background variables were examined in relation to teachers' responses in the survey: (a) gender, (b) years of teaching experience, and (c) the size of the school where the teacher worked, measured by two variables. All the background variables were self-reported (See Study III for more details).

Methods of analysis

Study I and II

Analyses of scores

The quantitative data in the form of scores were first analysed using descriptive statistics, including measures of central tendency and measures of dispersion. Then, Spearman rank order correlations and Kendall's Tau correlations were performed for the pair-wise ratings of the Swedish raters, in order to assess inter-rater reliability. Finally, Cronbach's alpha, which measures internal consistency for the whole rater group, was calculated. More information on the estimates and their uses are given in Study I. SPSS Statistics, Version 21.0 (IBM Corp., 2012) was used to compute the statistical analyses.

Qualitative analyses of rater comments

As regards the qualitative rater comments used in Study I and II, two content analyses were performed (Galaczi, 2013; Green, 1998; Krippendorff, 2013). The steps taken to develop the coding schemes are thoroughly explained and exemplified in the studies. To validate the analyses, and to reduce coder subjectivity, co-coders were employed in both studies. In Study I, the main researcher and an assistant researcher with long experience of the Swedish educational context and familiarity with the CEFR, independently coded a subset (10%) of the data. The discrepancies in the interpretation of some of the coding categories were resolved through discussions and the coding scheme was subsequently revised with amendments of the category descriptions where necessary. The whole dataset was then coded by the main researcher independently. The final coding scheme included ten main categories, pertaining to the different components of communicative competence outlined in the CEFR, as well as a few categories that emerged from the rater comments:

- Accuracy
- Coherence
- Fluency
- Intelligibility
- Interaction
- Other
- Production strategies
- Range
- Sociolinguistic appropriateness
- Task realisation

In Study II, two researchers with PhDs in applied linguistics functioned as co-coders. In this study, an extensive coding process was undertaken in two steps. The first cycle of coding involved identifying relevant passages of raters' holistic comments that pertained specifically to interactional competence. On the basis of this, a draft set of coding categories was devised, based on categories used in previous research (Ducasse & Brown, 2009; Galaczi, 2008, 2014). In addition, the features of performance described in the rating scales used in the study, including the three descriptor scales for interaction strategies in the CEFR (2001, pp. 86-87), proved useful for the more detailed description of sub-categories. In the second cycle, the segments relating to interactional competence were further segmented into units of analysis and coded according to the coding scheme. The co-coders and the main researcher coded 45% of the total dataset independently. Discrepancies in the interpretation of the coding categories were resolved through discussions and relevant amendments to the coding scheme were made. The coding scheme was thus revised and reduced in a cyclic process, the end result being five main categories. In the last step, the researcher independently coded the whole dataset, according to these five main categories (See final coding scheme, including subcategories, in Appendix 3 in Study II):

- Topic development moves
- Turn-taking management
- Interactive listening strategies
- Interactional roles
- Additional comments on interaction

The development of a coding scheme and the steps taken to code qualitative data are part of a cyclic, iterative process, which involves checking for interrater agreement and revising categories until they are applied in a consistent manner by all coders (Galaczi, 2013). This process was followed in both Study I and II until satisfactory inter-rater agreement was reached. Interrater agreement was calculated using percent agreement; > 80% agreement at the main category level was considered satisfactory.

In both studies, the software NVivo 10/11 was used to organise and analyse the data. Frequency counts were undertaken to serve as an index of the salience of the features. Furthermore, in both cases the content analysis was mainly carried out deductively on the basis of existing theory and previous research, a so-called directed approach (Hsieh & Shannon, 2005)

Study III

Study III included mainly quantitative data from the responses to the survey. The respondents could also provide open-ended comments to some questions, which were used to illustrated the quantitative results. SPSS Statistics, Version 25 (IBM Corp., 2017), was used to compute the statistical analyses, which mainly included descriptive statistics and tests of association in relation to the background variables (See Study III for more details).

Summing up, both qualitative analyses in the form of content analysis, and quantitative analyses of scores and survey items, were utilised for the purpose of exploring the issues addressed in the main research questions of the thesis, enabling a triangulation of the findings. Some of the advantages and limitations of the methods and sampling techniques used will be further addressed in the section on reliability, validity and generalisability below.

Analytical stages

The empirical analyses, guided by the main research questions of the thesis, provided different kinds of validity evidence with regard to the interpretation and uses of test scores in the national speaking test under investigation. In Chapter 4 Theoretical framework, different validation frameworks in the context of language testing were presented. Weir's socio-cognitive framework was chosen as one way to structure the analytical steps. The different aspects of validity described in Weir (2005) were thus used as a basis for the analysis in Chapter 7 Discussion.

Reliability, validity and generalisability

The thesis aligns with a unified view of validity, proposed by Messick (1989), defined in the following way:

> This unified concept of validity integrates considerations of content, criteria, and consequences into a construct framework for testing rational hypotheses about theoretically relevant relationships, including those of an applied as well as of a scientific nature. The essence of unified validity is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the unifying force behind this integration is the trustworthiness of empirically grounded score interpretation, that is, construct validity (p. 5).

Validity is thus seen as multi-componential; however, with construct validity as the overarching concept.

Messick (1989b) further summarises the concept of validity as "an integrated evaluative judgement 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" (p. 13), stressing that validity is a matter of degree: a claim can be more or less valid, depending on empirical evidence and theoretical support. Furthermore, test providers may not always provide evidence about all forms of validity at one time. Validation can thus be seen as an ongoing process. In the three empirical studies of the present thesis, raters' scores, qualitative rater comments, and teachers' responses to survey questions intended to provide different aspects of validity evidence with regard to the interpretation and use of test scores in the national EFL speaking tests, as used in the Swedish school context.

The methods used, as all research methods, have advantages and disadvantages. The combination of quantitative and qualitative data allows for a triangulation of the data, which is a strength. As concerns content analysis (Study I and II), the issue of coder subjectivity poses threats to reliability. Steps were taken to mitigate this by using co-coders. Also, since the research findings were largely in line with previous research, this could be seen as providing support for the reliability of the studies. In addition, non-probability samples, as was used in Study I and II, are limited with regard to generalization. Since they do not truly represent a population, we cannot make valid inferences about the larger group from which they are drawn. However, for reasons of time, considering participation in the research project, and also relating to the analyses, it was not possible to include a larger sample of raters or candidates. It was therefore hoped that Study III, which includes a larger, more representative sample of teachers, would, to some extent, balance this 'insufficiency'. In Study III, sample representativeness may be an issue. Although steps were taken to select 150 schools randomly, the teachers who responded to the survey were self-selected. However, as the results of the survey are largely supported by data from the annual surveys conducted by the test constructors, reliability is strengthened.

Summing up, the combination of quantitative and qualitative approaches used in the three empirical studies provided different forms of validity evidence in relation to the speaking test under investigation. Some potential limitations of the methods have been discussed in this section; the strength, nevertheless, lies in the triangulation of data, which allows validity evidence from different perspectives and research methodologies to be collected and analysed.

Ethical considerations

Informed consent and confidentiality

The ethical guidelines of The Swedish Research Council were followed in the studies included in this thesis. The participating raters and teachers received oral and written information in the case of Study I and II, and written information in Study III about the purpose of the studies and the conditions of participation (See letter of information and consent for Study I and II in Appendix A and for Study III in Appendix B). In the collection and analysis of data, the anonymity of individuals and schools was protected, names being replaced by numbers and codes. Furthermore, students who participated in the audio-recorded performances had given their consent to the material being used for research purposes.

Chapter Six: Results

Three studies were conducted for the purpose of exploring different types of validity evidence in relation to a paired speaking assessment, as administered in the context of a high-stakes national test of English at the upper secondary level of the Swedish educational system. In this chapter, the objectives of each of the studies are specified and the main results are summarised, followed by a synthesis of the results.

Study I

The main purpose of the first study was to investigate rater variability and raters' decision-making processes in relation to a national EFL speaking test at the upper secondary level in Sweden. Furthermore, these two aims were combined in an attempt to explore the relationship between scores and raters' justifications of these scores. In addition, a subordinate aim was to make a small-scale, tentative comparison of Swedish performance standards for EFL and CEFR levels. The research questions guiding the analyses were:

- **RQ1**: What can be noticed regarding variability of scores and consistency of rater behaviour?
- **RQ2**: What features of test-taker performance are salient to raters as they make their decisions?
- **RQ3**: What is the possible relationship between scores and raters' justifications of these scores?
- **RQ4**: At what levels of the CEFR do external raters judge the performances of the Swedish students to be?

To explore these research questions, six authentic, audio-recorded paired conversations, amounting to twelve individual performances, from a Swedish national test of English at the upper secondary level (with a minimal pass corresponding to CEFR level B2.1) were used. Raters from two contexts related to the CEFR participated in the study: (1) formally qualified teachers of English (*n* = 17) from 11 different upper-secondary schools in two regions of Sweden, and (2) raters from Finland and Spain (*n* = 14). The methodological choice to include external raters was motivated by the opportunity this provided to make a small-scale, tentative comparison between the national EFL standards in the Swedish school context and the reference levels in the CEFR. In addition to assigning holistic scores, the 31 raters provided written verbal reports on features of the performances that contributed to their judgements. The scores were analysed using (a) descriptive, (b) correlational and (c) reliability statistics. A content analysis was performed on the written comments, which were segmented and coded using NVivo 10 software. The coding scheme was based on some of the illustrative scales for the different components of communicative competences (linguistic competence, pragmatic competence and sociolinguistic competence) and communication strategies described in the CEFR. Additional coding categories were also added, based on features emerging from the rater comments. Frequency counts were computed to serve as an index of the salience of features.

With regard to RQ1, the results from the descriptive statistics for the Swedish teacher raters' scores showed certain degrees of variability. Distinct rater profiles with differences in rater severity/leniency were also evident. For example, the means (on a ten-point scale) for the Swedish raters ranged from 5.6 for the harshest rater to 8.0 for the most lenient. There were also examples of rater profiles with central tendency and restrictions of range (Wilson & Case, 2000). Furthermore, the standard deviations indicate that some performances were more difficult for raters to agree on than others. When looking at pair-wise correlations between the Swedish raters' scores, the median was .77, using Spearman's rho coefficient, and .66, using Kendall's *tau-b* coefficient. This indicates acceptable inter-rater reliability, albeit with clear room for improvement. Finally, Cronbach's alpha coefficient was computed for the Swedish group of raters and the results point to high internal consistency in the group: .98.

With regard to RQ4, the results showed that the rank ordering of performances was fairly similar between the Swedish and the CEFR raters. In addition, the means of the CEFR raters' scores were between B1+ and C1 for all performances but two, which were borderline cases for some of the Swedish raters too. The CEFR ratings thus generally indicated that the Swedish students' performances were at the intended levels of the test.

RQ2 was analysed using a content analysis of raters' written comments, which revealed that a wide array of performance features was taken into account in the holistic rating decision, the majority of which were in alignment with the rating criteria provided, contrary to what previous research has suggested (A. Brown, 2007; May, 2006; Orr, 2002). The most salient performance features pertained to test-takers' linguistic and pragmatic competence, as well as their interaction strategies, in that order of frequency. There were surprisingly few comments coded as sociolinguistic features, the third component of communicative competence in the CEFR. Sociolinguistic competence, as defined in the CEFR, refers mainly to sociocultural aspects of language use, in particular social conventions, which it was indicated that this test offered limited opportunities for showing. Further, as part of the decision-making process, raters were found to compare and contrast test-takers' performances, a result also observed in previous research on the rating of paired speaking tests (May, 2011b; Orr, 2002). Cross-groups comparisons of the frequency counts of the coding categories indicated that the Swedish and CEFR raters had somewhat different rater orientations. Linguistic aspects, in the form of accuracy and range, was a highly salient category to both the Swedish and the CEFR raters. However, the CEFR raters had a somewhat more even distribution of comments with regard to the different components of test-takers' communicative competence.

Finally, as regards RQ3, the tentative findings indicated that score variability could be explained by two main factors: (1) raters noticed similar features but evaluated them differently, or (2) raters partly focused on different features.

In summary, the results of study I revealed that the ratings of the Swedish teacher raters exhibited certain degrees of variability, however, in general acceptable inter-rater reliability, albeit with clear room for improvement. Furthermore, the CEFR raters judged the performances to be, on average, at the intended levels of the test. In terms of the oral communicative ability of the test-takers, raters seemed to include a wide range of features in their holistic rating decisions, indicating broad content coverage. Finally, a tentative analysis of performances with large variations in scores showed that raters either noticed similar features but evaluated them differently or focused partly on different aspects of the performance.

Study II

In this study, involving the same raters and the same six conversations from the national EFL speaking test as Study I, the primary aim was to analyse features of interactional competence, i.e. students' ability to interact meaningfully with other participants, that raters attended to when they scored paired performances. Two research questions were addressed:

- **RQ1**: What features of interactional competence do raters attend to as they judge performances in a paired speaking test?
- **RQ2**: According to the raters, what characterizes more or less successful interaction?

Additionally, a tentative analysis of the nature of, and relationship between, the descriptors for IC in the Swedish EFL performance standards and CEFR scales was made.

The raters listened to six audio-recorded paired conversations and provided written comments to explain their rating decision. The raters' comments were then analysed following procedures for content analysis (Galaczi, 2013; Krippendorff, 2013). Two external researchers functioned as co-coders and contributed to the development of a coding scheme in a successive validation process, until satisfactory interrater agreement was achieved. The final coding scheme consisted of five main categories:

- Topic development moves
- Turn-taking management
- Interactive listening strategies
- Interactional roles
- Additional comments on interaction

The findings of the content analysis align well with previous research, both earlier rater report studies (Ducasse & Brown, 2009; May, 2011b), as well as studies using conversation analytic methodology (Galaczi, 2008, 2014), showing that the two research methodologies provide complementary perspectives in terms of construct conceptualisation. It was indicated that raters paid attention to three main interactional resources employed by test-takers: *topic development moves*, *turn-taking management* and *interactive listening strategies*. These were seen as contributing to successful interaction when used in a collaborative and mutual manner, with test-takers actively monitoring and responding to their partner's speech. In comparison, less successful interaction strategies were characterized by weaker alignment between test-takers and a lower degree of collaborative and interpersonal moves.

The first category, *topic development moves*, was drawn from rater comments on test-takers' efforts to stimulate and develop the content of the conversation as an interlocutor, for example by initiating, developing and connecting topics in a mutual or cooperative manner that helped the discussion advance. The rater comments indicated that extensions of both self-initiated and partner-initiated ideas were important for successful interaction. This category also comprised comments on test-takers' use of questions as topic development moves.

The second category, *turn-taking management*, encompassed rater comments on test-takers' ability to initiate and maintain discourse in an appropriate way, as well as comments on how natural and automatic turn-taking was perceived to be. Raters used expressions like 'conversational fluency', 'natural turn-taking', 'keeping up and talking active part in the conversation', 'maintain the conversation flowing' to characterise interactional flow. Aspects connected to speaker change were also commented on, such as turn length.

The third interactional resource identified in the rater comments was *interactive listening strategies*, which included comments on test-takers' efforts to display attention or engagement while listening. Listening as part of a test-taker's interactive skills was divided into three subcategories. The first subcategory, *confirmations*, was the largest and comprised comments on test-takers' ability to actively monitor partner's speech and confirm mutual understanding, e.g. by giving feedback on and responding to their partner's contributions. The second subcategory, *clarifications*, was drawn from comments on test-takers' efforts to respond to interactional trouble by asking for or giving clarification or help. Finally, the last subcategory, *flexibility*, pertained mainly to Swedish raters' comments on test-takers' ability to accommodate speech to the situation and recipient; an aspect emphasized in the national EFL performance standards.

In addition, it was found that raters considered the impact of test-takers' interactional roles on scores. This was evident in one pair which was characterized by an asymmetric interaction pattern with one dominant and one passive speaker. Opinions among raters differed as to the effect of the dominant interactional style of the female speaker on the more passive partner's performance, highlighting the challenge of rating co-constructed interaction. Raters also paid attention to how test-takers performed in *comparison*, or in relation, to one another, and how test-takers' performances were interrelated, thereby acknowledging the inherently co-constructed and interpersonal nature of interactional competence.

Although not the main focus of the investigation, a cross-group comparison with regard to the relative salience of the coding categories was made, indicating some differences in rater orientations. For example, the Swedish raters made proportionally more comments on topic development moves, while the CEFR raters commented on turn-taking strategies more frequently than the Swedish raters. The reason for this may be related to differences in the wording of the descriptors for the national EFL performance standards and the CEFR scales, which emphasise slightly different aspects of interactional competence.

In summary, the findings of the study correspond, in a broad sense, to what has been shown in other studies of paired oral testing and further emphasize the need to take contextual as well as individual factors into account, thus including the variability inherent in social interaction as part of the construct conceptualisation.

Study III

Study III is a survey-based study, with the aim of providing a stakeholder perspective of the national EFL speaking tests in the Swedish school system by exploring self-report data from upper-secondary teachers of English regarding their implementation practices and views of practicality. The national tests are centrally designed and developed. However, since Sweden has a highly decentralized school system, the responsibility for the implementation of the oral national tests is entrusted to the head teacher who should plan the organisation together with his/her staff at the local school level. This means that the organisation may look different at different schools, which has implications for standardisation. With this in mind, the following research questions were addressed in Study III:

- **RQ1**: How do teachers implement the national EFL speaking tests in the Swedish upper secondary school?
- **RQ2**: What are teachers' views regarding the practicality of the national EFL speaking tests and what potential challenges do they identify?

- **RQ3**: Do teacher background variables, more specifically gender, teaching experience and the size of the school, relate to their practices and views of the national EFL speaking tests?

The survey was distributed to a sample of 150 randomly selected upper secondary schools; 267 teachers responded. The analyses were mainly based on descriptive statistics and tests of association with background variables. Three background variables, which were self-reported, were examined in order to find out whether teachers' practices and views differed with respect to (a) gender (a) teaching experience and (b) the size of the school where the respondent worked.

As regards RQ1 and RQ2, concerning implementation practices and challenges related to this, the findings indicate that there were variations in how the speaking tests were carried out and administered at the local school level, which is a result of the decentralized responsibility of the implementation. A majority of teachers conducted the tests during their regular English lessons (61%), which was a concern as this took time from teaching. Teachers working at schools where the tests were centrally organized and scheduled, with the help of the school management, seemed to find this solution less stressful. The organisation of the oral tests was thus shown to be a crucial issue.

Recording of the speaking tests is strongly recommended in the test guidelines, a main objective being that it makes re-listening possible and facilitates co-rating. Whereas the results indicated that almost half of the teachers recorded the oral tests, there was still a large group that did not. The main reason mentioned for not recording was lack of time for re-listening. Furthermore, the results of the survey revealed that the majority of the teachers in the sample grouped students in pairs, but it was also quite common to use groups consisting of three and in some cases four students. This variation in number of students per group clearly pose a challenge to standardisation, and potentially has consequences for students' results.

With regard to composing groups, teachers reported considering aspects such as students' proficiency level, their talkativeness and communication style, as well as inter-personal relations. A careful matching of students was seen to be an essential task to teachers. Previous research of the paired and group format is inconsistent as regards interlocutor effects and their potential effects on scores; however, it is indicated the matching of students is an aspect that needs special attention in the paired speaking test format. The study therefore recommends to include more explicit advice in the test guidelines regarding this aspect.

In terms of scoring, teachers generally found the assessment materials to be of good support. The analytic assessment factors and the benchmarked and commented samples of oral performances were perceived most favourably, whereas the national performance standards were seen to provide acceptable support. Furthermore, the results of the survey showed that it was most common for teachers to assess the oral tests alone without co-rating (42%), although a fair number of teachers reported that they assessed some (36%), many (6%) or all (13%) of the performances in collaboration with colleagues. In general, teachers were positive towards co-rating and thought it would contribute to a fairer assessment but many expressed that they did not have time due to heavy workload.

It was also found that a majority of the participating teachers perceived that they did not receive enough support from the school management which implies that they were left to solve the organisation of the oral tests on their own, contrary to the national directives of delegation. Many teachers pointed to the need of extra administrative support in terms of organising the oral tests, providing extra time for aspects such as co-rating, and taking in extra staff to supervise the class while the teacher administered the speaking tests. In terms of material resources, almost half of the teachers stated that there were enough rooms available at their school to carry out the national EFL speaking tests in an efficient way, whereas the other half claimed there were not. Teachers at schools where there was a shortage of group study rooms remarked that this was a stressful factor.

Two general questions were asked about teachers' perceptions of the practicality of the speaking tests (RQ2). It was indicated that teachers found the practical implementation to be somewhat problematic and quite time-consuming; however, with great variation in answers. Furthermore, the teachers generally found the instructions to be clear and easy to follow, although not always possible to adhere to in practice.

Statistical tests of association were undertaken to explore potential group differences (RQ3). Gender and teaching experience did not account for the variation in teachers' practices and views to any great extent; however, school size seemed to be more strongly related. It was suggested that teachers at smaller schools experienced more practical problems with the speaking tests and found them to be more time-consuming than teachers at larger schools, possibly related to the fact that at smaller schools the implementation of the oral tests is left to the individual teacher to a greater extent than at larger schools.

In summary, the results showed that there were variations in how the national speaking test was implemented at the local level. This has clear implications for standardisation, but must be considered in relation to the decentralized school system that the test is embedded in. Further, contrary to national directives, many teachers perceived that they did not receive enough support from the school management, indicating that clearer routines and administrative support are needed.

In Table 2, a summary of the three studies included in the thesis is offered.

Table 2. Summary of studies included in thesis

	Study I	Study II	Study III
Title	Looking beyond scores. A study of rater orientations and ratings of speaking	Assessing Interactional skills in a paired speaking test: Raters' interpretation of the construct	Evaluating a High-Stakes Speaking Test: Teachers' Practices and Views
Main purpose	to investigate rater variability and raters' decision-making processes in relation to a national speaking test of English at the upper secondary level in Sweden	to analyse features of interactional competence (IC) that raters attended to when they scored paired performances	to provide a stakeholder perspective of the national speaking tests of English in the Swedish school system by exploring self-report data from teachers of English regarding their implementation practices and views of practicality
Research questions	RQ1: What can be noticed regarding variability of scores and consistency of rater behaviour? RQ2: What features of test-taker performance are salient to raters as they make their decisions? RQ3: What is the possible relationship between scores and raters' justifications of these scores? RQ4: At what levels of the CEFR do external raters judge the performances of the Swedish students to be?	RQ1: What features of interactional competence do raters attend to as they judge performances in a paired speaking test? RQ2: According to the raters, what characterizes more or less successful interaction?	RQ1: How do teachers implement the national EFL speaking tests in the Swedish upper secondary school? RQ2: What are teachers' views regarding the practicality of the national EFL speaking tests and what challenges do they identify? RQ3: Do teacher background variables, more specifically gender, teaching experience and the size of the school, relate to their practices and views of the national EFL speaking tests?
Methods of analysis	a) descriptive, correlational and reliability statistics b) content analysis of raters' written comments	content analysis of raters' written comments	Survey-based analyses: a) descriptive statistics b) tests of association c) thematic analysis of open-ended comments
Main findings	• the test format has the potential of eliciting a wide range of a test-taker's oral communicative competence • inter-rater agreement results point to certain degrees of variability, however, in general acceptable inter-rater reliability, albeit with room for improvement.	• raters paid attention to three main interactional resources: topic development moves, turn-taking management and interactive listening strategies • results suggest that assessing co-constructed discourse pose challenges with regard to asymmetric interaction	• variations in how the national speaking test is implemented at the local level, in line with the decentralised responsibility; however with implications for standardisation • indication that more resources are needed for crucial aspects such as the organisation of the speaking tests, and co-rating

Synthesis of results

The overall purpose of the present thesis was to explore different aspects of validity evidence in relation to a paired speaking assessment, as administered in the context of a high-stakes national test at the upper secondary level of the Swedish educational system.

The results of study I suggest that the paired speaking test format has the potential of eliciting a wide range of a test-taker's oral communicative competence, not least interactional skills. With regard to inter-rater agreement, results point to certain degrees of variability, however in general, acceptable inter-rater reliability, albeit with obvious room for improvement.

The results of study II, investigating the assessment of students' interactional skills, indicated that raters paid attention to three main interactional resources, *topic development moves*, *turn-taking management* and *interactive listening strategies*, which were viewed as positive when employed by test-takers in a mutual and reciprocal way. Further, results suggest that assessing co-constructed discourse posed challenges with regard to asymmetric interaction.

Finally, the results of Study III, highlighting the administration and scoring of the oral national EFL test, showed that there were variations in how the national speaking test was implemented at the local level, which is in line with the decentralised responsibility of the implementation, but has implications for standardisation. It was indicated that many teachers did not receive enough support with regard to the organisation of the speaking tests, which may be a bigger issue at smaller schools. Further, the results point to a need of more resources for crucial aspects such as co-rating.

In the next chapter, the results of the studies are discussed.

Chapter Seven: Discussion

The overall purpose of the present thesis was to explore different aspects of validity evidence from the raters' perspective in relation to a paired speaking assessment, as administered in the context of a high-stakes national test at the upper secondary level of the Swedish educational system. More specifically, three areas were investigated: (1) the scoring process, (2) the construct underlying the test format, and (3) the setting and test administration. These main areas will be discussed using validity evidence gathered from the three empirical studies undertaken in the present thesis.

As validity is a multifaceted concept, different types of validity evidence, gained from both quantitative and qualitative approaches, are necessary in order to support or refute claims made regarding the validity of test score interpretations and use. In line with this thinking, validity should be viewed as a continuum, not as an 'all-or-nothing proposition', and validation is an ongoing process. This is a presupposition of the following analysis, in which an *overall* evaluative judgement of validity evidence will be made.

Weir's (2005) socio-cognitive framework of language test validation (see Chapter 4 Theoretical Framework) was considered a useful methodological tool and was therefore used to structure the discussion of the different aspects of validity evidence gathered. It should be kept in mind that the discussion is limited to the type of validity evidence collected in the present thesis, which focuses on the raters' perspective; hence, a comprehensive investigation of construct validity cannot be made. In this sense, it is a partial validation, in which relevant components of the socio-cognitive framework are used to guide the analysis. It should also be emphasised that the validity evidence discussed is collected *a posteriori*, in other words after the test has been operationalized. In this way, the analysis can be used to inform test development. However, in Weir (2005), context, cognitive and scoring validity are described as *a priori* aspects of validation, typically collected during the test development process, whereas criterion and consequential validity are characterized as *a posteriori* aspects. In the updated version of the model (O'Sullivan, 2011b; O'Sullivan & Weir, 2011), this was slightly changed with regard to consequential validity, which is now considered an *a priori* aspect, "seeing all decisions taken in the development process from the perspective of their impact on the test taker" (O'Sullivan, 2013, p. 3).

Weir (2005, p. 48-49) argues that test developers are obliged to seek to address all of the following questions, related to the different kinds of validity:

- How are the physical/physiological, psychological and experiential characteristics of candidates catered for by this test? (Test taker)
- Are the characteristics of the test task(s) and its administration fair to the candidates who are taking them? (Context validity)
- Are the cognitive processes required to complete the tasks appropriate? (Cognitive/Theory-based validity)
- How far can we depend on the scores on the test? (Scoring validity)
- What effects does the test have on its various stakeholders? (Consequential validity)
- What external evidence is there outside of the test scores themselves that it is doing a good job? (Criterion-related validity)

It was not possible to address all these questions in a comprehensive way within the scope of the present thesis. Especially the first question, regarding the characteristics of the test taker, has not been directly addressed. Since the empirical basis for the current thesis focuses on the raters' perspective, the validity evidence discussed is obviously restricted to this viewpoint. Furthermore, as pointed out by Weir (2005), "[t]here is a symbiotic relationship between context- and theory-based validity and both are influenced by, and in turn influence, the criteria used for marking which are dealt with as part of scoring validity" (p. 20). In other words, the validity components of the framework – especially context, cognitive and scoring validity – overlap and influence each other. The discussion is structured according to the five aspects of validity described in the framework; however, as they interact, more than one validity component may be addressed in the same section.

Context validity

The response format used in the national EFL speaking test in the Swedish school context is a paired speaking test, involving peer-peer interaction. The purpose of the speaking test is to test oral production and interaction, i.e. the ability to express oneself and communicate in spoken English. The test task consists of a conversation in which students should "speak about, develop their thoughts on, and discuss" a given theme "on their own and in interaction with others". The test consists of two parts: in Part 1, focusing on oral production and interaction, students present something to their partner, for example a short text they have read during the preparation, followed by a discussion in the pair. Part 2 consists of a more general discussion about the given topic, focusing on interaction.[6] The speaking test task is intended to elicit a broad range of language functions: *informational*, e.g. expressing opinions and elaborating; *interactional*, e.g. agreeing and disagreeing; and *interaction management* functions, e.g. initiating and reciprocating (O'Sullivan et al., 2002).

As pointed out in previous research (Brooks, 2009; Galaczi, 2008; O'Sullivan et al., 2002), the paired speaking test format has been found to elicit a richer speech sample and more varied interaction between participants than the singleton interview format. Albeit not without complications, the paired speaking test format may also allow for a more authentic and direct representation of spoken interaction and production, as it is likely to resemble natural conversations in real-life. In other words, there is good reason to believe that this format has advantages in terms of content representativeness and situational authenticity (Bachman & Palmer, 1996). In addition, the test specifications for the national EFL tests explicitly state that "the ambition is that the test should have a high degree of authenticity. This means on the one hand that the test materials are as authentic as possible and on the other hand that test tasks are possible to imagine in non-test situations as far as possible" [translated from Swedish] (Swedish National Agency for Education, 2017a, p. 13).

Weir (2005) emphasises that context validity "relates to linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed arising from both the task itself and its administrative setting" (p. 19). In light of this, the following section discusses the validity evidence gathered from the empirical results in Study I, II and III regarding a) the construct underlying the test and b) the setting and test administration.

[6] On the National Assessment web page, sample tests are provided for reference: https://nafs.gu.se/prov_engelska/exempel_provuppgifter.

Construct conceptualisation

In Study I, raters' decision-making with regard to attention to different features of test-takers' performances in a paired speaking test was analysed through a content analysis of raters' written justifications of scores. It was found that raters observed a wide range of students' oral competence. The results of Study I thus seem to support the assumption that the test format of peer-peer interaction allows for a broad representation of the construct of oral proficiency, as conceptualised in theoretical models of communicative and interactional competence (e.g., Bachman & Palmer, 1996; Canale & Swain, 1980; Council of Europe, 2001; Young, 2011). Furthermore, the raters' comments did not indicate in any way that the speaking test was framed as an 'assessment event' rather than a meaningful communicative exchange by students, which has been found to be a threat to the authenticity of paired/group speaking tests in school-based assessments in previous research (L. He & Dai, 2006; Lam, 2015; Luk, 2010). On the contrary, raters' comments, particularly evident in Study II, indicated that students were capable of interacting in a meaningful way together.

In addition, the results of Study I and Study II confirm the view that the construct of interactional competence should be seen "both as a cognitive and a social interactional trait, with emphasis not just on the knowledge and processing dimension of language use, as seen in the Bachman and Palmer (1996) model, but also on the social, interactional nature of speaking, which has as its primary focus the individual in interaction" (Galaczi & Taylor, 2018, p. 3). In other words, through raters' comments it was possible to observe both cognitive, individual and context-dependent, co-constructed features of test-takers' performances. This also implies that the variability inherent in peer-peer interaction should not primarily be considered as a source of *construct-irrelevant variance* but as part of the construct. This holds obvious challenges for assessment, which will be further explored below.

Guided by the rating scales used in the study, which are both based on the CEFR (Council of Europe, 2001), it was shown that the Swedish and CEFR raters paid attention to test-takers' linguistic competence, by commenting on aspects of accuracy and range, as well as their pragmatic competence, by commenting on aspects of fluency and coherence. Furthermore, raters made frequent reference to students' use of communication strategies, both in the form of interaction strategies and production strategies (The cognitive processes involved in strategy use will be further explored in the following section on cognitive validity). However, there were surprisingly few comments pertaining to test-takers' socio-linguistic competence, which is the third component of communicative competence in the CEFR (Council of Europe, 2001).

According to the CEFR, sociolinguistic competence is concerned with social conventions of language use, such as "linguistic markers of social relations, politeness conventions, register differences; and dialect and accent" (Council of Europe, 2018, p. 137). This result may therefore suggest that a paired speaking test task, as used in the context of a school setting with two non-native speakers of the same age, offers limited opportunities for demonstrating socio-linguistic competence, if not used with a prompt that will specifically elicit such language use. As mentioned in Plough, Banerjee, and Iwashita (2018), task design is an area in need of further exploration in the context of interactive speaking tests, in order to maximize the interactional features elicited from the format.

When cross-group comparisons were made, somewhat different rater orientations were found in both Study I and II, which may be related to the fact that the two rater groups used different rating scales. Rating scales and criteria are elements typically addressed as aspects of scoring validity in the socio-cognitive framework, but will be discussed as an element of context validity here, as they are closely related to construct conceptualisation. Study I showed that linguistic aspects, in the form of both accuracy and range, was a highly salient category to both the Swedish and the CEFR raters. However, the CEFR raters had a somewhat more even distribution of comments with regard to the different components of communicative competence (accuracy, coherence, fluency, interaction and range). In previous research of foreign/second language oral tests, it has often been demonstrated that "across levels grammatical accuracy is the principal determining factor for raters assigning a global score, with some variations in contribution of other factors depending on level" (Iwashita, Brown, McNamara, & O'Hagan, 2008, p. 27). However, taken together, the results of the qualitative coding of rater comments in Study I demonstrate that both the Swedish and CEFR raters took a wide array of features into account in their holistic rating decision, pointing to a comprehensive and broad view of test-takers' communicative competence. In the case of the CEFR raters, the more balanced distribution of comments in relation to communicative competence is very likely a result of the CEFR scales used, which include a more detailed and comprehensive conceptualisation of the construct (see Appendix 4, Study 1) than the EFL performance standards in the Swedish school context, which are expressed in a generic way (see Appendix 2 and 3 in Study I).

In Study II, a cross-group comparison of the relative salience of aspects of interactional competence also demonstrated differences in rater orientations between the teacher raters from Sweden, who were guided by the national EFL performance standards, and the group of external CEFR raters, who were guided by CEFR scales, suggesting that the rating scales emphasise slightly different facets of interactional competence. Whereas *topic development moves* and *interactive listening strategies* were more frequently mentioned by the teacher raters from Sweden, the CEFR raters made proportionally more comments in the categories *turn-taking management* and *additional comments on interaction*. The reason for this may be found in differences in the wording of the rating scale descriptors with regard to the conceptualisation of interactional competence. For example, in the Swedish assessment materials, the development of ideas is highlighted in terms of *complexity and variation – that test-takers should be able to give different examples and perspectives of the topics discussed* and *use communicative strategies to develop and advance the conversation*. In comparison, turn-taking is a prominent feature in the CEFR scales, whereas this feature is not explicitly articulated in the EFL performance standards in the Swedish context.

Similar to what has been demonstrated in previous rater orientation studies (May, 2011b; Orr, 2002), Study II showed that raters' interpretation of interactional competence provided a more comprehensive view of the construct than was reflected in the rating scales. As Brooks (2009) noted: "there was a lot more going on in the paired format than the rating scale captured" (p. 361). In light of this, rating scales need be further developed, representing the reciprocal and mutually constructed characteristics of interaction, as well as illustrating the progression of IC skills more clearly.

On the other hand, the results of Study I and II also indicated that raters focused mainly on the criteria expressed in the rating scales, which is contrary to what previous research studies on the rating of speaking tests have suggested (A. Brown, 2007; May, 2006; Orr, 2002). The group of non-criterion features identified in the content analysis was small, and mainly comprised comments that were relevant to a valid interpretation and use of test scores. Despite this, it should be remembered that the tentative analysis of the relationship between raters' comments and scores for performances with a large degree of variability in Study I highlighted raters' differential evaluations of the same performance.

This implies that although raters pay attention to similar criterion features, they may still interpret the descriptors in different ways in relation to student performances. Two patterns that could potentially explain score variability were identified: (1) raters noticed similar features but evaluated them differently, or (2) raters partly focused on different features.

It can thus be seen that rating scales play an important role in construct conceptualisation. As there seems to be great potential in enhancing teachers' assessment literacy in connection with their involvement in high-stake testing (Harlen, 2005; Malone, 2017; Xerri & Vella Briffa, 2018), the findings of the present thesis suggest that policy makers should invest resources into rater training and professional development (see, e.g., Daly et al., 2011), so that teachers can regularly meet and discuss the grading of student performances in relation to the criteria, preferably with the help of benchmarks or exemplars, in order to develop a shared understanding of the standards. This could also mitigate tendencies to weight some criteria over others. In a similar vein, Graham et al. (2012) maintain that the "[c]urrent thinking about rater training emphasizes developing a common understanding among evaluators so that they will apply the rating system as consistently as possible" (p. 15). This common understanding is often called Frame of Reference (FOR) training, and has the potential of addressing some of the main sources of rater bias identified in previous research (Hoyt & Kerns, 1999), namely "lack of overlap among what is observed, discrepant interpretations of descriptor meanings, and personal beliefs or biases" (Graham et al., 2012, p. 15).

In summary, the empirical evidence in Study I and II seems to support the assumption that the test format allows for a broad and authentic representation of the construct of oral proficiency. This is further strengthened by the close connection between the test format and the emphasis on oral interaction and production in the foreign language syllabuses in the Swedish school system.

Interactional roles and co-constructed interaction

In both Study I and II, it was found that raters, as part of their decision-making process, compared and contrasted test-takers' performances. This has been demonstrated in previous research on paired orals, e.g. in Orr (2002) and May (2006, 2011a). The authors in these studies were concerned that comparisons could be seen as a form of relative judgement, in that test-takers' performances were assessed in comparison with one another instead of in relation to the criteria. However, the fact that raters make comparisons can also be seen as a way to acknowledge the co-constructed and intersubjective nature of the construct. The construct of interactional competence has been described and studied by several scholars, e.g. by Young (2011), who stresses that "IC is not to be described in the knowledge and actions of an individual participant in an interaction; instead, IC is the construction of a shared mental context through the collaboration of all interactional partners" (p. 428). In light of this, considering test-takers' performances in relation to the other group participants seems justified as part of raters' operationalisation of the construct.

Both Study I and II showed that raters reflected on the impact of test-takers' interactional roles on scores. This was most evident in one of the conversations which was characterised by 'asymmetric interaction', where the female student was perceived to be dominating the discussion, whereas the boy had a more passive role. As demonstrated both in May (2009) and the present investigation, raters had difficulty agreeing on whether to penalise or compensate test-takers for their roles in asymmetric interaction (Galaczi, 2008). Some raters noted that the dominant interactional style of the female speaker interfered with her partner's capacity to demonstrate his full potential and could thus be perceived as disadvantageous. However, opinions differed.

The issue of 'interlocutor effects', where an individual test-taker's performance may be affected by the way the conversation is co-constructed with the partner they are interacting with, can be addressed from different perspectives. As Brooks (2009) states, with reference to Canagarajah (2006): "Perhaps rather than being viewed as a threat to construct validity, variability in interaction should be embraced as being more reflective of real world communication" (p. 361). This view is supported by the socio-cognitive definition of interactional competence where "the interlocutors and the host of variables they bring to the interactional event become part of the construct of L2 interaction and have implications for the validity considerations supporting the assessment" (Galaczi & Taylor, 2018, p. 3). On the other hand, we cannot base a validity argument on the authenticity or representativeness of the test format alone. We also need to consider the consequences in terms of reliability, standardisation and fairness. As interlocutor effects may also be seen as a source of *construct-irrelevant variance*, Van Moere (2013) stresses that test developers need

> to be aware of, and preferably eliminate or reduce, the factors which might affect the performance of each individual test taker. This might be accomplished by: designing the test tasks so as to reduce unwanted effects; scheduling candidate groupings or pairings in advance, where possible, to ensure that candidates are not tested together if they might advantage or disadvantage one another unfairly; or by conducting research to show that perceived sources of unwanted variance do not negatively impact performance in the testing context. (p. 3)

Clearly, more research is needed to show if and how the interactional roles of test-takers impact performance and ultimately scores in the context of asymmetric interaction patterns. For example, in the present investigation, the speaker perceived as dominant received the highest average grade in the sample of twelve performances, and the more passive speaker also performed above average. It may even be the case that the passive speaker's performance was helped by the more dominant partner, as was claimed by some raters. This is an interesting difference in comparison with Galaczi's study (2008), in which collaborative pairs, with test-takers collaborating equally and mutually in the conversation, achieved the highest average grades. The complexity of asymmetric interactions was further highlighted in Davis (2009), where one of three candidates who participated in both collaborative and asymmetric interactions, received a higher score when engaged in dominant asymmetric interaction. Davis concluded that "[a]lthough collaborative interaction was generally associated with higher level examinees and scores, there did not appear to be a penalty in terms of score when an examinee's interlocutor was unable to maintain a collaborative interaction." (p. 387). It should also be remembered that national test results in the Swedish educational system have an advisory function. The national test results are intended to be used in combination with teachers' continuous assessment in the classroom. In other words, students have more than one chance of demonstrating their interactional skills, most likely in different group constellations.

Summing up, it is suggested by Weir (2005) that context validity can be addressed by asking the question whether the characteristics of the test tasks are fair to candidates who are taking them. The evidence presented here does not unequivocally support this claim. To build a stronger argument, more empirical evidence is needed to demonstrate how and if the interactional roles of speakers affect test scores. However, it seems clear that interlocutor effects have to be considered when results are interpreted, and guidelines for raters need to be elaborated, including conceptually grounded reasoning as well as commented examples.

Test administration

As pointed out by Weir (2005) in relation to test administration: "Primary considerations affecting validity are the circumstances under which the test takes place. These conditions need to be similar across sites or the processing will differ. If the test is not well administered, unreliable results may occur" (p. 82). In evaluating context validity, the evaluator should therefore ask the following question: Was the test administered in the same manner across sites?

To answer this question, some of the evidence gathered in Study III will be discussed, focusing on the educational context of the current investigation. Study III showed that there were variations in how the national speaking tests were implemented at the local level (See Study III for examples), which has implications for standardisation. This is obviously an issue that needs to be treated from different angles, not least including possible consequences for students. Nevertheless, the results must also be considered in relation to the educational system that the national tests are embedded in.

Sweden has a decentralised school system, where national tests are centrally developed but internally marked by teachers. The responsibility for the implementation of the oral national tests is delegated to the head teacher who should plan the organisation together with his/her staff and decide on the most appropriate solution. The National Agency for Education (2018a) therefore concludes that "[t]he most suitable organisation of the oral national tests may look different at different schools". The national tests are thus embedded in a decentralised school system, requiring local decisions to be made and local responsibility to be taken. As Bachman and Palmer (2010) emphasise, the context of a test is complex: "Not only may differing stakeholder groups have different values, but in many contexts assessments are subject to a variety of different laws and regulations. These often operate at different levels (e.g., school, district, state, nation), and are sometimes in conflict with each other and with societal or educational values" (p. 257). An important question raised in the current study is therefore how far a centrally designed paired speaking test can be standardised in terms of uniform administration procedures when carried out in a decentralised school setting.

Furthermore, the results in Study III indicated that many teachers perceived that they did not receive enough support from the school management in the organisation of the oral tests, and it was also suggested that this may be a bigger problem at smaller schools, where the implementation of the speaking tests was perhaps left to the individual teacher to a greater extent. The organisation of the oral tests was a stressful factor for many teachers, not least because many had to conduct their regular English lessons and at the same time administer the tests. Clearer routines and administrative support are needed to make sure test administration is carried out under stable circumstances for both teachers and students.

It should also be kept in mind that "[t]here are enormous practical constraints on the large-scale testing of spoken language proficiency. These include the administrative costs and difficulties and the sheer logistics of testing large numbers of candidates either individually or in very small groups" (Weir, 2005, p. 191). As a result, in many European countries, speaking skills are either not tested or they are internally developed and administered at schools (European Commission/EACEA/Eurydice, 2015). As pointed out by Roever (2004), we need to consider the consequences of not administering the test, as this would very likely result in a narrowing of the construct, as well as a corresponding limitation to the generalizability of test scores. Furthermore, teachers in general express very positive attitudes towards the national EFL speaking test in relation to other validity aspects than practicality, as shown for example in the annual questionnaires conducted by the test constructors. In a validity argument, it is therefore important to consider both the advantages, in this case for example in terms of construct representativeness, and the disadvantages, such as practicality, and present suggestions on improvements that will help improve validity claims. In Study III, some critical aspects of the setting and administration of the national EFL speaking tests were highlighted and discussed, and suggestions that may strengthen validity claims were made.

In short, the validity evidence collected to highlight the issue of administration and setting indicates that the decentralised responsibility for the implementation of the oral national tests has implications for standardisation; however, the effects this may have on test scores are not known and therefore require further analyses.

Cognitive validity

Cognitive validity, or theory-based validity, is concerned with the cognitive operations involved in accessing executive resources (Weir, 2005). According to Field (2011), "[a]n important consideration in establishing the validity of tests that aim to measure performance is the extent to which the task, test content and prevailing conditions require the taker to replicate the cognitive processes which would prevail in a natural (i.e. non-test) context" (p. 66). Establishing cognitive validity in speaking can thus be achieved by evaluating the activation of executive resources and processes prompted by the task. In Field (2011), this was done by applying Levelt's (1989, 1999) model of speech production to the test in question. This was not possible in the current investigation, as the evidence collected is limited to data collected from raters' verbal protocols. In this discussion, the focus will therefore be on test-takers' *strategic competence* (Bachman & Palmer, 1996) and *interactional competence* (Young, 2011), which was possible to observe through raters' comments on strategy use.

Study I and Study II allowed for the collection of validity evidence regarding test-takers' various communication and production strategies. In Study I, two of the coding categories used in the analyses referred specifically to strategy use: *interactional strategies* and *production strategies*. In Study II, an investigation into raters' perceptions of the construct of interactional competence was made and the interactional strategies noted by raters were more thoroughly described and analysed.

Starting with production strategies, raters noted two types of strategy use: self-monitoring strategies (*monitoring and repair*), which was the largest group, and compensatory strategies (*compensating*). Monitoring and repair referred to raters' comments on test-takers' ability to monitor their own speech and backtrack and correct slips and errors. Compensating pertained to comments on students' ability to use circumlocution and paraphrase to compensate for gaps in vocabulary and structure. The two categories applied to the ability to use strategies in relation to both own and partner's speech.

Furthermore, in Study II, three main categories of interactional strategies were evident: topic development moves, turn-taking management and interactive listening strategies. These were seen as contributing to interaction in a positive way when used mutually and reciprocally by test-takers, contributing to what has been described as co-constructed *spoken fluency* (McCarthy, 2010) or *fluency between people* (Jacoby & Ochs, 1995). Another characteristic of the cognitive processes activated in the test format, as illustrated in the category *interactive listening strategies*, is the overlap of receptive and productive skills when test-takers switch between the role of listener and speaker. This is obviously cognitively demanding and was therefore seen as a skill requiring high interactional proficiency. It was indicated that the development of both *turn-taking skills* and *interactive listening skills* was related to test-takers' increased efficiency in simultaneously monitoring their partner's speech and constructing a response.

In summary, the investigation of raters' interpretation of the speaking construct thus revealed that the test format allows for a range of cognitive processes to be activated by test-takers and observed by raters, including both monitoring functions as well as co-construction of meaning in a social context. In addition, the interactional strategies used by test-takers seem to be representative of communication in L1 speech, i.e. speakers' first language (see, e.g., Zhu, Cai, Fan, Chan, & Cheong, 2017), strengthening the claims made with regard to the authenticity and representativeness of the test content. Field (2011) notes that "[t]he goal is to establish whether the tasks proposed by a test designer elicit mental processes resembling those which a language user would actually employ when undertaking similar tasks in the world beyond the test. (p. 67). Thus, in answer to Weir's question concerning cognitive validity; "Are candidates likely to use the same cognitive processes as they would in a 'real world' context?", the validity evidence presented in Study I and II, although limited in scope, indicate that this is likely the case.

Scoring validity

Evidence regarding scoring validity was gathered mainly in Study I and III. In Study I, the results from the descriptive statistics of the Swedish raters' scores showed certain degrees of variability, which is not surprising considering the fact that the test is performance-based (McNamara, 1996). In addition, as is commonly the case in performance-based assessment, rater profiles with differences in rater severity/leniency were found. It was also indicated that raters had more difficulty agreeing on the scoring of some performances than others, which could tentatively be explained by two factors: 1) raters noticed similar features of the performance but evaluated them differently, e.g. one rater evaluated a feature positively and the other rater viewed the same feature negatively, or 2) raters focused on partly different features of the student's performance, e.g. one rater focused more on grammatical accuracy while another focused on fluency. Similar findings have been made in previous rater cognition studies in the context of paired speaking tests (Orr, 2002).

With regard to inter-rater agreement, the correlation coefficients in Study I pointed to acceptable inter-rater reliabilities, albeit with clear room for improvement. It was generally found that raters agreed on the ranking of performances, although rater severity varied. In addition, the results are in line with previous rater studies of group oral assessment (Bonk & Ockey, 2003; Shohamy, Reves, & Bejarano, 1986; Van Moere, 2006), where it has been found that inter-rater agreement is generally somewhat lower than for the individual interview format. This is most likely related to the greater degree of variability present in the paired and group format. However, with regard to the Swedish national speaking test of English, a study by Erickson (2009) points out that correlations in the reference groups that assess speaking tests during the test development process are generally very high.

Furthermore, in Study III, it was found that raters perceived that they had good support for rating from the assessment materials available in conjunction with the national test. The analytic assessment factors, as well as the benchmarked sample performances, were most favourably perceived, pointing to the positive impact of such assessment materials. In Study III, it was also indicated that it was common for teachers to assess the oral national EFL tests on their own, even though co-rating is strongly recommended in the test instructions as a measure to increase inter-rater reliability. Considering this, it is still encouraging to see that as many as 55% of the teachers in the current study reported that they co-rated some, many or all of the performances. Whereas many teachers expressed positive attitudes towards co-rating, lack of time and heavy workload were the main reasons for this not taking place. Co-rating is usually done by two teachers sitting in on the test together, or by sharing recordings, which may be more time-consuming than the co-rating of the written essay-based subtest. The fact that recording is not mandatory and practices thus vary complicates the issue further, even though the tendency seems to be that teachers increasingly use recordings – about 70% of teachers reported recording the speaking test in a recent report (National Assessment Project, 2017).

To sum up, in relation to scoring validity, Weir (2005) asks to what extent we can depend on the scores on the test. The validity evidence presented here suggests that the potential variability of teacher ratings as well as the formal organisation of co-rating are issues in need of further attention in order to further strengthen validity claims. In an investigation of the Swedish school system carried out by the OECD (Nusche et al., 2011), it is argued that 'it is vital' to increase the reliability of the teacher-rated national tests. The authors of the OECD report suggest external moderation, teacher training and professional development as possible measures:

> External moderation is essential to ensure consistency, comparability and equity of the teacher-based assessments. There are several options of doing this, such as employing a second grader (a teacher in the same subject) in addition to the students' own teachers, employing professionals for systematic external grading and/or moderation, or introducing a checking procedure by a competent authority or examination board. In any of these options, high quality training for all graders is essential to ensure professional assessment competencies (p. 11).

This should be related to current on-going activities at the national level, where the Swedish government recently proposed *external rating* of national tests, carried out by a teacher other than the student's own, preferably from another school unit, and a form of *co-rating*, whereby two teachers, one of whom holds the main responsibility and is not the student's own teacher, independently mark the test (Swedish Ministry of Education and Research, 2017a). External rating and co-rating are presently being tried out in a pilot project coordinated by the NAE. However, as pointed out in the OECD report, "high quality training for all graders is essential to ensure professional assessment competencies". As a complement to a more formalised organisation of co-rating, it therefore seems motivated to consider investing resources in professional development, including issues of assessment and rating, which works to maintain and support teachers' assessment literacy (Malone, 2017; Xerri & Vella Briffa, 2018).

Criterion-related validity

Weir (2005) described three forms of criterion-related validity: (1) cross-test comparability, (2) comparison with different versions of the same test and (3) comparison with external standard. In Study I, a small-scale comparison of the performance standards in the Swedish educational system, which are related and comparable to the reference levels in the CEFR, and ratings using CEFR scales was made. In other words, this is an example of a comparison with an external standard. Fourteen raters from Finland and Spain (EFL education professionals working at schools/universities and/or ministries), who were familiar with and experienced in using CEFR scales, rated the same twelve performances as the Swedish raters did. The results indicate that the Swedish and CEFR raters, despite coming from different educational contexts, agreed on the ranking of the 12 performances to a large extent, which strengthens scoring validity claims.

The average ratings of the CEFR group also showed that the CEFR raters rated all performances but two between B1+ and C1, which is in line with the intended pass level of the test (for the course English 6), which is B2-. Furthermore, the two performances that had been assigned an average rating below B1+ by the CEFR raters were also seen as problematic by many of the Swedish raters. Some of the Swedish raters had rated these performances as a 'Fail', indicating that they were border-line cases.

Overall, the validity evidence collected in Study I as regards criterion-related validity seems to point to positive results. However, a more large-scale investigation is obviously needed to follow this up. Also, studies investigating cross-test comparability and comparisons with different versions of the same test need to be conducted, in order to explore different kinds of criterion-related validity.

Consequential validity

Consequential validity is concerned with the impact of a test on individuals, institutions and society, and with the use that is made of test results (Weir, 2005). This aspect of validity was mainly investigated in Study III of the present thesis. First of all, it should be noted that more than 95% of teachers have been shown to express positive opinions on the national EFL tests in the annually conducted surveys by the test constructors, both to the principle of national testing as such and to the national assessment materials (Erickson, 2017b); however, concerns about work load have been raised, not least with regard to the oral tests. Against this background, Study III explored teachers' implementation practices and views of practicality in connection with the speaking subtest.

The results indicate that teachers found the practical implementation to be somewhat problematic and quite time-consuming; however, with great variation in answers. On the other hand, teachers in general found the instructions to be clear and easy to follow, although not always possible to adhere to in practice. Furthermore, they perceived that they had good support from the assessment materials. One main implication of the study is that the oral national tests are an important concern for the whole school and the implementation should not be left to be solved by the individual teachers who are conducting the tests. Clearer routines and administrative support are therefore needed. The validity evidence thus clearly points to the fact that the implementation, as organised presently, may have negative effects on stakeholders, in this case teachers, in terms of workload and working conditions. The findings of Study III partly answer the question posed in Weir (2005) regarding consequential validity: "What effect does the test have on its various stakeholders? However, obviously other stakeholder groups need to be included in further studies, including students and school management. It is also worth emphasising that the results pointing to negative consequential validity for teachers should be considered in a larger context, where, as mentioned above, the very large majority of the teachers consider the testing of oral language proficiency within the national tests important and positive (see, e.g., National Assessment Project, 2017).

To sum up this section, Weir's socio-cognitive validation framework proved a useful tool in the analysis of the three empirical studies, enabling a discussion of both strengths and weaknesses of the national EFL speaking tests, as used in the Swedish educational context. Furthermore, applying the socio-cognitive framework to the results of the three empirical studies highlighted the need to take contextual as well as individual, cognitive factors into account in the validation process. It is hoped that the thesis contributes to the field of speaking assessment in two ways: firstly, by showing how a theoretical framework can be used to support the validation process, and secondly, by providing a concrete example of validation of a high-stakes speaking test, highlighting strengths and weaknesses, and providing suggestions for test development.

In the following, and final section, some concluding remarks will be made.

Chapter Eight: Concluding remarks

The overall purpose of the present thesis was to explore different aspects of validity evidence in relation to a paired speaking assessment, as administered in the context of a high-stakes national test at the upper secondary level of the Swedish educational system. Some main conclusions and pedagogical implications based on the results will now be presented.

Firstly, it should be remembered that Sweden does not have a school-leaving exam system with a one-time final assessment format. Final grades are based on national test results in combination with teachers' continuous assessment during the course. However, in light of the fact that the Swedish Government has recently introduced an amendment in the Education Act, which gives more weight to national test results in relation to final grades (Swedish Ministry of Education and Research, 2017b), their high-stakes nature has become more pronounced. This obviously puts higher demands on the system of national assessment in terms of fairness and reliability, and there are ongoing government-initiated activities to address these issues.

In light of the process of change that the national assessment system is going through at the moment, which also includes digitalisation of the national tests, some main strengths and challenges of the national EFL speaking test at the upper secondary level, as illustrated in the present thesis, are worth highlighting. To start with, the thesis indicates that the speaking test format of peer-peer interaction allows for a rich representation of the construct of oral competence in general, and test-takers' interactional competence in particular. This should be considered in relation to the emphasis given to spoken interaction in the foreign language syllabus, as well as teachers' positive views towards the national EFL speaking test in general. As Weir (2005) contends:

> Clearly, if we wish to test spoken interaction, a valid test must include reciprocity conditions. This contrasts with the traditional interview format in which the interviewer asks the questions and the interviewee answers. So if we are interested in the candidate's capacity to take part in spoken interaction, there should be reciprocal exchanges where both interlocutor and candidate have to adjust vocabulary and message and take each other's contributions into account. (p. 72)

However, the complexity involved in rating co-constructed speech, and the variability inherent in the construct, pose challenges for high-stakes testing. Therefore, interlocutor effects have to be considered when results are interpreted, and guidelines for raters as well as for teachers have to be elaborated, including conceptually grounded reasoning as well as commented examples.

Furthermore, it seems likely that the assessment materials and guidelines provided for teachers in connection with the national EFL speaking tests, including for example benchmarked and commented performances, in combination with co-rating, contribute to enhancing teachers' assessment literacy in a positive way. However, as it has been emphasised both nationally and internationally that the inter-rater reliability of the national assessment system needs to be improved, policy makers should consider investing more resources in teacher training and professional development, in order to further strengthen teachers' assessment literacy (Malone, 2017; Xerri & Vella Briffa, 2018). Hughes (2002) rightfully points out:

> The accurate measurement of oral ability is not easy. It takes considerable time and effort, including training, to obtain valid and reliable results. Nevertheless, where a test is high stakes, or backwash is an important consideration, the investment of such time and effort may be considered necessary. (p. 134)

In addition, the oral national tests must be seen as an important concern for the whole school and the implementation should not be left to be solved by individual teachers. Clearer routines and administrative support are needed. In addition, if co-rating is considered an important measure to increase inter-rater reliability, a more formalised organisation of this system is desirable to ensure that student performances are assessed under similar circumstances (Swedish Ministry of Education and Research, 2017a). This could also have a positive effect on teacher training and professional development.

Finally, it needs to be emphasised again that national assessment in Sweden is in a process of considerable change, not least following a decision to digitalise the system within a few years' time. This will undoubtedly affect the assessment of oral language competence in several ways. In this, input from different stakeholders, among which teachers is an important group, seems an essential aspect of the development of valid and quality-assured products and procedures.

Methodological issues

In Study I and II, content analysis was used to explore raters' decision-making processes. Shapiro and Markoff (1997) claim that content analysis itself is only valid and meaningful if the results are related to other measures. In order to cross-validate the findings, a discourse analysis of students' performances would have been a useful source of triangulation. For reasons of time, however, this was not possible to undertake in the present investigation. Nevertheless, in the analyses and presentation of the results, references and comparisons have been made in relation to discourse-based studies, e.g. Galaczi (2008, 2014), thus strengthening the interpretation of the results.

In Study I, 17 Swedish raters' scores on twelve student performances were used as a basis for the inter-rater reliability estimates. Obviously, a larger sample, both of raters and of student performances, is needed for generalizability. However, in relation to previous raters studies, and considering the fact that each rater had to invest one working day to participate, the sample was still deemed adequate for the present purposes. Furthermore, in order to account for features affecting estimates of inter-rater reliability in speaking and writing tests, it is now common to use sophisticated IRT statistical models, for example Multifaceted Rasch (MFR). This is something that is planned for further studies.

Future research

Finally, some suggestions for future research will be made. To start with, the findings of the current study suggest that the complexity involved in assessing asymmetric test-taker interaction is an area that warrants further investigation. It was shown in both May's (2009) and the current study that raters tried to "unravel the extent to which a candidate's interactional style had impacted on his/her partner" (p. 416). The issue of rating co-constructed interaction in paired speaking tests, where an individual test-taker's performance may be affected by the way the discourse is co-constructed with the partner they are interacting with, is, as has been shown in previous research on interlocutor effects (e.g., O'Sullivan, 2002), a crucial issue for the test format.

One way of addressing the context-dependent nature of IC is proposed by Young (2011), who suggests making systematic comparisons of the interactional resources used in testing and corresponding non-testing contexts. If these are found to be similar, a stronger claim in relation to the generalisability of test results can be made. This is also an avenue for future research.

In addition, the present thesis (Study III) found that the number of students included in test groups varied, from two or three to four in some cases. This issue, namely possible effects of two or three test takers in the assessment of oral interaction, needs to be further researched to find out to what extent the number has, or does not have, a significant effect on results. Nakatsuhara (2011) investigated this issue with regard to groups of three and four students. In the Swedish context, it would be valuable to compare the effects of groups of two and three students.

Furthermore, as the intention was to keep the investigation in the present thesis as close to the authentic rating procedures as possible, only audio-recorded paired speaking tests were used. Previous research (Ducasse & Brown, 2009; May, 2011b; Nakatsuhara, 2011) has given strong indications that non-verbal features, such as body language, facial expressions and gaze, are part of the construct of interactional competence, and this was not possible to investigate in the present analyses. This is an avenue for further research in the context of the paired speaking task used in the Swedish national test of English.

Another interesting development of the present thesis would be to focus more specifically on *washback effects*, i.e. intended or unintended effects of a high-stakes test on teaching and learning (L Cheng, Watanabe, & Curtis, 2004; Muñoz & Álvarez, 2010; Taylor, 2005). This is considered an important aspect of consequential validity. There are good reasons to believe that the national EFL speaking test in the Swedish school context has positive washback effects on the teaching and learning of oral skills in the language classroom. However, further empirical research is needed to confirm this.

Finally, it needs to be emphasised that the issues dealt with in the current thesis have a strong relation to the classroom, hence to the practices of teachers in teaching and assessing continuously oral language proficiency. Here as well, aspects of validity in the wide sense, as expressed in the different frameworks used in the thesis, should be considered and concretized/implemented in everyday pedagogical work.

Svensk sammanfattning (Swedish summary)

Inledning

Den kommunikativa språkundervisningens genomslag har lett till en ökning av autentiska och interaktiva muntliga provformat, såsom par- och gruppsamtal. I linje med denna utveckling har även begreppet muntlig språkfärdighet vidgats till att inbegripa både kognitiva och sociala dimensioner av språkanvändning, i samspel med varandra (Galaczi & Taylor, 2018; McNamara & Roever, 2006). Fokus för denna avhandling är den muntliga delen av de nationella proven i engelska i gymnasieskolan. I detta prov prövas muntlig produktion och interaktion, dvs. förmåga att formulera sig och kommunicera på engelska i tal. Provet genomförs i form av ett samtal, företrädesvis mellan två elever, där eleverna får ”uttrycka, utveckla och diskutera ett innehåll på egen hand och i samspel med andra” utifrån ett givet tema.

Att pröva muntlig förmåga i par eller grupp har många fördelar. Par- och gruppsamtal möjliggör till exempel en mer autentisk bedömning av muntlig interaktion än intervjuer med en provdeltagare och en intervjuare/examinator. I parsamtal får provdeltagarna möjlighet att visa upp en större bredd av språkliga funktioner och interaktionsstrategier än i intervjusituationen (Brooks, 2009; O'Sullivan et al., 2002), där det finns en tydlig hierarkisk struktur mellan provdeltagare och intervjuare. Det finns dock svårigheter med par- och gruppsamtal ur bedömningssynpunkt. För det första kan olika bakgrundsvariabler hos provdeltagarna, t.ex. personlighet, kön och språklig nivå, påverka samtalspartnerns prestation (Foot, 1999; Norton, 2005; O'Sullivan, 2002). De studier som har gjorts är dock inte entydiga beträffande om eller hur dessa s.k. *interlocutor effects* påverkar betyget. En ytterligare utmaning vid par- och gruppsamtal är att individuella bedömningar görs av en gemensamt skapad prestation (*co-construction*) (McNamara, 1997). Den variabilitet och oförutsägbarhet som finns inbyggd i par- och gruppinteraktion kan antingen ses som ett hot mot validiteten, en källa till så kallad *konstruktirrelevant varians* (”construct irrelevant variance”) (Messick, 1989a), eller som en ingående del av det *konstukt*, eller den förmåga, som ska bedömas, nämligen muntlig interaktion.

Vilket synsätt man än väljer får det konsekvenser för hur provresultaten och användningen av dessa tolkas och förstås. Med tanke på den komplexitet som provformatet inrymmer finns det ett behov av ytterligare forskning för att belysa dess operationalisering i olika kontexter och ur olika perspektiv.

Bakgrund

De flesta nationella språkprov i Europa, liksom de svenska, är relaterade till *Gemensam europeisk referensram för språk: lärande, undervisning och bedömning* (GERS) (Skolverket, 2009), vilket är ett ramverk för undervisning, lärande och bedömning av språk. I GERS beskrivs sex generella språknivåer: A1, A2, B1, B2, C1 och C2, från nybörjare till avancerad språkanvändare. Ämnet engelska, liksom moderna språk, är inordnat i ett system av språkfärdighetsnivåer, eller steg. Stegen är påbyggbara och möjliga att jämföra med språknivåerna i GERS. Till exempel ska nivån för det lägsta betyget i kurserna Engelska 5 och Engelska 6 i gymnasieskolan, som denna undersökning handlar om, jämföras med ett högt B1 (B1.2) respektive ett lågt B2 (B2.1) i GERS (Swedish National Agency for Education, 2018b).

En undersökning som Europeiska kommissionen (European Commission, 2015) har genomfört angående nationella prov i främmande språk i Europa visar att de fyra färdigheterna (läsa, lyssna, tala och skriva) testas i olika stor utsträckning. Läsförmåga prövades i störst utsträckning medan muntliga förmåga var den färdighet som prövades i minst utsträckning. Detta menade författarna till rapporten berodde på komplexiteten i att pröva muntlig förmåga och de höga kostnaderna som detta innebär ("Highlights Report: Languages in Secondary Education," 2015, p. 2), vilket gör att en del länder väljer att inte ha med ett muntligt delprov i de nationella språkproven medan andra länder väljer en lösning som innebär att de muntliga proven utvecklas på lokal nivå. De muntliga nationella proven i engelska i det svenska skolsystemet, som utvecklas centralt men genomförs och bedöms internt på den lokala skolnivån, är därför intressanta att undersöka.

I de enkäter som årligen genomförs av provkonstruktörerna vid Göteborgs universitet framgår det att lärare generellt är mycket nöjda med de muntliga nationella delproven, speciellt vad gäller deras tydliga koppling till kursplanen och det stöd för betygssättningen som proven ger. Den kritik som framförts gäller framför allt arbetsbelastning.

Syfte

Det övergripande syftet med avhandlingen är att undersöka bedömningen av de muntliga nationella delproven i engelska i gymnasieskolan utifrån olika validitetsaspekter. Mer specifikt undersöks följande tre delområden: (1) bedömningsprocessen, (2) *konstruktet*, det vill säga den underliggande förmåga som provformatet avser fånga och (3) det praktiska genomförandet. Avhandlingen bidrar till tidigare forskning inom muntlig språkbedömning genom att undersöka både sociala, kontextuella aspekter och kognitiva färdigheter och processer som aktiveras i provformatet. Den ansluter därmed till en socio-kognitiv modell för validering av språkprov (O'Sullivan & Weir, 2011; Weir, 2005). Följande övergripande forskningsfrågor har utgjort grunden för analysen:

- Vad kan uppmärksammas vad gäller bedömarvariation och bedömarprofiler? (Studie I)
- Vilka aspekter av provdeltagares muntliga prestation uppmärksammas av bedömare? (Studie I och II)
- Hur genomförs och bedöms de muntliga nationella delproven i engelska på den lokala skolnivån? (Studie III)
- Vad anser lärarna om det praktiska genomförandet? (Studie III)

Kontextuell bakgrund

Huvudsyfte med nationella prov i det svenska skolsystemet är att vara ett tydligt och starkt stöd för lärare inför betygssättningen, och därmed bidra till en likvärdig och rättvis bedömning och betygssättning. För att skapa goda förutsättningar för detta rekommenderar Skolverket att man kan arbeta med sam- och medbedömning, vilket innebär att ”lärare tillsammans diskuterar och bedömer elevprestationer utifrån bedömningsanvisningarna”.

Efter att det målstyrda betygssystemet infördes 1994 visade det sig snart att det svenska skolsystemet led av problem, både vad gäller betygsinflation (Cliffordson, 2004) och konsekventa skillnader mellan lärares betyg och elevernas resultat på nationella provet (Swedish National Agency for Education, 2007). Det nationella provsystemet har under de senaste åren därför utsatts för granskning både nationellt och internationellt (Nusche et al., 2011; OECD, 2015). De omrättningar av nationella prov som Skolinspektionen genomfört visar på skillnader mellan de ursprungliga lärarbedömningarna och

Skolinspektionens bedömningar för delproven med mer omfattande elevproduktion (Swedish Schools Inspectorate, 2010, 2011, 2012, 2013, 2015, 2016, 2017). De muntliga delarna har dock inte undersökts, eftersom inspelning är frivilligt och det därför inte finns något underlag att samla in. Detta visar på ytterligare behov av undersökningar av bedömareffekter i muntliga prov.

Som en följd av den kritik som riktats mot de nationella proven har en statlig utredning genomförts (Swedish Ministry of Education and Research, 2016). Baserat på denna har regeringen föreslagit eller vidtagit åtgärder för att öka likvärdigheten i bedömningen av de nationella proven, t.ex. digitalisering av proven, avidentifiering av elevsvar, och en försöksverksamhet med extern bedömning av lärare som inte är elevens undervisande lärare (Swedish Ministry of Education and Research, 2017b).

Teoretisk inramning

Definitionen av begreppet validitet har genomgått en genomgripande förändring sedan mitten på 1900-talet. Tidigare var det vanligt att dela in validitet i tre olika typer: innehållsvaliditet (i vilken utsträckning provinnehållet är representativt för den förmåga man har för avsikt att mäta), kriterierelaterad (i vilken utsträckning provet överensstämmer med andra prov som avser mäta samma sak, eller med kriterier i verkliga livet, t.ex. studieframgång) och begreppsvaliditet (i vilken utsträckning provet mäter det begrepp eller den förmåga som det utger sig för att mäta) (Cronbach & Meehl, 1955). Begreppsvaliditet uppfattades ofta inbegripa de andra två typerna av validitet. Under denna period sågs validitet som en egenskap hos själva provet. Messick (1989) förändrade uppfattningen om validitet genom att argumentera för en samlad syn; Han framhöll att validitet är mångfacetterat, vilket innebär att både empiriska belägg och teoretiska resonemang är viktiga delar i en valideringsprocess och kan bidra i olika omfattning. I Messicks beskrivningen av validitet ingår två olika aspekter. Den första aspekten har att göra med provets resultat där tolkningen av provresultatets nytta och relevans diskuteras. Den andra aspekten har att gör med hur provets användning motiveras utifrån värderings- och konsekvensaspekter. Messick flyttade alltså fokus från själva provet till provresultatets tolkning och användning.

Inom språkbedömning finns ett flertal teoretiska ramverk för validering. I denna avhandling har en socio-kognitiv validringsmodell, utvecklad av Weir (2005) och av O'Sullivan and Weir (2011), använts för att diskutera resultaten.

Det socio-kognitiva perspektivet ansluter sig till Messicks beskrivning av validitet som ett samlat begrepp och framhåller att valideringsprocessen innebär att bevis bör samlas från olika perspektiv och med hjälp av olika forskningsmetoder. Ramverket är socio-kognitivt eftersom det tar hänsyn till både den sociala kontext som provet genomför inom och de kognitiva förmågor och processer som aktiveras i provet. De främsta aspekterna av validitet som bör undersökas i valideringsprocessen, enligt Weirs (2005) modell, är följande, komplementära komponenter: kontextsvaliditet (*context validity*), kognitiv validitet (*cognitive validity*), bedömningsvaliditet (*scoring validity*), kriterierelaterad validitet (*criterion-related validity*) och konsekvensvaliditet (*consequential validity*).

Material och metod

Tre empiriska studier genomfördes för att samla in belägg som kunde användas i valideringen av det muntliga nationella delprovet i engelska i gymnasiet.

Deltagare

I Studie I och II bygger analyserna på ett material av sex ljudinspelade elevsamtal som kommer från en utprövning av det muntliga delprovet i kursen Engelska 6 i gymnasiet. Provdeltagarna var gymnasieelever från olika skolor i Sverige, en flicka och en pojke i varje samtal, för att underlätta för bedömarna att skilja dem åt. Samtalen hade valts ut för att spegla olika färdighetsnivåer. Inga andra bakgrundsvariabler samlades in.

Bedömarna valdes ut genom bekvämlighetsurval. Den första gruppen bestod av 17 gymnasielärare i engelska från elva olika skolor, både kommunala och fristående skolor, i två olika län i Sverige. Av deltagarna var fyra män och 17 kvinnor. Undervisningserfarenheten varierade från 1-29 år (*M*=12 år). Fyra av deltagarna hade kortare undervisningserfarenhet (< 5 år) och övriga hade arbetat mer än 5 år (6-29 år).

Den andra gruppen bestod av 14 internationellt rekryterade bedömare från två europeiska länder, Finland (n = 7) och Spanien (n = 7). De arbetade på skolor/universitet och/eller skolmyndigheter och hade erfarenhet av att använda GERS-skalor i bedömningssammanhang. Det metodologiska valet att inkludera 'externa' bedömare motiverades av att detta möjliggjorde en småskalig, tentativ validering/jämförelse av kunskapskraven i kursen Engelska 6 med GERS referensnivåer. Arbetet med att sammanlänka de svenska kursplanerna i främmande språk med GERS språknivåer har nämligen framförallt gjorts genom textuella jämförelser, och det finns följaktligen ett behov av empiriska undersökningar.

I Studie III medverkade 267 engelsklärare från olika gymnasieskolor i Sverige, som svarade på en webbenkät som bland annat undersökte det praktiska genomförandet av det muntliga delprovet. Av dessa var 75 % kvinnor. Medelåldern var 47 år, med en spännvidd på 26 till 68 år. Deltagarnas undervisningserfarenhet var i genomsnitt 16 år (spännvidd 1-42 år, *SD*=10 år). Nittiosex procent av lärarna i urvalet uppgav att de hade lärarlegitimation som inkluderade engelska.

Bedömningsskalor

I Studie I och II bedömdes samma tolv elevprestationer av både de svenska och GERS-bedömarna, med hjälp av olika bedömningsskalor. De svenska bedömarna använde kunskapskraven i kursen Engelska 6 för muntlig produktion och interaktion (se Appendix 2 i Studie I). De hade även tillgång till de analytiska bedömningsfaktorer som bifogas i lärarinstruktionerna till de nationella proven och som anger olika kvaliteter av muntlig språkförmåga, uppdelat i två områden: innehåll och språk och uttrycksförmåga (se Appendix 3 i Studie I). GERS-bedömarna använde två kompletterande skalor, en holistisk och en analytisk från *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment – A Manual* (Council of Europe, 2009, s. 184-186). Skalorna sträcker sig över alla referensnivåerna i GERS, inklusive de s.k. plus-nivåerna. Den analytiska skalan inkluderar fem aspekter: omfång (range), korrekthet (accuracy), interaktion (interaction) koherens (coherence), och flyt (fluency). Båda bedömargrupperna satte alltså holistiska betyg men var hjälpta av analytiska kriterier och deskriptorer i bedömningsprocessen.

Datainsamling

Data till Studie I och II samlades in under endagsseminarier som hölls med de olika bedömargrupperna i juni, september och november 2013. Strukturen på seminarierna var identiska. De inleddes med en introduktion med information om studien och ett kort träningstillfälle. Sedan lyssnade bedömarna med hörlurar på de sex samtalen och bedömde dem individuellt. Bedömarna ombads att skriva ett sammanfattande omdöme för varje elevprestation där framträdande aspekter betonades.

I Studie III skickades en webbenkät ut under vårterminen 2017. Ett slumpmässigt urval av 150 skolor gjordes från en databas som innehåller Sveriges samtliga gymnasieskolor (vuxenutbildning ingick inte i urvalet). Listan är sammanställd av Statistiska Centralbyrån och publiceras på Skolverkets hemsida. Information om enkäten skickades ut till administrativ personal och rektorer på de 150 skolorna med en förfrågan om att enkäten skulle vidarebefordras till engelsklärare på skolan. 267 svar kom sammanlagt in. Svarsfrekvensen på skolnivå var god (79 %). Det uppnådda urvalet var också representativt i förhållande till geografisk spridning och fördelning av kommunala och fristående skolor. I Studie III undersöktes ett urval av frågor från enkäten som handlar om det praktiska genomförandet av provet.

Analysmetoder

I Studie I analyserades de betyg som bedömarna hade satt på de tolv elevprestationerna med hjälp av deskriptiv statistik. Vidare undersöktes bedömarsamstämmighet i den svenska bedömargruppen genom rangkorrelationskoefficienter: Spearman's rho och Kendall's Tau. Den interna konsistensen i gruppen beräknades med hjälp av Cronbach's alpha. SPSS Statistics, Version 21.0 (IBM Corp., 2012) användes för de statistiska analyserna.

I Studie I och II analyserades de skriftliga kommentarer med hjälp av framför allt kvalitativ innehållsanalys (Galaczi, 2013; Green, 1998; Krippendorff, 2013). Materialet segmenterades och kodades. Resultatet illustrerades även kvantitativt för att visa på förekomst av de olika kodningskategorierna. Processen för att arbeta fram kodningsscheman är detaljerat beskriven i studierna. I båda fallen utvecklades kodningsschemat främst genom ett deduktivt angreppssätt som bygger på tidigare forskning och teori (Galaczi, 2013; Hsieh & Shannon, 2005). Mjukvaran NVivo 11 användes för att organisera och analysera materialet. För att validera kodningen dubbelkodades en viss del av materialet, 10% i studie I och 45% i Studie II. I Studie I anlitades en medkodare med lång erfarenhet av provutveckling och goda kunskaper om GERS, och i Studie II två medkodare med doktorsexamen i tillämpad språkvetenskap. Kodningsschemat utvecklades i en iterativ process tills en bedömarsamstämmighet på >80% hade nåtts. Därefter kodades resten av materialet självständigt av forskaren.

Slutligen bestod Studie III av enkätdata med både slutna och öppna frågor. Beskrivande statistik och sambandsanalys användes för att analysera de kvantitativa svaren. För att belysa och förklara resultaten användes exempel från de öppna kommentarerna. Tre bakgrundsvariabler inkluderades för att undersöka en möjlig relation till lärarnas svar i enkäten: (1) kön, (2) undervisningserfarenhet i år och (3) storleken på skolan som läraren arbetade på (två variabler: antal elever på skolan och antal engelsklärarkollegor). Alla bakgrundsvariabler var självrapporterade. SPSS, Version 25 (IBM Corp., 2017) användes för att analysera enkätsvaren.

Resultat av de tre studierna

Studie I

Huvudsyftet med den första studien var att undersöka bedömarsamstämmighet och bedömarnas beslutsprocesser i relation till det muntliga nationella delprovet i engelska i gymnasieskolan. Ett småskaligt försök gjordes även att utforska sambandet mellan betyg och bedömarnas skriftliga motiveringar. Dessutom var ett sekundärt syfte att göra en tentativ, empirisk jämförelse av de svenska, nationella kunskapskraven och referensnivåerna i GERS.

Analyserna av de 17 svenska bedömarnas betyg för de tolv elevprestationerna visade på viss variabilitet. Vidare fanns det tydliga bedömarprofiler med olika grad av stränghet. Till exempel varierade medelbetygen för bedömarna mellan 5,6 och 8,0 på den tiogradiga skalan. Det framkom också av standardavvikelserna att vissa elevprestationer var mer svårbedömda än andra och därmed hade större variabilitet. Medianen för de parvisa korrelationerna mellan bedömarna låg på på .77 med Spearman's rho och .66 med Kendall's tau, vilket kan ses som relativt god samstämmighet, dock med en uppenbar förbättringspotential. Cronbach's alpha, som mäter den interna konsistensen i gruppen, var hög: .98.

Resultaten visade också att de europeiska bedömarna i genomsnitt bedömde elevprestationerna på den nivå i GERS som provet avser mäta. Medelvärdena för de europeiska bedömarna låg mellan B1+ och C1 med något enstaka undantag. Dessutom jämfördes den svenska och europeiska gruppens rankning av elevprestationer och resultaten visar på stora likheter.

Innehållsanalysen av bedömarnas skriftliga kommentarer visade att de tog hänsyn till en stor bredd av provdeltagarnas kommunikativa kompetens i sin holistiska bedömning. De lingvistiska och pragmatiska aspekterna, samt elevernas interaktionsstrategier var mest framträdande. Det var dock få kommentarer som handlade om förmågan att anpassa språket efter olika sociokulturella förhållanden och sociala konventioner, t.ex. artighetsregler, vilket benämns som sociolingvistisk kompetens i GERS.

Bedömarna höll sig väl till bedömningskriterierna, och kommenterade andra aspekter i relativt liten utsträckning, vilket motsäger resultat i tidigare bedömarstudier av muntliga prov (A. Brown, 2007; May, 2006; Orr, 2002), där det visat sig att bedömare tar hänsyn till sådant som inte specifikt beskrivs i bedömningskriterierna. Det fanns även en viss skillnad i bedömarprofiler mellan de svenska och europeiska bedömarna, med en mer jämn fördelning av kategorierna hos de europeiska bedömarna jämfört med de svenska som hade en stor andel kommentarer om de lingvistiska aspekterna.

Bedömarna reflekterade även över olika aspekter som har med *co-construction* i samtalet att göra, till exempel hur elevens prestation påverkades av den andra partnern. De gjorde också jämförelser mellan eleverna i paret, angående likheter och skillnader, språklig nivå och hur väl interaktionen mellan eleverna fungerade. Liknande iakttagelser har gjorts i tidigare studier av muntliga prov med parsamtal (May, 2011b; Orr, 2002). En tentativ jämförelse mellan bedömarnas kommentarer och betyg visade också att bedömare som satte lågt respektive högt betyg på samma elevprestation antingen uppmärksammade samma aspekter men värderade dem olika (alltså som positivt eller negativt), eller uppmärksammade delvis olika aspekter.

Studie II

Studie II bygger på samma material och urval av bedömare som i Studie I. Huvudsyftet var att undersöka hur bedömare tolkar och uppfattar provdeltagarnas interaktionella kompetens, det vill säga deras förmåga att delta i och bidra till ett samtal på ett meningsfullt sätt. I definitionen av interaktionell kompetens är två aspekter framträdande: för det första skapas social interaktion gemensamt (*co-construction*) mellan individer och för det andra är den kontextberoende, det vill säga den interaktionella kompetensen varierar med den kontext eller praktik som interaktionen genomförs i (A. W. He & Young, 1998; Young, 2000). Dessa två karakteristika utgör en svårighet vid bedömning, vilket gör att det finns ett behov av studier som undersöker konstruktet interaktionell kompetens mer ingående i olika kontexter.

Resultaten visade att bedömarna tog hänsyn till tre huvudkategorier av interaktionell kompetens: (1) strategier för att utveckla samtalets innehåll (*topic development moves*), (2) turtagningsstrategier (*turn-taking management*) och (3) interaktiva lyssnandestrategier (*interactive listening strategies*). Bedömarna upplevde det som positivt när dessa strategier användes på ett kollaborativt och reciprokt sätt, vilket bidrog till att samtalet utvecklades tillsammans av eleverna. Det uppfattades mer negativt om eleverna saknade strategier för att föra samtalet framåt, genom att till exempel inte bygga vidare på eller bekräfta det partnern sa. Dessa tre huvudkategorier sammanfaller väl med tidigare forskning av parsamtal, både bedömarstudier (Ducasse & Brown, 2009; May, 2011b), och konversationsanalys av provsamtal (Galaczi, 2008, 2014).

Resultaten visade också att bedömarna tog hänsyn till effekten av provdeltagranas interaktionella roller (*interactional roles*) i bedömningen. Detta var tydligt i ett samtal med ett så kallat asymmetriskt interaktionsmönster, då en eleverna hade en mer dominerande roll och den andra var mer passiv och därmed fick mindre talutrymme i samtalet. Bedömarna var inte ense om huruvida den mer passiva talaren hjälptes eller påverkades negativt av den mer dominanta talarens roll, och hur detta i sin tur påverkade betyget, vilket visar på komplexiteten i att bedöma samtal som bygger på *co-construction*. Bedömarna lade även märke till hur eleverna presterade i *jämförelse* med, eller i *relation* till varandra, vilket belyser den inter-subjektiva dimensionen av konstruktet.

En jämförelse gjordes även mellan de svenska och GERS-bedömarnas kommentarer, vilken visade på vissa skillnader. Till exempel kommenterade de svenska bedömarna mer frekvent elevernas strategier för att utveckla samtalets innehåll, medan GERS-bedömarna lade mer fokus på turtagningsstrategier. Detta kunde relateras till de olika formuleringarna, så kallade deskriptorer, i bedömningskriterierna. I kunskapskraven och bedömningsfaktorerna i den svenska kontexten betonas vikten av att utveckla ett innehåll, både på egen hand och tillsammans med andra, medan GERS-skalorna mer ingående beskriver turtagningsstrategier.

Studie III

Studie III är en enkätbaserad studie som hade syftet att undersöka lärares synpunkter på det praktiska genomförandet av det muntliga nationella delprovet i engelska. De nationella proven utvecklas centralt, men den praktiska implementeringen av de muntliga proven är delegerad till rektor som har ansvar för att med sin personal organisera genomförandet så att det gagnar elever och lärare på bästa sätt. Därför framhåller Skolverket att den lämpligaste organisationen kan se olika ut på olika skolor och att de muntliga proven ska ses som hela skolans angelägenhet. Mot bakgrund av detta genomfördes Studie III.

Resultaten visade att det fanns en variation i hur lärarna genomförde och bedömde de muntliga nationella delproven, vilket har uppenbara konsekvenser för möjligheten till standardisering, men som är i enlighet med politiska direktiv om decentraliserat ansvar. Det framkom även en del utmaningar med provet i lärarsvaren. Till exempel var det vanligast att genomföra proven på lektionstid (61%) jämfört med utanför lektionstid. Många lärare påpekade att det var stressigt och tidskrävande att genomföra proven under ordinarie engelsklektioner, och de var bekymrade över att detta tog tid från undervisningen. Lärare som arbetade på skolor där proven organiserades mer centralt, som en schemabrytande aktivitet, verkade vara mer nöjda med den lösningen. Inspelning av de muntliga proven rekommenderas starkt i instruktionerna, bland annat eftersom det möjliggör sam- och medbedömning och att man kan lyssna igen på samtalen. Resultaten visade att närmare hälften av lärarna i urvalet spelade in de muntliga delproven, medan ungefär 40% inte spelade in alls. Den huvudsakliga anledningen till att inte spela in var brist på tid, både vad gäller att lyssna igenom samtal igen efteråt och att sam- eller medbedöma med kollegor. Ett ytterligare exempel på skillnader i genomförandet är antalet elever per grupp. Enkätsvaren visar att det var vanligast att gruppera eleverna i par, men grupper om tre, och ibland fyra elever var också vanligt.

Vad gäller bedömning av samtalen i det muntliga provet, framkom det att lärarna i urvalet generellt tyckte att bedömningsmaterialet som finns att tillgå utgjorde ett bra stöd. De analytiska bedömningsfaktorerna och de inspelade och kommenterade exempelsamtalen gav mest stöd, ansåg lärarna, medan kunskapskraven för muntlig produktion och interaktion skattades något lägre. Det visade sig också vara vanligast att bedöma de muntliga proven ensam utan sam- eller medbedömning (42%), även om det också fanns ett relativt stort antal lärare som bedömde alla (13%), många (6%) eller några (36%) av samtalen i samarbete med kollegor. Lärarna i urvalet var generellt positivt inställda till sam-och medbedömning, men brist på tid och en stor arbetsbelastning under perioden med nationella prov gjorde att detta inte hanns med.

Trots direktiven att rektor ska planera genomförandet av de muntliga nationella delproven tillsammans med sin personal, framkom det i enkätsvaren att en majoritet av lärarna ansåg att de inte fick tillräckligt med stöd från skolledningen (62%). Många lärare uppmärksammade behovet av mer administrativt stöd för att kunna genomföra de muntliga proven på ett optimalt sätt. De framgick även att hälften av lärarna som deltog ansåg att det inte fanns tillräckligt med lokaler och grupprum på skolan för att genomföra de muntliga proven, vilket var stressande för många.

Slutligen visade enkäten att lärarna generellt ansåg att provet var relativt tidskrävande och i viss mån problematiskt att genomföra, men att lärarinstruktionerna var lätta att förstå och följa, även om det inte alltid var praktiskt möjligt. De statistiska sambandsanalyser som gjordes mellan lärarnas svar och bakgrundsvariablerna visade att storleken på skolan verkade ha ett visst samband med variationen i enkätsvaren. Det framkom att de muntliga delproven upplevdes som något mer problematiska och tidskrävande på mindre skolor än på större, vilket kan betyda att lärare på mindre skolor ansvarar för att genomföra de muntliga proven på egen hand i större utsträckning än på större skolor.

Diskussion

Det socio-kognitiva ramverket för validering av språkprov (Weir, 2005) användes för att strukturera analysen och diskussionen av resultaten.

Kontextvaliditet

Kontextvaliditet berör både lingvistiska och innehållsliga krav som provformatet ställer på provdeltagarna och de kontextuella ramarna för genomförandet av provet. Resultaten från Studie I, som undersökte aspekter av elevprestationerna som var framträdande för bedömarna, indikerade att provformatet gör det möjligt att bedöma en stor bredd av elevernas muntliga kommunikativa kompetens (Bachman & Palmer, 1996; Canale & Swain, 1980; Hymes, 1972). Provformatet verkar således möjliggöra en bred innehållslig representation av konstruktet 'muntlig förmåga', om man ser till hur det operationaliseras i det muntliga nationella delprovet i den svenska skolkontexten. Detta stöds av tidigare forskning som har visat att par- och gruppinteraktion kan framkalla ett större spektrum av språkliga funktioner, speciellt interaktionella språkfunktioner, såsom att *jämföra*, *beskriva*, *föreslå* och *utveckla*, än intervjuformatet som genomförs med en elev och en examinator (ffrench, 2003; O'Sullivan et al., 2002). Även om bedömarnas sammanlagda kommentarer visar på en allsidig uppfattning och beskrivning av elevernas kommunikativa kompetens skönjdes en viss tendens att uppmärksamma lingvistiska aspekter (i form av både språklig korrekthet och bredd) i särskilt hög grad, vilket bekräftas av tidigare forskning som visar att det är vanligt att bedömare av muntliga prov lägger stor vikt vid grammatisk korrekthet, vilket är en del av den lingvistiska kompetensen (Iwashita et al., 2008). En slutsats av detta är att det är viktigt att lärare ges möjlighet att regelbundet delta i sam- och medbedömning, där elevprestationer diskuteras i förhållande till betygskriterier och bedömningsmaterial, för att på så sätt utveckla en gemensam förståelse av kunskapskraven (Daly et al., 2011). Detta kan möjligen också minska risken för att vissa kriterier viktas mer än andra.

Resultaten tyder även på att beskrivningen av interaktionell kompetens i bedömningsskalor kan utvecklas. Bedömarnas kommentarer i Studie II visade att de hade en bredare syn på konstruktet än vad som avspeglades i bedömningsskalorna, vilket även har framkommit i tidigare bedömarstudier av parsamtal (May, 2011b; Orr, 2002). Med tanke på detta är det önskvärt att bedömningsskalorna/betygskriterierna för muntlig interaktion utvecklas ytterligare så att aspekter av *co-construction* framgår tydligare. Det är även önskvärt att utvecklingen av interaktionell kompetens på olika språkliga nivåer tydliggörs.

Både Studie I och II fann att bedömarna reflekterade över elevernas interaktionella roller och hur detta påverkade prestationerna, speciellt i ett fall av asymmetrisk interaktion då en provdeltagare hade en mer dominant roll i samtalet och tog över, medan partner var mer passiv och inte fick så mycket talutrymme. Bedömarna var inte helt överens om eller hur detta skulle inverka på betyget (cf. May, 2009). Att bedömare har svårigheter att bedöma asymmetrisk interaktion har även framkommit i tidigare studier (May, 2009). Resultaten visar därför på ytterligare behov av forskning för att undersöka *om* och *hur* provdeltagares interaktionella roller påverkar bedömningen. I läraranvisningarna till provet framhålls det att läraren ska påpeka "för eleverna att de ska hjälpas åt att hålla igång samtalet" och "uppmuntra eleverna att ge varandra ungefär lika mycket utrymme". Det betonas även att det "är viktigt att eleverna ges tillfälle att visa vad de kan". Trots att det alltså redan finns vissa riktlinjer tyder resultaten på att det behövs mer explicit information och råd kring interaktionella roller och bakgrundsvariablers påverkan (*interlocutor effects*) på interaktionen.

Inom analysen av begreppet kontextvaliditet ingår även att undersöka provets praktiska genomförande. Resultaten från Studie III visade att det fanns en variation i hur de muntliga proven genomfördes på den lokala skolnivån. Detta är i linje med det decentraliserade ansvaret för de muntliga nationella proven, men har uppenbara konsekvenser för den standardisering som är önskvärd vid *high-stakes-prov*. Den effekt detta eventuellt får på elevresultat kräver därför vidare analys och forskning. En fråga som lyfts är också hur långt det är möjligt att standardisera genomförandet av ett muntligt språkprov som genomförs i en lokal skolkontext. Trots tydliga direktiv att rektor har yttersta ansvar för genomförandet av de muntliga proven upplevde många lärare att de inte fick tillräckligt med stöd från skolledningen för att organisera de muntliga delproven. Detta visar att det behövs tydligare rutiner och mer administrativt stöd vid genomförandet för att genomförandet ska ses som ”hela skolans angelägenhet”.

Kognitiv validitet

I Studie I och II framkom att provformatet med parinteraktion ger möjlighet för bedömare att uppmärksamma ett flertal kognitiva processer (cognitive validity), i form av provdeltagarnas *strategiska kompetens* (Bachman & Palmer, 1996). I Studie I la bedömarna märke till när eleverna använde så kallade *produktionsstrategier* (Skolverket, 2009), vilket kan handla om förmågan att korrigera felsägningar och misstag eller kompensera för språkliga brister genom att omformulera. I Studie II beskrevs dessutom provdeltagarnas användning av tre huvudkategorier av *interaktionsstrategier* (Skolverket, 2009): (1) strategier för att utveckla samtalets innehåll (*topic development moves*), (2) turtagningsstrategier (*turn-taking management*) och (3) interaktiva lyssnandestrategier (*interactive listening strategies*). Bedömarna upplevde det som positivt när dessa strategier användes på ett kollaborativt och reciprokt sätt, vilket bidrog till att samtalet utvecklades tillsammans av eleverna. För att lyckas med detta krävs en relativt krävande kognitiv process, nämligen att talaren aktivt lyssnar på det partnern säger, samtidigt som han/hon planerar sitt svar.

Det är positivt för tolkningen av provets validitet om de kognitiva processer som eleverna använder i provet är samma eller liknar de kognitiva processer som används i naturlig interaktion utanför provsituationen. Det finns indikationer på att de huvudkategorier av interaktionsstrategier som beskrivs i Studie II också förkommer i förstaspråksanvändning (Zhu et al., 2017), vilket alltså är positivt för validiteten, även om ytterligare studier krävs.

Bedömningsvaliditet

Vad gäller bedömningsvaliditet (scoring validity) visade Studie I att interbedömarreliabiliteten var relativt god bland de 17 svenska bedömarna, speciellt med tanke på att det muntliga provet är ett så kallat *performance-baserat* prov. Det finns dock uppenbart utrymme för förbättring för att nå ännu högre samstämmighet. Studie III antydde dessutom att de muntliga nationella delproven sambedöms i mindre uträckning än de skriftliga uppsatsproven Det faktum att inspelning inte är obligatorisk och därmed varierar gör situationen mer komplex, även om det visat sig att inspelning har ökat starkt under de senaste åren (National Assessment Project, 2017).

I utvärderingar av det svenska skolsystemet som OECD utfört (Nusche et al., 2011) betonas att det är mycket viktigt att öka reliabiliteten för de nationella proven. Författarna föreslår åtgärder såsom medbedömning med en 'second grader', och kompetensutveckling inom bedömning. Det verkar troligt att det bedömningsstöd som tillhandahålls i anslutning till de nationella proven, t.ex. de kommenterade elevprestationer, i kombination med sam- och medbedömning, bidrar på ett positivt sätt till lärares bedömningskompetens. Men med tanke på att det har uppmärksammats både nationellt och internationellt att interbedömarreliabiliteten i det nationella provsystemet bör förbättras, bör mer resurser satsas på kompetensutveckling inom bedömning för att ytterligare stärka lärares bedömningskompetens (Malone, 2017; Xerri & Vella Briffa, 2018). Eftersom Skolverket uppmanar till sam- och medbedömning som ett sätt att öka reliabilitet i de nationella proven, är det även önskvärt att få till stånd en mer systematisk och enhetlig organisation av denna verksamhet. Som det ser ut idag verkar förhållandena skilja sig mellan skolor, vilket påverkar likvärdighet. Dessutom pekar resultaten av denna studie på att sam- och medbedömning förekommer i mindre utsträckning för det muntliga delprovet jämfört med det skriftliga uppsatsprovet, vilket visar på ett extra behov av resurser.

Kriterierelaterad och konsekvensvaliditet

Resultaten i Studie I visar också att de europeiska bedömarna i genomsnitt bedömde elevprestationerna på den nivå i GERS som provet avser mäta, vilket är en form av kriterierelaterad validering. Eftersom detta är en ytterst småskalig jämförelse måste dock resultaten ses som högst tentativa, och det finns följaktligen ett behov av mer storskalig empirisk validering av det muntliga provet gentemot yttre kriterier.

Slutligen undersöktes konsekvensvaliditet (consequential validity) i Studie III. Lärarna uttrycket generellt att provet var relativt tidskrävande och i viss mån problematiskt att genomföra, men att lärarinstruktionerna var lätta att förstå och följa, även om det inte alltid var praktiskt möjligt. Det framkom att de muntliga delproven upplevdes som något mer problematiska och tidskrävande på mindre skolor än på större, vilket kan tyda på att lärare på mindre skolor får ansvara för genomförandet av de muntliga proven på egen hand i större utsträckning än på större skolor.

Dessa resultat bör dock ställas mot det faktum att en stor majoritet av lärare uttrycker positiva åsikter om provets innehåll och format i de årliga enkäter som genomförs av provutvecklarna (se, t.ex. National Assessment Project, 2017). En slutsats är att genomförandet av de muntliga nationella delproven måste bli en gemensam angelägenhet på skolnivå och inte överlämnas till enskilda lärare. Tydligare rutiner och mer administrativt stöd för genomförandet av det muntliga delprovet är önskvärt.

Slutord

Det övergripande syftet med avhandlingen var att utforska olika validitetsaspekter av det muntliga nationella delprovet i engelska för gymnasiet. Några övergripande slutsatser presenteras här, vilka bör ställas i förhållande till de förändringar i det nationella provsystemet som nu pågår för att öka reliabiliteten och likvärdigheten (Swedish Ministry of Education and Research, 2017b).

Om man ser till provets styrkor och svagheter kan två tydliga slutsatser dras. För det första möjliggör det muntliga provet att pröva elevers kommunikativa muntliga förmåga på ett brett och representativt sätt, vilket är positivt. Den tydliga kopplingen mellan det muntliga nationella delprovet och det fokus på muntlig produktion och interaktion som framkommer i ämnes- och kursplaner stärker ytterligare validiteten av provets användande. Svårigheten ligger dock i att bedöma par- och gruppsamtal på ett reliabelt och konsekvent sätt, eftersom provformatet med elevinteraktion medför en viss grad av variabilitet och oförutsägbarhet. Avhandlingen pekar därför på ett behov av kompetensutveckling och bedömarträning vad gäller muntlig språkfärdighet. Organiserade former för sam- och medbedömning behöver också utvecklas för att systemet ska bli mer likvärdigt. En ytterligare svaghet som framkom i valideringen är att det decentraliserade ansvaret för genomförandet av de muntliga proven ibland brister. Det är alltså viktigt att organisationen av de muntliga delproven, i enlighet med Skolverkets anvisningar, blir en angelägenhet för hela skolan och inte lämnas åt enskilda lärare att ta hand om.

Till slut bör nämnas att de aspekter av det muntliga nationella provet som har diskuterats i denna avhandling också har en tydlig koppling till undervisning och bedömning av muntlig språkfärdighet i klassrummet. Förhoppningen är att resultaten kan bidra till att stärka lärares bedömningskompetens inom muntlig språkfärdighet, och visa på behov av fortsatt kompetensutveckling. I detta arbete bör validitetsaspekter, som de som beskrivs i den modell som använts i denna avhandling, även tas hänsyn till och implementeras i daglig praktik.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington DC: American Council on Education.

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly, 2*(1), 1-34. doi:10.1207/s15434311laq0201_1

Bachman, L. F. (2007). 3. What is the Construct? The Dialectic of Abilities and Contexts in Defining Constructs in Language Assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language Testing Reconsidered.* Ottawa: Les Presses de l'Université d'Ottawa/University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

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

Berry, V. (1993). Personality characteristics as a potential source of language test bias. In A. Huhta, K. Sajavaara, & S. Takala (Eds.), *Language testing: New openings* (pp. 115-124). Jyvaskyla, Finland: Institute for Educational research.

Berry, V. (2007). *Personality differences and oral test performance.* Frankfurt: Peter Lang.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110. doi:10.1191/0265532203lt245oa

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341-366. doi:10.1177/0265532209104666

Brooks, L., & Swain, M. (2014). Contextualizing Performances: Comparing Performances During TOEFL iBT TM and Real-Life Academic Speaking Activities. *Language Assessment Quarterly, 11*(4), 353-373. doi:10.1080/15434303.2014.947532

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS Collected Papers: Research in Speaking and Writing Assessment* (pp. 98-141). Cambridge: UCLES/Cambridge University Press.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.

Canagarajah, S. (2006). Changing Communicative Needs, Revised Assessment Objectives: Testing English as an International Language. *Language Assessment Quarterly, 3*(3), 229-242. doi:10.1207/s15434311laq0303_1

Canale, M., & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics, I*(1), 1-47. doi:10.1093/applin/I.1.1

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20*(4), 369-383. doi:10.1191/0265532203lt264oa

Chalhoub-Deville, M., & Deville, C. (2005). A look back and forward to what language testers measure. In E. Hinkel (Ed.), *Handbook of research in second langauge teaching and learning* (pp. 815-832). Mahwah, NJ: Erlbaum.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). New York: Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language.* New York: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an Argument-Based Approach to Validity Make a Difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13. doi:doi:10.1111/j.1745-3992.2009.00165.x

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20*(4), 409-439. doi:10.1191/0265532203lt266oa

Chapelle, C. A., & Voss, E. (2013). Evaluation of Language Tests Through Validation Research. In A. J. Kunnan (Ed.), *The Companion to Language Assessment.*

Cheng, L., & Sun, Y. (2015). Interpreting the Impact of the Ontario Secondary School Literacy Test on Second Language Students Within an Argument-Based Validation Framework. *Language Assessment Quarterly, 12*(1), 50-66. doi:10.1080/15434303.2014.981334

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in Language Testing. research contexts and methods.* London: Lawrence Erlbaum.

Cliffordson, C. (2004). Betygsinflation i de målrelaterade gymnasiebetygen [Grade inflation in the criterion-referenced grades in upper secondary school]. *Pedagogisk forskning i Sverige, 9*(1), 1-14.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). Retrieved from http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp

Council of Europe. (2018). Common European Framework of Reference for Languages: Learning, Teaching, Assesment. Companion Volume with New Descriptors. Retrieved from http://www.coe.int/en/web/common-european-framework-reference-languages

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the Valid Use of Assessments. *Assessment in Education: Principles, Policy &amp; Practice, 3*(3), 265-286. doi:10.1080/0969594960030302

Csépes, I. (2009). *Measuring oral proficiency through paired-task performance*. Frankfurt am Main: Peter Lang.

Daly, A., Billington, L., Chamberlain, S., Meyer, L., Stringer, N., Taylor, M., & Tremain, K. (2011). *Principles of moderation of internal assessment*. Manchester, NH: AQA.

Davies, A. (1997). Introduction: the limits of ethics in language testing. *Language Testing, 14*(3), 235–241.

Davies, A. (2012). Kane, validity and soundness. *Language Testing, 29*(1), 37-42. doi:10.1177/0265532211417213

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396. doi:10.1177/0265532209104667

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. doi:10.1177/0265532215582282

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423-443. doi:10.1177/0265532209104669

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers’ viewpoints on assessment reform. *Language Testing, 32*(1), 101-120. doi:doi:10.1177/0265532214544393

Egyud, G., & Glover, P. (2001). Readers respond. Oral testing in pairs - secondary school perspective. *ELT Journal, 55*(1), 70-76. doi:10.1093/elt/55.1.70

Enright, M., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater(R) scoring. *Language Testing, 27*(3), 317-334. doi:10.1177/0265532210363144

Erickson, G. (1999). Från Sp 8 till Äp 9. Om utvecklingen av ett nytt nationellt prov i engelska i grundskolan. [From Sp 8 to Äp 9. On the development of a new national test of English in comprehensive school]. In *Papers on Language Learning Teaching Assessment. Festskrift till Torsten Lindblad [Papers* *on Language Learning Teaching Assessment. Festschrift to Torsten Lindblad]*. Gothenburg: University of Gothenburg.

Erickson, G. (2006). Bedömning av och för lärande - En kollaborativ ansats i arbetet med nationella prov i språk. [Assessment of and for Learning: A collaborative approach to the development of national language tests] (D. o. Education, Trans.). In U. Tornberg (Ed.), *Mångkulturella aspekter på språkundervisningens kommunikativa praktiker*. (pp. 187-207). Örebro, Sweden: Örebro University.

Erickson, G. (2009). *Nationella prov i engelska - en studie av bedömarsamstämmighet [National tests of English - a study of interrater agreement]*. Stockholm: The Swedish National Agency for Education.

Erickson, G. (2017a). Experiences with Standards and Criteria in Sweden. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education – The Nordic Countries in an International Perspective*. Cham, Switzerland: Springer International Publishing.

Erickson, G. (2017b, Nov.). National assessment of foreign languages in Sweden. Retrieved from https://nafs.gu.se/digitalAssets/1671/1671355_national_assessm_of_foreign_lang_in_sweden2017.pdf

Erickson, G., & Pakula, H.-M. (2017). Den gemensamma europeiska referensramen för språk: Lärande, undervisning, bedömning – ett nordiskt perspektiv [Common European Framework of Reference for Languages: Learning, teaching, assessment - a Nordic perspective]. *Acta Didactica Norge, 11*(3). doi: http://dx.doi.org/10.5617/adno.4789

Erickson, G., & Åberg-Bengtsson, L. (2012). A Collaborative Approach to National Test Development. In D. Tsagari & I. Csépes (Eds.), *Collaboration in Language Testing and Assessment* (pp. 93-108). Frankfurt am Main: Peter Lang.

European Commission. (2015). *Study on comparability of language testing in Europe. Final report September 2015*. Luxembourg: Publications Office of the European Union Retrieved from http://ec.europa.eu/dgs/education_culture/repository/languages/library/documents/edl-report_en.pdf.

European Commission/EACEA/Eurydice. (2009). *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results*. Brussels: Education, Audiovisual and Culture Executive Agency (EACEA P9 Eurydice).

European Commission/EACEA/Eurydice. (2015). *Languages in Secondary Education: An Overview of National Tests in Europe – 2014/15. Eurydice Report* Luxembourg: Publications Office of the European Union.

ffrench, A. (1999). *Study of qualitative differences between CPE individual and paired test formats (Internal UCLES EFL report)*. Cambridge, UK: University of Cambridge Local Examinations Syndicate.

ffrench, A. (2003). The change process at the paper level - Paper 5, Speaking. In C. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913-2002* (pp. 367-471). Cambridge: Cambridge University Press.

Field, J. (2011). Cognitive validity. In M. Milanovic & C. Weir (Eds.), *Examining Speaking. Research and practice in assessing second langauge speaking* (pp. 65-111). Cambridge: Cambridge University Press.

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53*(1), 36-41. doi:10.1093/elt/53.1.36

Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing, 29*(3), 345-369. doi:10.1177/0265532211424479

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing, 13*(1), 23-51. doi:10.1177/026553229601300103

Fulcher, G. (2003). *Testing second language speaking.* London [u.a.]: Longman.

Fulcher, G. (2004). Are Europe's tests being built on an 'unsafe' framework? *Guardian Weekly.*

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book.* London & New York: Routledge.

Fulcher, G., & Davidson, F. (2012). *The Routledge handbook of language testing.* Milton Park, Abingdon, Oxon ; New York: Routledge.

Galaczi, E. D. (2008). Peer-Peer Interaction in a Speaking Test: The Case of the First Certificate in English Examination. *Language Assessment Quarterly, 5*(2), 89-119. doi:10.1080/15434300801934702

Galaczi, E. D. (2013). Content Analysis. In *The Companion to Language Assessment*: John Wiley & Sons, Inc.

Galaczi, E. D. (2014). Interactional Competence across Proficiency Levels: How do Learners Manage Interaction in Paired Speaking Tests? *Applied Linguistics, 35*(5), 553-574. doi:10.1093/applin/amt017

Galaczi, E. D., & Taylor, L. (2018). Interactional Competence: Conceptualisations, Operationalisations, and Outstanding Questions. *Language Assessment Quarterly, 15*(3), 219-236. doi:10.1080/15434303.2018.1453816

Galaczi, E. D., & Vidakovic, I. (2010). Testing legal English: Insights from the International legal English test. *Professional and Academic English, 35*(March).

Gan, Z., Davison, C., & Hamp-Lyons, L. (2009). Topic Negotiation in Peer Group Oral Assessment Situations: A Conversation Analytic Approach. *Applied Linguistics, 30*(3), 315-334. doi:10.1093/applin/amn035

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Retrieved from https://files.eric.ed.gov/fulltext/ED532068.pdf

Green, A. (1998). *Verbal protocol analysis in language testing research: a handbook*. Cambridge, UK: Cambridge University Press.

Gustafsson, J.-E. (2013). Förändringar i kunskapbedömning på individ- och systemnivå i den svenska skolan under 25 år [Changes in educational assessment at the invidual and system level in the Swedish school during 25 years]. In I. Wernersson & I. Gerrbo (Eds.), *Differentieringens janusansikte: En antologi från Institutionen för Pedagogik och Specialpedagogik vid Göteborgs universitet [The Janus face of differentiation: An anthology from the Department of Education and Special Education at the University of Gothenburg]* Gothenburg: University of Gothenburg.

Gustafsson, J.-E., Cliffordson, C., & Erickson, G. (2014). *Likvärdig kunskapsbedömning i och av den svenska skolan : problem och möjligheter [Equal assessment of knowledge in and of the Swedish school system – problems and possibilities]* Stockholm: SNS förlag.

Gustafsson, J.-E., & Erickson, G. (2013). To trust or not to trust?—teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability, 25*(1), 69-87. doi:10.1007/s11092-013-9158-x

Hall, J. K. (1993). The Role of Oral Practices in the Accomplishment of Our Everyday Lives: The Sociocultural Dimension of Interaction with Implications for the Learning of Another Language1. *Applied Linguistics, 14*(2), 145-166. doi:10.1093/applin/14.2.145

Hall, J. K. (1995). (Re)creating our Worlds with Words: A Sociohistorical Perspective of Face-to-Face Interaction. *Applied Linguistics, 16*(2), 206-232. doi:10.1093/applin/16.2.206

Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. *Curriculum Journal, 16*(2), 207-223. doi:10.1080/09585170500136093

He, A. W., & Young, R. (1998). Language Proficiency Interviews: A discourse approach. In Y. R. & H. A. W (Eds.), *Talking and testing : discourse approaches to the assessment of oral proficiency*. Amsterdam; Philadelphia, Pa.: J. Benjamins Pub. Co.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing, 23*(3), 370-401. doi:10.1191/0265532206lt333oa

Highlights Report: Languages in Secondary Education. (2015, September 29). Retrieved from http://www.european-net.org/2015/09/eurydice-report-languages-in-secondary-education/

Holmström, C. (2018, 18 April, 2018). Friskolor i Sverige [Independent Schools in Sweden]. Retrieved from https://www.ekonomifakta.se/Fakta/Valfarden-i-privat-regi/Skolan-i-privat-regi/Antal-friskolor-i-Sverige/

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and Moderators of Bias in Observer Ratings: A Meta-Analysis. *Psychological Methods, 4*(4), 403-424. doi:10.1037/1082-989X.4.4.403

Hsieh, H.-F., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research, 15*(9), 1277-1288. doi:10.1177/1049732305276687

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

Hughes, A. (2002). *Testing for Language Teachers* (2 ed.). Cambridge: Cambridge University Press.

Hulstijn, J. H. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency 1. *Modern Language Journal, 91*(4), 663-667. doi:10.1111/j.1540-4781.2007.00627_5.x

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolingustics: Selected readings* (pp. 269-293). Harmondsworth: Penguin.

IBM Corp. (2012). IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

IBM Corp. (2017). IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Iwashita, N. (2001). The effect of learner proficiency on interactional moves and modified output in nonnative–nonnative interaction in Japanese as a foreign language. *System, 29*(2), 267-287. doi:10.1016/S0346-251X(01)00015-X

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics, 29*(1), 24-49. doi:10.1093/applin/amm017

Jacoby, S., & Ochs, E. (1995). Co-Construction: An Introduction. *Research on Language and Social Interaction, 28*(3), 171-183. doi:10.1207/s15327973rlsi2803_1

Johnson, M. (2001). *The art of non-conversation: A re-examination of the validity of the oral proficiency interview*. New Haven, CT: Yale University Press.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1-73. doi:doi:10.1111/jedm.12000

Kane, M., Crooks, T., & Cohen, A. (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17. doi:10.1111/j.1745-3992.1999.tb00010.x

Kantarcioglu, E. (2012). *Relating an Institutional Proficiency Examination to the CEFR: a case study.* (Unpublished PhD thesis), University of Roehampton, UK.

Khalifa, H., & Weir, C. (2009). *Examining Reading: Research and Practice in Assessing second language reading, Studies in Language testing 29*. Cambridge: Cambridge University Press.

Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing*. doi:10.1177/0265532217710049

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing, 16*(2), 163-188. doi:10.1177/026553229901600203

Kramsch, C. (1986). From Language Proficiency to Interactional Competence. *The Modern Language Journal, 70*(4), 366-372. doi:10.2307/326815

Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology* (3rd ed.). Thousand Oaks, CA: Sage.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book*. London: Longman.

Lam, D. M. K. (2015). *Assessing interactional competence: The case of school-based speaking assessment in Hong Kong.* University of Edingburgh, UK, Unpublished doctoral thesis.

Lindblad, T. (1992). Oral tests in Swedish schools: A five-year experiment. *System, 20*(3), 279-292. doi:http://dx.doi.org/10.1016/0346-251X(92)90040-A

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy. *The Modern Language Journal, 91*(4), 645-655. doi:doi:10.1111/j.1540-4781.2007.00627_2.x

Little, D. (2009). *The European Language Portfolio: where pedagogy and assessment meet.* Strasbourg: Council of Europe Retrieved from http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Publications/ELP_pedagogy_assessment_Little_EN.pdf.

Luk, J. (2010). Talking to Score: Impression Management in L2 Oral Assessment and the Co-Construction of a Test Discourse Genre. *Language Assessment Quarterly, 7*(1), 25-53. doi:10.1080/15434300903473997

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54-71. doi:10.1177/026553229501200104

Malone, M. E. (2017). Training in Language Assessment. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language Testing and Assessment* (pp. 225-239). Cham: Springer International Publishing.

Marklund, S. (1987). *Skolsverige 1950-1975 [School in Sweden 1950-1957]* (Vol. 5 Curriculum). Stockholm: Swedish National Board of Education.

May, L. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Langugae Testing, 1*, 29-51.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*(3), 397-421. doi:10.1177/0265532209104668

May, L. (2011a). *Interaction in Paired Speaking Test: The Rater's Perspective*. Frankfurt, Germany: Peter Lang.

May, L. (2011b). Interactional Competence in a Paired Speaking Test: Features Salient to Raters. *Language Assessment Quarterly, 8*(2), 127-145. doi:10.1080/15434303.2011.565845

McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal, 1*(E4). doi:10.1017/S2041536210000012

McNamara, T. (1996). *Measuring Second Language Performance*. London and New York: Addison Wesley Longman.

McNamara, T. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446-466. doi:10.1093/applin/18.4.446

McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing, 18*(4), 333-349. doi:10.1177/026553220101800402

McNamara, T. (2006). Validity in Language Testing: The Challenge of Sam Messick's Legacy. *Language Assessment Quarterly, 3*(1), 31-51. doi:10.1207/s15434311laq0301_3

McNamara, T. (2010). The use of language tests in the service of policy: issues of validity. *Revue française de linguistique appliquée 2010/1, Vol. XV*, 7-23.

McNamara, T., & Roever, C. (2006). *Language testing: the social dimension*. Oxford: Blackwell Publishing.

Mehrens, W. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-18.

Messick, S. A. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89*, 575-588.

Messick, S. A. (1989a). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher, 18*(2), 5-11. doi:10.3102/0013189x018002005

Messick, S. A. (1989b). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Mons, N. (2009). *Theoretical and real effects of standardised assessment. Background paper to the study National Testing of Pupils in Europe: Objectives, Organisation and Use of Results.* Brussels: Eurydice

Moss, P. A. (2003). Reconceptualizing Validity for Classroom Assessment. *Educational Measurement: Issues and Practice, 22*(4), 13-25. doi:10.1111/j.1745-3992.2003.tb00140.x

Moss, P. A. (2013). Validity in Action: Lessons From Studies of Data Use. *Journal of Educational Measurement, 50*(1), 91-98. doi:10.1111/jedm.12003

Muñoz, A. P., & Álvarez, M. E. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing, 27*(1), 33-49. doi:10.1177/0265532209347148

Nakatsuhara, F. (2006). The impact of proficiency level on conversational styles in paired speaking tests. *Cambridge ESOL Research Notes, 25*, 15-20.

Nakatsuhara, F. (2009). *Conversational styles in group oral tests: How is the conversation constructed?* (Unpublished PhD thesis), University of Essex, UK.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483-508. doi:10.1177/0265532211398110

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Frankfurt am Main, Germany: Peter Lang.

National Assessment Project. (2017). Kursproven i engelska för gymnasieskolan [National tests of English for upper secondary school]. Retrieved from https://nafs.gu.se/prov_engelska/engelska_gymn/resultat

Newton, P. E., & Shaw, S. D. (2014). *Validity in Educational and Psychological Assessment*. London: SAGE Publications.

Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT Journal, 59*(4), 287-297. doi:10.1093/elt/cci057

Nusche, D., Halász, G., Looney, J., Santiago, P., & Shewbridge, C. (2011). *OECD Reviews of Evaluation and Assessment in Education: Sweden*. Paris: OECD.

O'Sullivan, B. (2013). Assessing Speaking. In The Companion to Language Assessment: John Wiley & Sons, Inc. Retrieved from http://dx.doi.org/10.1002/9781118411360.wbcla084. doi:10.1002/9781118411360.wbcla084

O’Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192. doi:10.1191/0265532202lt226oa

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*(3), 373-386. doi:10.1016/S0346-251X(00)00018-X

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277-295. doi:10.1191/0265532202lt205oa

O'Sullivan, B. (2005). *Levels Specification Project Report.* Internal report, Zayed University: United Arab Emirates.

O'Sullivan, B. (2011a). The City & Guilds Communicator Examination linking project: a brief overview with reflections on the process. In W. Martyniuk (Ed.), *Relating Language Examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual.* Cambridge: Cambridge University Press.

O'Sullivan, B. (2011b). Language testing. In J. Simpson (Ed.), *The Routledge Handbook of Applied Linguistics* (pp. 259-273). Abingdon, England: Routledge.

O'Sullivan, B., & Weir, C. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices.* Basingstoke: Palgrave Macmillan.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56. doi:10.1191/0265532202lt219oa

Ockey, G. J. (2001). Is the oral interview superior to the group oral? . *Working Papers, International University of Japan*, 22-40.

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing, 26*(2), 161-186. doi:10.1177/0265532208101005

OECD. (1998). *Education at a glance. OECD indicators 1998*. Paris: Centre for Educational Research and Innovation, OECD.

OECD. (2015). *Improving Schools in Sweden: An OECD Perspective*. Paris: OECD.

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System, 30*(2), 143-154. doi:10.1016/S0346-251X(02)00002-7

Plough, I., Banerjee, J., & Iwashita, N. (2018). Revisiting the speaking construct: The question of interactional competence. *Language Testing, 35*(3), 325-329. doi:10.1177/0265532218772322

Popham, J. (1997). Consequential Validity: Right Concern—Wrong Concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Roever, C. (2004). Difficulty and Practicality in Tests of Interlanguage Pragmatics. In D. Boxer & A. Cohen (Eds.), *Studying Speaking to Inform Second Language Learning* (pp. 283 - 301). Clevedon: Multilingual Matters.

Rönnberg, L. (2011). Exploring the Intersection of Marketisation and Central State Control through Swedish National School Inspection. *Education Inquiry, 2*(4), 689-707. doi:10.3402/edui.v2i4.22007

Shapiro, G., & Markoff, J. (1997). 'A Matter of Definition'. In C. W. Roberts (Ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shaw, S., & Weir, C. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing, Studies in Language Testing 26*. Cambridge: Cambridge University Press and Cambridge ESOL.

Shepard, L. A. (1993). Evaluating Test Validity. *Review of Research in Education, 19*, 405. doi:10.2307/1167347

Shohamy, E. (2001). *The Power of Tests*. Harlow: Pearson Education Limited: Pearson.

Shohamy, E., Reves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *ELT Journal, 40*(3), 212-220. doi:10.1093/elt/40.3.212

Skolverket. (2009). *Gemensam europeisk referensram för språk: lärande, undervisning och bedömning*. Stockholm: Skolverket.

Spence-Brown, R. (2001). The eye of the beholder: authenticity in an embedded assessment task. *Language Testing, 18*(4), 463-481. doi:10.1177/026553220101800408

Spolsky, B. (1976). *Language Testing: Art of Science?* Paper presented at the Proceedings of the 4th International Congress of Applied Linguistics, Stuttgart.

Swedish Ministry of Education and Research. (1942). *Betänkande om betygssättningen i folkskolan. SOU 1942:11 [Report on grading in comprehensive school. Report number: SOU 1942:11]*. Stockholm.

Swedish Ministry of Education and Research. (2007a). *Tydlig och öppen - Förslag till en stärkt skolinspektion. SOU 2007:101 [Clear and open - Suggestions for a strengthened Schools Inpsectorate. Government Report, publication number SOU 2007:101]*. Stockholm: Swedish Ministry of Education and Research.

Swedish Ministry of Education and Research. (2007b). *Tydliga mål och kunskapskrav i grundskolan - Förslag till nytt mål- och uppföljningssystem. SOU 2007:28. [Clear goals and knowledge requirements in compulsory school. Government Report, publication number SOU 2007:28]*. Stockholm: Swedish Ministry of Education and Research.

Swedish Ministry of Education and Research. (2016). *Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning. SOU 2016:25. [Equivalent, fair and efficient – a new national system for knowledge assessment. Government Report, publication number SOU 2016:25.]*. Stockholm.

Swedish Ministry of Education and Research. (2017a). *Förordning (2017:1106) om en försöksverksamhet med datorbaserade nationella prov, extern bedömning och medbedömning [Ordinance (2017:1106) regarding a pilot project involving digitalised national tests, external rating and co-assessment]*. Stockholm: Swedish Ministry of Education and Research.

Swedish Ministry of Education and Research. (2017b). *Nationella prov – rättvisa, likvärdiga, digitala. Prop. 2017/18:14 [National tests - fair, equivalent and digital. Government Proposition, publication number Prop. 2017/18:14.]*. Stockholm: Swedish Ministry of Education and Research.

Swedish National Agency for Education. (1994). *Läroplan för det obligatoriska skolväsendet och de frivilliga skolformerna. Lpo 94 Lpf 94 [The curriculum for the compulsory school system and the non-compulsory school forms. Lpo 94 Lpf 94]*. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2007). *Provbetyg – Slutbetyg- Likvärdig bedömning? En statistisk analys av sambandet mellan nationella prov och slutbetyg i grundskolans årskurs 9, 1998-2006 [Grades on national tests – Final grades - Equivalent assessment, A statistical analysis of the relationship between national tests and final grades in compulsory school Grade 9, 1998–2006]*. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2011). *Läroplan, examensmål och gymnasiegemensamma ämnen för gymnasieskola 2011 [Curriculum, degree objectives and upper secondary foundation subjects for the upper secondary school 2011]*. Stockholm: Swedish National Agency for Education, .

Swedish National Agency for Education. (2016a). *English 5, Spring Term 2016. Lärarinformation - inklusive bedömningsanvisningar til Delprov A Focus: Speaking [English 5, Spring Term 2016. Teacher information - including scoring guidelines for Subtest A Focus: Speaking]*. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2016b). *Utvärdering av den nya betygsskalan samt kunskapskravens utformning [An evaluation of the new grading scale and the design of the knowledge requirements]* (Dnr 2014:892). Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2017a). *English 6, Spring Term 2017. Lärarinformation - inklusive bedömningsanvisningar till Delprov A Focus: Speaking [English 6, Spring Term 2017. Teacher information - including scoring guidelines for Subtest A Focus: Speaking]*. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2017b). *Skolverkets systemramverk för nationella prov [The Swedish National Agency for Education. A common framework for national tests]* Stockholm: Swedish National Agency for Education. Retrieved from https://www.skolverket.se/om-skolverket/publikationer/visa-enskildpublikation?_xurl_=http%3A%2F%2Fwww5.skolverket.se%2Fwtpub%2Fws%2Fskolbok%2Fwpubext%2Ftrycksak%2FRecord%3Fk%3D3890

Swedish National Agency for Education. (2018a). Att organisera muntliga delprov [Organising the oral national subtests]. Retrieved from https://www.skolverket.se/undervisning/gymnasieskolan/nationella-prov-i-gymnasieskolan/genomfora-och-bedoma-prov-i-gymnasieskolan

Swedish National Agency for Education. (2018b). *Ämneskommentar. Om ämnet Engelska [Subject commentary. About the subject English].* Retrieved from https://www.skolverket.se/download/18.6011fe501629fd150a28916/1536831518394/Kommentarmaterial_gymnasieskolan_engelska.pdf.

Swedish Schools Inspectorate. (2010). *Kontrollrättning av nationella prov i grundskolan och gymnasieskolan. Redovisning av regeringsuppdrag Dnr. U2009/4877/G [Control marking of national tests for comprehensive school and upper secondary education. Report from a government commission].* Retrieved from http://www.skolinspektionen.se > Publikationer.

Swedish Schools Inspectorate. (2011). *Lika eller olika? Omrättning av nationella prov i grundskolan och gymnasieskolan. Redovisning av regeringsuppdrag Dnr. U2009/4877/G [The same or different? Remarking of national tests in compusory school and upper secondary school. Report from a government commission].* Retrieved from http://www.skolinspektionen.se > Publikationer.

Swedish Schools Inspectorate. (2012). *Lika för alla? Omrättning av nationella prov i grundskolan och gymnasieskolan under tre år [The same for all? Re-marking of national tests in compulsory school and upper secondary school during three years].* Retrieved from http://www.skolinspektionen.se > Publikationer

Swedish Schools Inspectorate. (2013). *Olikheterna är för stora. Omrättning av nationella prov i grundskolan och gymnasieskolan, 2013 [The differences are too great. Remarkings of national tests in compulsory and upper secondary school, 2013].* Retrieved from http://www.skolinspektionen.se > Publikationer.

Swedish Schools Inspectorate. (2015). *Ombedömning av nationella prov 2014: "Processerna spelar roll" [Remarking of national tests 2014: ”The procedures matter”. Report from government commisssion].* Retrieved from http://www.skolinspektionen.se > Publikationer.

Swedish Schools Inspectorate. (2016). *Samverkan för en likvärdig bedömning. Ombedömning av nationella prov 2015. Redovisning av regeringsuppdrag Dnr U2014/7535/GV [Collaboration for a reliable assessment. Remarkings of national tests 2015. Report from a government commission. Ref U2014/7535/GV].* Retrieved from http://www.skolinspektionen.se > Publikationer.

Swedish Schools Inspectorate. (2017). *Bedömningsprocessernas betydelse för likvärdigheten - Ombedömning av nationella prov 2016 [The importance of the* *scoring procedures for equity - Remarkings of national tests 2016].* Retrieved from http://www.skolinspektionen.se > Publikationer.

Tarone, E. (1998). Research on interlanguage variation: implications for language testing. In L. F. Bachman & A. Cohen (Eds.), *Interfaces between Second Language Acquisition and Language Testing Research* (pp. 71-89). Cambridge: Cambridge University Press.

Taylor, L. (2003). The Cambridge approach to speaking assessment. *Research Notes, 13*, 2-4.

Taylor, L. (2005). Washback and impact. *ELT Journal, 59*(2), 154-155. doi:10.1093/eltj/cci030

Taylor, L. (Ed.) (2011). *Examining speaking: research and practice in assessing second language speaking / edited by Lynda Taylor.* Cambridge: Cambridge University Press.

Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing, 26*(3), 325-339. doi:10.1177/0265532209104665

Tholin, J. (2006). *Att kunna klara sig i ökänd natur: En studie av betyg och betygskriterier – historiska betingelser och implementering av ett nytt system [To find the way in unknown areas: A study of grades and grading criteria].* (Doctoral Thesis), University of Gothenburg, Sweden, Retrieved from https://gupea.ub.gu.se/handle/2077/16892

Timpe-Laughlin, V., & Choi, I. (2017). Exploring the Validity of a Second Language Intercultural Pragmatics Assessment Tool. *Language Assessment Quarterly, 14*(1), 19-35. doi:10.1080/15434303.2016.1256406

Toulmin, S. E. (1958). *The Uses of Argument* (Vol. 34): Cambridge University Press.

van Ek, J. A. (1975). *The Threshold level.* Strasbourg: Council of Europe.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*(23), 489-508.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*(4), 411-440. doi:10.1191/0265532206lt336oa

Van Moere, A. (2007). *Group oral test: How does task affect candidate performance and test score?* . (Unpublished doctoral thesis), Lancaster University.

Van Moere, A. (2013). Paired and Group Oral Assessment. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics.*

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

Weir, C. J., & Shaw, S. (2005). Establishing the validity of Cambridge ESOL writing tests: Towards the implementation of a socio-cognitive model for test validation. *Cambridge ESOL: Research Notes, 21*, 10-14.

Wilkins, D. A. (1976). *Notional syllabuses.* Oxford: Oxford University Press.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson & G. Engelhard Jr. (Eds.), *Objective Measurement: Theory into Practice* (Vol. 5, pp. 113-133). Stamford, Connecticut: Greenwood Publishing Group.

Wu, R. (2011). *Establishing the validity of the General English Proficiency Test Reading Component through a critical evaluation on alignment with the Common European Framework of Reference* (Unpublished PhD thesis), University of Bedfordshire, UK.

Xerri, D., & Vella Briffa, P. (2018). Introduction. In D. Xerri & P. Vella Briffa (Eds.), *Teacher Involvement in High-Stakes Language Testing* (pp. 1-7). Cham: Springer International Publishing.

Xi, X., & Davis, L. (2016). Quality factors in language assessment. In B. Jayanti & D. Tsagari (Eds.), *Handbook of second language assessment.* Berlin: De Gruyter Mouton.

Xi, X., & Sawaki, Y. (2017). Methods of Test Validation. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language Testing and Assessment* (pp. 193-209). Cham: Springer International Publishing.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199-225. doi:10.1177/0265532214557113

Young, R. (1995). Conversational Styles in Language Proficiency Interviews. *Language Learning, 45*(1), 3-42. doi:10.1111/j.1467-1770.1995.tb00961.x

Young, R. (2000). *Interactional Competence: Challenges for Validity*. Paper presented at the Annual meeting of the American Association for Applied Linguistics and the Language Testing Research Colloquium, Vancouver, BC, Canada.

Young, R. (2008). *Language and interaction. An advanced resource book*. London and New York: Routledge.

Young, R. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426-443). London & New York: Routledge.

Zhu, X., Cai, Y., Fan, J.-Q., Chan, S.-D., & Cheong, C.-M. (2017). Validation of the Oral Interaction Strategy Scale for speakers of Chinese as the first language in elementary schools. *L1 Educational Studies in Language and Literature, 17, SI ReadWritMult*(SI ReadWritMult), 1-25. doi:10.17239/L1ESLL-2017.17.03.02

Appendices

Appendix A: Letter of information and consent Study I and II

The following letter of information was distributed to the teachers who participated in Study I and II:

Bedömning av muntlig språkfärdighet i det nationella provet i Engelska 6

Studiens syfte
Bedömning är en aktuell och viktig fråga i skolan, inte minst idag, då nya styrdokument och en ny betygsskala nyss införts. I denna studie undersöks hur lärare bedömer muntlig språkfärdighet i det nationella provet i kursen Engelska 6. Syftet är att beskriva och analysera bedömningsprocessen och att undersöka bedömarsamstämmighet.

Studien har också en koppling till GERS (Gemensam europeisk referensram för språk) som Europarådet har tagit fram och som ämnesplanerna i moderna språk och engelska är relaterade till.

Studiens uppläggning och genomförande
I studien bedömer engelsklärare ett antal inspelade elevsamtal från den muntliga delen av det nationella provet i Engelska 6 från vårterminen 2013. Tiden för varje lärares deltagande är **en heldag**, som fyller en kompetensutvecklande funktion. Datum för undersökningen är den 10/6 2013. Först görs en individuell bedömning av elevsamtalen. Efter detta hålls en gruppdiskussion om bedömning och betygssättning kopplat till de nyss genomförda bedömningarna.

I studien deltar även fem finska och fem spanska lärare. Dessa lärare bedömer, under motsvarande dagar i sina respektive länder, samma elevsamtal som de svenska lärarna men utifrån referensnivåerna i GERS.

Urval och frivilligt deltagande
Allt deltagande sker på frivillig basis. Skolor och deltagare anonymiseras och kodas under bearbetning och analys av data. Såväl lärare som skolor kommer följaktligen att förbli anonyma i redovisningen av studien.

Möjlighet till reflektion
Hela undersökningen kan ses som ett led i kompetensutveckling kring bedömning och betygssättning. Jag hoppas att ni som deltar ska uppleva att medverkan i studien ger er stöd i det bedömningsuppdrag som vi lärare och skolledare har.

Med vänlig hälsning
Linda Borger
GÖTEBORGS UNIVERSITET
Institutionen för pedagogik och specialpedagogik

Appendix B: Letter of information and consent Study III

The following letter of information was distributed to the respondents of the questionnaire in Study III:

Enkät om lärares åsikter om det muntliga nationella delprovet i Engelska 5 och 6

En studie av lärares professionella uppfattningar
Bedömning och betygssättning är en aktuell och viktig fråga i skolan och en statlig utredning om de nationella proven har nyligen genomförts (SOU 2016:25). Mot bakgrund av den centrala roll som lärare har vid bedömning av nationella prov är det angeläget att belysa lärares professionella uppfattningar om provens innehåll, relevans och brukbarhet.

Den bifogade webbenkäten har syftet att undersöka lärares åsikter om olika aspekter av bedömningen och det praktiska genomförandet av det muntliga nationella delprovet i Engelska 5 och Engelska 6 i gymnasieskolan. Resultaten lämnar ett betydelsefullt bidrag till valideringen av de muntliga proven. **Genom att besvara enkäten bidrar du med värdefull information om lärares uppfattningar om bedömningen och användningen av de muntliga nationella delproven i engelska.**

Enkäten ingår som underlag för en delstudie i min avhandling om bedömning av muntlig språkfärdighet i det muntliga nationella delprovet i engelska. Jag, Linda Borger, är doktorand vid Institutionen för pedagogik och specialpedagogik vid Göteborgs universitet. Mina handledare är Gudrun Erickson, professor i pedagogik med inriktning mot språk och bedömning och Monica Rosén, professor i pedagogik. Jag har en bakgrund som gymnasielärare i engelska och svenska.

Urval och deltagande
Undersökningen baseras på enkätsvar från lärare som undervisar i engelska vid 150 slumpvis valda gymnasieskolor runt om i Sverige, och din skola är en av dem. För att få ett så rikt och representativt underlag som möjligt är varje enskilt svar viktigt. Din medverkan är alltså central för resultatens tillförlitlighet, men helt frivillig. Dina svar behandlas konfidentiellt. Det innebär att kommuner, skolor och deltagare anonymiseras och kodas under bearbetning och analys av data. Såväl lärare som kommuner/skolor kommer följaktligen att förbli anonyma i redovisningen av studiens resultat. I enkäten ställs bakgrundsfrågor om din skola. Denna information behövs för att kunna ta ställning till de resultat som kommer in. Kraven på anonymitet vid redovisning av resultaten kvarstår naturligtvis. Allt material hanteras endast av mig och mina handledare, och kommer inte andra till del.

Enkäten
Det tar ca 20 minuter att besvara enkäten, som innehåller tre frågeområden: 1) det praktiska genomförandet av de muntliga delproven, 2) bedömning i relation till styrdokument och syfte och 3) åsikter om provinnehåll och den muntliga uppgiftstypen. Du når enkäten via följande webbadress: https://sunet.artologik.net/gu/Survey/1403. Enkäten fylls lättast i på dator, men det är möjligt även via mobilen.

Stort tack på förhand för din medverkan!

Med vänlig hälsning
Linda Borger
GÖTEBORGS UNIVERSITET
Inst. för pedagogik och specialpedagogik

Studies I-III

Study I

Borger, Linda. (2014). *Looking Beyond Scores: A Study of Rater Orientations and Ratings of Speaking* (Licentiate thesis, University of Gothenburg, Gotheburg, Sweden). Retrived from: https://gupea.ub.gu.se/handle/2077/38158

Study II

Borger, Linda. (2018). Assessing Interactional Skills in a Paired Speaking Test: Raters' Interpretation of the Construct. Manuscript submitted for publication.

Study III

Borger, Linda. (2018). Evaluating a High-Stakes Speaking Test: Teachers' Practices and Views. Manuscript submitted for publication.

Tidigare utgåvor:

Editors: Kjell Härnqvist and Karl-Gustaf Stukát

1. KARL-GUSTAF STUKÁT *Lekskolans inverkan på barns utveckling.* Stockholm 1966

2. URBAN DAHLLÖF *Skoldifferentiering och undervisningsförlopp.* Stockholm 1967

3. ERIK WALLIN *Spelling. Factorial and experimental studies.* Stockholm 1967

4. BENGT-ERIK ANDERSSON *Studies in adolescent behaviour. Project Yg, Youth in Göteborg.* Stockholm 1969

5. FERENCE MARTON *Structural dynamics of learning.* Stockholm 1970

6. ALLAN SVENSSON *Relative achievement. School performance in relation to intelligence, sex and home environment.* Stockholm 1971

7. GUNNI KÄRRBY *Child rearing and the development of moral structure.* Stockholm 1971

Editors: Urban Dahllöf, Kjell Härnqvist and
Karl-Gustaf Stukát

8. ULF P. LUNDGREN *Frame factors and the teaching process. A contribution to curriculum theory and theory on teaching.* Stockholm 1972

9. LENNART LEVIN *Comparative studies in foreign-language teaching.* Stockholm 1972

10. RODNEY ÅSBERG *Primary education and national development.* Stockholm 1973

11. BJÖRN SANDGREN *Kreativ utveckling.* Stockholm 1974

12. CHRISTER BRUSLING *Microteaching - A concept in development.* Stockholm 1974

13. KJELL RUBENSON *Rekrytering till vuxenutbildning. En studie av kortutbildade yngre män.* Göteborg 1975

14. ROGER SÄLJÖ *Qualitative differences in learning as a function of the learner's conception of the task.* Göteborg 1975

15. LARS OWE DAHLGREN *Qualitative differences in learning as a function of content-oriented guidance.* Göteborg 1975

16. MARIE MÅNSSON *Samarbete och samarbetsförmåga. En kritisk granskning.* Lund 1975

17. JAN-ERIC GUSTAFSSON *Verbal and figural aptitudes in relation to instructional methods. Studies in aptitude - treatment interactions.* Göteborg 1976

18. MATS EKHOLM *Social utveckling i skolan. Studier och diskussion.* Göteborg 1976

19. LENNART SVENSSON *Study skill and learning.* Göteborg 1976

20. BJÖRN ANDERSSON *Science teaching and the development of thinking.* Göteborg 1976

21. JAN-ERIK PERNEMAN *Medvetenhet genom utbildning.* Göteborg 1977

Editors: Kjell Härnqvist, Ference Marton and
Karl-Gustaf Stukát

22. INGA WERNERSSON *Könsdifferentiering i grundskolan.* Göteborg 1977

23. BERT AGGESTEDT & ULLA TEBELIUS *Barns upplevelser av idrott.* Göteborg 1977

24. ANDERS FRANSSON *Att rädas prov och att vilja veta.* Göteborg 1978

25. ROLAND BJÖRKBERG *Föreställningar om arbete, utveckling och livsrytm.* Göteborg 1978

26. GUNILLA SVINGBY *Läroplaner som styrmedel för svensk obligatorisk skola. Teoretisk analys och ett empiriskt bidrag.* Göteborg 1978

27. INGA ANDERSSON *Tankestilar och hemmiljö.* Göteborg 1979

28. GUNNAR STANGVIK *Self-concept and school segregation.* Göteborg 1979

29. MARGARETA KRISTIANSSON *Matematikkunskaper Lgr 62, Lgr 69.* Göteborg 1979

30. BRITT JOHANSSON *Kunskapsbehov i omvårdnadsarbete och kunskapskrav i vårdutbildning.* Göteborg 1979

31. GÖRAN PATRIKSSON *Socialisation och involvering i idrott.* Göteborg 1979

32. PETER GILL *Moral judgments of violence among Irish and Swedish adolescents.* Göteborg 1979

33. TAGE LJUNGBLAD *Förskola - grundskola i samverkan. Förutsättningar och hinder.* Göteborg 1980

34. BERNER LINDSTRÖM *Forms of representation, content and learning.* Göteborg 1980

35. CLAES-GÖRAN WENESTAM *Qualitative differences in retention.* Göteborg 1980

36. BRITT JOHANSSON *Pedagogiska samtal i vårdutbildning. Innehåll och språkbruk.* Göteborg 1981

37. LEIF LYBECK *Arkimedes i klassen. En ämnespedagogisk berättelse.* Göteborg 1981

38. BIÖRN HASSELGREN *Ways of apprehending children at play. A study of pre-school student teachers' development.* Göteborg 1981

39. LENNART NILSSON *Yrkesutbildning i nutidshistoriskt perspektiv. Yrkesutbildningens utveckling från skråväsendets upphörande 1846 till 1980-talet samt tankar om framtida inriktning.* Göteborg 1981

40. GUDRUN BALKE-AURELL *Changes in ability as related to educational and occupational experience.* Göteborg 1982

41. ROGER SÄLJÖ *Learning and understanding. A study of differences in constructing meaning from a text.* Göteborg 1982

42. ULLA MARKLUND *Droger och påverkan. Elevanalys som utgångspunkt för drogundervisning.* Göteborg 1983

43. SVEN SETTERLIND *Avslappningsträning i skolan. Forskningsöversikt och empiriska studier.* Göteborg 1983

44. EGIL ANDERSSON & MARIA LAWENIUS *Lärares uppfattning av undervisning.* Göteborg 1983

45. JAN THEMAN *Uppfattningar av politisk makt.* Göteborg 1983

46. INGRID PRAMLING *The child's conception of learning.* Göteborg 1983

47. PER OLOF THÅNG *Vuxenlärarens förhållningssätt till deltagarerfarenheter. En studie inom AMU.* Göteborg 1984

48. INGE JOHANSSON *Fritidspedagog på fritidshem. En yrkesgrupps syn på sitt arbete.* Göteborg 1984

49. GUNILLA SVANBERG *Medansvar i undervisning. Metoder för observation och kvalitativ analys.* Göteborg 1984

50. SVEN-ERIC REUTERBERG *Studiemedel och rekrytering till högskolan.* Göteborg 1984

51. GÖSTA DAHLGREN & LARS-ERIK OLSSON *Läsning i barnperspektiv.* Göteborg 1985

52. CHRISTINA KÄRRQVIST *Kunskapsutveckling genom experimentcentrerade dialoger i ellära.* Göteborg 1985

53. CLAES ALEXANDERSSON *Stabilitet och förändring. En empirisk studie av förhållandet mellan skolkunskap och vardagsvetande.* Göteborg 1985

54. LILLEMOR JERNQVIST *Speech regulation of motor acts as used by cerebral palsied children. Observational and experimental studies of a key feature of conductive education.* Göteborg 1985

55. SOLVEIG HÄGGLUND *Sex-typing and development in an ecological perspective.* Göteborg 1986

56. INGRID CARLGREN *Lokalt utvecklingsarbete.* Göteborg 1986

57. LARSSON, ALEXANDERSSON, HELMSTAD & THÅNG *Arbetsupplevelse och utbildningssyn hos icke facklärda.* Göteborg 1986

58. ELVI WALLDAL *Studerande vid gymnasieskolans vårdlinje. Förväntad yrkesposition, rollpåverkan, självuppfattning.* Göteborg 1986

Editors: Jan-Eric Gustafsson, Ference Marton and
Karl-Gustaf Stukát

59. EIE ERICSSON *Foreign language teaching from the point of view of certain student activities.* Göteborg 1986

60. JAN HOLMER *Högre utbildning för lågutbildade i industrin.* Göteborg 1987

61. ANDERS HILL & TULLIE RABE *Psykiskt utvecklingsstörda i kommunal förskola.* Göteborg 1987

62. DAGMAR NEUMAN *The origin of arithmetic skills. A phenomenographic approach.* Göteborg 1987

63. TOMAS KROKSMARK *Fenomenografisk didaktik.* Göteborg 1987

64. ROLF LANDER *Utvärderingsforskning - till vilken nytta?* Göteborg 1987

65. TORGNY OTTOSSON *Map-reading and wayfinding.* Göteborg 1987

66. MAC MURRAY *Utbildningsexpansion, jämlikhet och avlänkning.* Göteborg 1988

67. ALBERTO NAGLE CAJES *Studievalet ur den väljandes perspektiv.* Göteborg 1988

68. GÖRAN LASSBO *Mamma - (Pappa) - barn. En utvecklingsekologisk studie av socialisation i olika familjetyper.* Göteborg 1988

69. LENA RENSTRÖM *Conceptions of matter. A phenomenographic approach.* Göteborg 1988

70. INGRID PRAMLING *Att lära barn lära.* Göteborg 1988

71. LARS FREDHOLM *Praktik som bärare av undervisnings innehåll och form. En förklaringsmodell för uppkomst av undervisningshandlingar inom en totalförsvarsorganisation.* Göteborg 1988

72. OLOF F. LUNDQUIST *Studiestöd för vuxna. Utveckling, utnyttjande, utfall.* Göteborg 1989

73. BO DAHLIN *Religionen, själen och livets mening. En fenomenografisk och existensfilosofisk studie av religionsundervisningens villkor.* Göteborg 1989

74. SUSANNE BJÖRKDAHL ORDELL *Socialarbetare. Bakgrund, utbildning och yrkesliv.* Göteborg 1990

75. EVA BJÖRCK-ÅKESSON *Measuring Sensation Seeking.* Göteborg 1990

76. ULLA-BRITT BLADINI *Från hjälpskolelärare till förändringsagent. Svensk speciallärarutbildning 1921-1981 relaterad till specialundervisningens utveckling och förändringar i speciallärarens yrkesuppgifter.* Göteborg 1990

77. ELISABET ÖHRN *Könsmönster i klassrumsinteraktion. En observations- och intervjustudie av högstadieelevers lärarkontakter.* Göteborg 1991

78. TOMAS KROKSMARK *Pedagogikens vägar till dess första svenska professur.* Göteborg 1991

Editors: Ingemar Emanuelsson, Jan-Eric Gustafsson
and Ference Marton

79. ELVI WALLDAL *Problembaserad inlärning. Utvärdering av påbyggnadslinjen Utbildning i öppen hälso- och sjukvård.* Göteborg 1991

80. ULLA AXNER *Visuella perceptionssvårigheter i skolperspektiv. En longitudinell studie.* Göteborg 1991

81. BIRGITTA KULLBERG *Learning to learn to read.* Göteborg 1991

82. CLAES ANNERSTEDT *Idrottslärarna och idrottsämnet. Utveckling, mål, kompetens - ett didaktiskt perspektiv.* Göteborg 1991

83. EWA PILHAMMAR ANDERSSON *Det är vi som är dom. Sjuksköterskestuderandes föreställningar och perspektiv under utbildningstiden.* Göteborg 1991

84. ELSA NORDIN *Kunskaper och uppfattningar om maten och dess funktioner i kroppen. Kombinerad enkät- och intervjustudie i grundskolans årskurser 3, 6 och 9.* Göteborg 1992

85. VALENTIN GONZÁLEZ *On human attitudes. Root metaphors in theoretical conceptions.* Göteborg 1992

86. JAN-ERIK JOHANSSON *Metodikämnet i förskollärarutbildningen. Bidrag till en traditionsbestämning.* Göteborg 1992

87. ANN AHLBERG *Att möta matematiska problem. En belysning av barns lärande.* Göteborg 1992

88. ELLA DANIELSON *Omvårdnad och dess psykosociala inslag. Sjuksköterskestuderandes uppfattningar av centrala termer och reaktioner inför en omvårdnadssituation.* Göteborg 1992

89. SHIRLEY BOOTH *Learning to program. A phenomenographic perspective.* Göteborg 1992

90. EVA BJÖRCK-ÅKESON *Samspel mellan små barn med rörelsehinder och talhandikapp och deras föräldrar - en longitudinell studie.* Göteborg 1992

91. KARIN DAHLBERG *Helhetssyn i vården. En uppgift för sjuksköterskeutbildningen.* 1992

92. RIGMOR ERIKSSON *Teaching Language Learning. In-service training for communicative teaching and self directed learning in English as a foreign language.* 1993

93. KJELL HÄRENSTAM *Skolboks-islam. Analys av bilden av islam i läroböcker i religionskunskap.* Göteborg 1993.

94. INGRID PRAMLING *Kunnandets grunder. Prövning av en fenomenografisk ansats till att utveckla barns sätt att uppfatta sin omvärld.* Göteborg 1994.

95. MARIANNE HANSSON SCHERMAN *Att vägra vara sjuk. En longitudinell studie av förhållningssätt till astma/allergi.* Göteborg 1994

96. MIKAEL ALEXANDERSSON *Metod och medvetande.* Göteborg 1994

97. GUN UNENGE *Pappor i föräldrakooperativa daghem. En deskriptiv studie av pappors medverkan.* Göteborg 1994

98. BJÖRN SJÖSTRÖM *Assessing acute postoperative pain. Assessment strategies and quality in relation to clinical experience and professional role.* Göteborg 1995

99. MAJ ARVIDSSON *Lärares orsaks- och åtgärdstankar om elever med svårigheter.* Göteborg 1995

100. DENNIS BEACH *Making sense of the problems of change: An ethnographic study of a teacher education reform.* Göteborg 1995.

101. WOLMAR CHRISTENSSON *Subjektiv bedömning - som besluts och handlingsunderlag.* Göteborg 1995

102. SONJA KIHLSTRÖM *Att vara förskollärare. Om yrkets pedagogiska innebörder.* Göteborg 1995

103. MARITA LINDAHL *Inlärning och erfarande. Ettåringars möte med förskolans värld.* Göteborg. 1996

104. GÖRAN FOLKESTAD *Computer Based Creative Music Making - Young Peoples´ Music in the Digital Age.* Göteborg 1996

105. EVA EKEBLAD *Children • Learning • Numbers. A phenomenographic excursion into first-grade children's arithmetic.* Göteborg 1996

106. HELGE STRÖMDAHL *On mole and amount of substance. A study of the dynamics of concept formation and concept attainment.* Göteborg 1996

107. MARGARETA HAMMARSTRÖM *Varför inte högskola? En longitudinell studie av olika faktorers betydelse för studiebegåvade ungdomars utbildningskarriär.* Göteborg 1996

108. BJÖRN MÅRDÉN *Rektorers tänkande. En kritisk betraktelse av skolledarskap.* Göteborg 1996

109. GLORIA DALL'ALBA & BIÖRN HASSELGREN (EDS) *Reflections on Phenomenography - Toward a Methodology?* Göteborg 1996

110. ELISABETH HESSLEFORS ARKTOFT *I ord och handling. Innebörder av "att anknyta till elevers erfarenheter", uttryckta av lärare.* Göteborg 1996

111. BARBRO STRÖMBERG *Professionellt förhållningssätt hos läkare och sjuksköterskor. En studie av uppfattningar.* Göteborg 1997

112. HARRIET AXELSSON Våga lära. *Om lärare som förändrar sin miljöundervisning.* Göteborg 1997

113. ANN AHLBERG *Children's ways of handling and experiencing numbers.* Göteborg 1997

114. HUGO WIKSTRÖM *Att förstå förändring. Modellbyggande, simulering och gymnasieelevers lärande.* Göteborg 1997

115. DORIS AXELSEN *Listening to recorded music. Habits and motivation among high-school students.* Göteborg 1997.

116. EWA PILHAMMAR ANDERSSON *Handledning av sjuksköterskestuderande i klinisk praktik.* Göteborg 1997

117. OWE STRÅHLMAN *Elitidrott, karriär och avslutning.* Göteborg 1997

118. AINA TULLBERG T*eaching the 'mole'. A phenomenographic inquiry into the didactics of chemistry.* Göteborg 1997.

119. DENNIS BEACH *Symbolic Control and Power Relay Learning in Higher Professional Education.* Göteborg 1997

120. HANS-ÅKE SCHERP *Utmanande eller utmanat ledarskap. Rektor, organisationen och förändrat undervisningsmönster i gymnasieskolan.* Göteborg 1998

121. STAFFAN STUKÁT *Lärares planering under och efter utbildningen.* Göteborg 1998

122. BIRGIT LENDAHLS ROSENDAHL *Examensarbetets innebörder. En studie av blivande lärares utsagor.* Göteborg 1998

123. ANN AHLBERG *Meeting Mathematics. Educational studies with young children.* Göteborg 1998

124. MONICA ROSÉN *Gender Differences in Patterns of Knowledge.* Göteborg 1998.

125. HANS BIRNIK *Lärare- elevrelationen. Ett relationistiskt perspektiv.* Göteborg 1998

126. MARGRETH HILL *Kompetent för "det nya arbetslivet"? Tre gymnasieklasser reflekterar över och diskuterar yrkesförberedande studier.* Göteborg 1998

127. LISBETH ÅBERG-BENGTSSON *Entering a Graphicate Society. Young Children Learning Graphs and Charts.* Göteborg 1998

128. MELVIN FEFFER *The Conflict of Equals: A Constructionist View of Personality Development.* Göteborg 1999

129. ULLA RUNESSON *Variationens pedagogik. Skilda sätt att behandla ett matematiskt innehåll.* Göteborg 1999

130. SILWA CLAESSON *"Hur tänker du då?" Empiriska studier om relationen mellan forskning om elevuppfattningar och lärares undervisning.* Göteborg 1999

131. MONICA HANSEN *Yrkeskulturer i möte. Läraren, fritidspedagogen och samverkan.* Göteborg 1999

132. JAN THELIANDER *Att studera arbetets förändring under kapitalismen. Ure och Taylor i pedagogiskt perspektiv.* Göteborg 1999

133. TOMAS SAAR *Musikens dimensioner - en studie av unga musikers lärande.* Göteborg 1999

134. GLEN HELMSTAD *Understandings of understanding. An inquiry concerning experiential conditions for developmental learning.* Göteborg 1999

135. MARGARETA HOLMEGAARD *Språkmedvetenhet och ordinlärning. Lärare och inlärare reflekterar kring en betydelsefältsövning i svenska som andraspråk.* Göteborg 1999

136. ALYSON MCGEE *Investigating Language Anxiety through Action Inquiry: Developing Good Research Practices.* Göteborg 1999

137. EVA GANNERUD *Genusperspektiv på lärargärning. Om kvinnliga klasslärares liv och arbete.* Göteborg 1999

138. TELLERVO KOPARE *Att rida stormen ut. Förlossningsberättelser i Finnmark och Sápmi.* Göteborg 1999

139. MAJA SÖDERBÄCK *Encountering Parents. Professional Action Styles among Nurses in Pediatric Care.* Göteborg 1999

140. AIRI ROVIO - JOHANSSON *Being Good at Teaching. Exploring different ways of handling the same subject in Higher Education.* Göteborg 1999

141. EVA JOHANSSON *Etik i små barns värld. Om värden och normer bland de yngsta barnen i förskolan.* Göteborg 1999

142. KENNERT ORLENIUS *Förståelsens paradox. Yrkeserfarenhetens betydelse när förskollärare blir grundskollärare.* Göteborg 1999.

143. BJÖRN MÅRDÉN *De nya hälsomissionärerna – rörelser i korsvägen mellan pedagogik och hälsopromotion.* Göteborg 1999

144. MARGARETA CARLÉN *Kunskapslyft eller avbytarbänk? Möten med industriarbetare om utbildning för arbete.* Göteborg 1999

145. MARIA NYSTRÖM *Allvarligt psykiskt störda människors vardagliga tillvaro.* Göteborg 1999

146. ANN-KATRIN JAKOBSSON *Motivation och inlärning ur genusperspektiv. En studie av gymnasieelever på teoretiska linjer/program.* Göteborg 2000

147. JOANNA GIOTA *Adolescents' perceptions of school and reasons for learning.* Göteborg 2000

148. BERIT CARLSTEDT *Cognitive abilities – aspects of structure, process and measurement.* Göteborg 2000

149. MONICA REICHENBERG *Röst och kausalitet i lärobokstexter. En studie av elevers förståelse av olika textversioner.* Göteborg 2000

150. HELENA ÅBERG *Sustainable waste management in households – from international policy to everyday practice. Experiences from two Swedish field studies.* Göteborg 2000

151. BJÖRN SJÖSTRÖM & BRITT JOHANSSON *Ambulanssjukvård. Ambulanssjukvårdares och läkares perspektiv.* Göteborg 2000

152. AGNETA NILSSON *Omvårdnadskompetens inom hemsjukvård – en deskriptiv studie.* Göteborg 2001

153. ULLA LÖFSTEDT *Förskolan som lärandekontext för barns bildskapande.* Göteborg 2001

154. JÖRGEN DIMENÄS *Innehåll och interaktion. Om elevers lärande i naturvetenskaplig undervisning.* Göteborg 2001

155. BRITT MARIE APELGREN *Foreign Language Teachers' Voices. Personal Theories and Experiences of Change in Teaching English as a Foreign Language in Sweden.* Göteborg 2001

156. CHRISTINA CLIFFORDSON *Assessing empathy: Measurement characteristics and interviewer effects.* Göteborg 2001

157. INGER BERGGREN *Identitet, kön och klass. Hur arbetarflickor formar sin identitet.* Göteborg 2001

158. CARINA FURÅKER *Styrning och visioner – sjuksköterskeutbildning i förändring.* Göteborg 2001

159. INGER BERNDTSSON *Förskjutna horisonter. Livsförändring och lärande i samband med synnedsättning eller blindhet.* Göteborg 2001

160. SONJA SHERIDAN *Pedagogical Quality in Preschool. An issue of perspectives.* Göteborg 2001

161. JAN BAHLENBERG *Den otroliga verkligheten sätter spår. Om Carlo Derkerts liv och konstpedagogiska gärning.* Göteborg 2001

162. FRANK BACH *Om ljuset i tillvaron. Ett undervisningsexperiment inom optik.* Göteborg 2001

163. PIA WILLIAMS *Barn lär av varandra. Samlärande i förskola och skola.* Göteborg 2001

164. VIGDIS GRANUM *Studentenes forestillinger om sykepleie som fag og funksjon.* Göteborg 2001

165. MARIT ALVESTAD *Den komplekse planlegginga. Førskolelærarar om pedagogisk planlegging og praksis.* Göteborg 2001

166. GIRMA BERHANU *Learning-In-Context. An Ethnographic Investigation of Mediated Learning Experiences among Ethiopian Jews in Israel.* Göteborg 2001.

167. OLLE ESKILSSON *En longitudinell studie av 10 – 12-åringars förståelse av materiens förändringar.* Göteborg 2001

168. JONAS EMANUELSSON *En fråga om frågor. Hur lärares frågor i klassrummet gör det möjligt att få reda på elevernas sätt att förstå det som undervisningen behandlar i matematik och naturvetenskap.* Göteborg 2001

169. BIRGITTA GEDDA *Den offentliga hemligheten. En studie om sjuksköterskans pedagogiska funktion och kompetens i folkhälsoarbetet.* Göteborg 2001

170. FEBE FRIBERG *Pedagogiska möten mellan patienter och sjuksköterskor på en medicinsk vårdavdelning. Mot en vårddidaktik på livsvärldsgrund.* Göteborg 2001

171. MADELEINE BERGH *Medvetenhet om bemötande. En studie om sjuksköterskans pedagogiska funktion och kompetens i närståendeundervisning.* Göteborg 2002

172. HENRIK ERIKSSON *Den diplomatiska punkten – maskulinitet som kroppsligt identitetsskapande projekt i svensk sjuksköterskeutbildning.* Göteborg 2002

173. SOLVEIG LUNDGREN *I spåren av en bemanningsförändring. En studie av sjuksköterskors arbete på en kirurgisk vårdavdelning.* Göteborg 2002

174. BIRGITTA DAVIDSSON *Mellan soffan och katedern. En studie av hur förskollärare och grundskollärare utvecklar pedagogisk integration mellan förskola och skola.* Göteborg 2002

175. KARI SØNDENÅ *Tradisjon og Transcendens – ein fenomenologisk studie av refleksjon i norsk førskulelærarutdanning.* Göteborg 2002

176. CHRISTINE BENTLEY *The Roots of Variation of English-Teaching. A Phenomenographic Study Founded on an Alternative Basic Assumption.* Göteborg 2002

177. ÅSA MÄKITALO *Categorizing Work: Knowing, Arguing, and Social Dilemmas in Vocational Guidance.* Göteborg 2002

178. MARITA LINDAHL *VÅRDA – VÄGLEDA – LÄRA. Effektstudie av ett interventionsprogram för pedagogers lärande i förskolemiljön.* Göteborg 2002

179. CHRISTINA BERG *Influences on schoolchildren's dietary selection. Focus on fat and fibre at breakfast.* Göteborg 2002

180. MARGARETA ASP *Vila och lärande om vila. En studie på livsvärldsfenomenologisk grund.* Göteborg 2002

181. FERENCE MARTON & PAUL MORRIS (EDS) *What matters? Discovering critical contitions of classroom learning.* Göteborg 2002

182. ROLAND SEVERIN *Dom vet vad dom talar om. En intervjustudie om elevers uppfattningar av begreppen makt och samhällsförändring.* Göteborg 2002

Editors: Björn Andersson, Jan Holmer and
Ingrid Pramling Samuelsson

183. MARLÉNE JOHANSSON *Slöjdpraktik i skolan – hand, tanke, kommunikation och andra medierande redskap.* Göteborg 2002

184. INGRID SANDEROTH *Om lust att lära i skolan: En analys av dokument och klass 8y*. Göteborg 2002

185. INGA-LILL JAKOBSSON *Diagnos i skolan. En studie av skolsituationer för elever med syndromdiagnos*. Göteborg 2002

186. EVA-CARIN LINDGREN *Empowering Young Female Athletes – A Possible Challenge to the Male Hegemony in Sport. A Descriptive and Interventional Study*. Göteborg 2002

187. HANS RYSTEDT *Bridging practices. Simulations in education for the health-care professions*. Göteborg 2002

188. MARGARETA EKBORG *Naturvetenskaplig utbildning för hållbar utveckling? En longitudinell studie av hur studenter på grunskollärarprogrammet utvecklar för miljöundervisning relevanta kunskaper i naturkunskap*. Göteborg 2002

189. ANETTE SANDBERG *Vuxnas lekvärld. En studie om vuxnas erfarenheter av lek*. Göteborg 2002

190. GUNLÖG BREDÄNGE *Gränslös pedagog. Fyra studier om utländska lärare i svensk skola*. Göteborg 2003

191. PER-OLOF BENTLEY *Mathematics Teachers and Their Teaching. A Survey Study*. Göteborg 2003

192. KERSTIN NILSSON *MANDAT – MAKT – MANAGEMENT. En studie av hur vårdenhetschefers ledarskap konstrueras*. Göteborg 2003

193. YANG YANG *Measuring Socioeconomic Status and its Effects at Individual and Collective Levels: A Cross-Country Comparison*. Göteborg 2003

194. KNUT VOLDEN *Mediekunnskap som mediekritikk*. Göteborg 2003.

195. LOTTA LAGER-NYQVIST *Att göra det man kan – en longitudinell studie av hur sju lärarstudenter utvecklar sin undervisning och formar sin lärarroll i naturvetenskap*. Göteborg 2003

196. BRITT LINDAHL *Lust att lära naturvetenskap och teknik? En longitudinell studie om vägen till gymnasiet*. Göteborg 2003

197. ANN ZETTERQVIST *Ämnesdidaktisk kompetens i evolutionsbiologi. En intervjuundersökning med no/biologilärare*. Göteborg 2003

198. ELSIE ANDERBERG *Språkanvändningens funktion vid utveckling av kunskap om objekt*. Göteborg 2003.

199. JAN GUSTAFSSON *Integration som text, diskursiv och social praktik. En policyetnografisk fallstudie av mötet mellan skolan och förskoleklassen*. Göteborg 2003.

200. EVELYN HERMANSSON *Akademisering och professionalisering – barnmorskans utbildning i förändring*. Göteborg 2003

201. KERSTIN VON BRÖMSSEN *Tolkningar, förhandlingar och tystnader. Elevers tal om religion i det mångkulturella och postkoloniala rummet*. Göteborg 2003

202. MARIANNE LINDBLAD FRIDH *Från allmänsjuksköterska till specialistsjuksköterska inom intensivvård. En studie av erfarenheter från specialistutbildningen och från den första yrkesverksamma tiden inom intensivvården*. Göteborg 2003

203. BARBRO CARLI *The Making and Breaking of a Female Culture: The History of Swedish Physical Education 'in a Different Voice'*. Göteborg 2003

204. ELISABETH DAHLBORG-LYCKHAGE *"Systers" konstruktion och mumifiering – i TV-serier och i studenters föreställningar*. Göteborg 2003

205. ULLA HELLSTRÖM MUHLI *Att överbrygga perspektiv. En studie av behovsbedömningssamtal inom äldreinriktat socialt arbete*. Göteborg 2003

206. KRISTINA AHLBERG *Synvändor. Universitetsstudenters berättelser om kvalitativa förändringar av sätt att erfara situationers mening under utbildningspraktik*. Göteborg 2004

207. JONAS IVARSSON *Renderings & Reasoning: Studying artifacts in human knowing*. Göteborg 2004

208. MADELEINE LÖWING *Matematikundervisningens konkreta gestaltning. En studie av kommunikationen lärare – elev och matematiklektionens didaktiska ramar*. Göteborg 2004

209. PIJA EKSTRÖM *Makten att definiera. En studie av hur beslutsfattare formulerar villkor för specialpedagogisk verksamhet*. Göteborg 2004

210. CARIN ROOS *Skriftspråkande döva barn. En studie om skriftspråkligt lärande i förskola och skola*. Göteborg 2004

211. JONAS LINDEROTH *Datorspelandets mening. Bortom idén om den interaktiva illusionen*. Göteborg 2004

212. ANITA WALLIN *Evolutionsteorin i klassrummet. På väg mot en ämnesdidaktisk teori för undervisning i biologisk evolution*. Göteborg 2004

213. EVA HJÖRNE *Excluding for inclusion? Negotiating school careers and identities in pupil welfare settings in the Swedish school*. Göteborg 2004

214. MARIE BLIDING *Inneslutandets och uteslutandets praktik. En studie av barns relationsarbete i skolan*. Göteborg 2004

215. LARS-ERIK.JONSSON *Appropriating Technologies in Educational Practices. Studies in the Contexts of Compulsory Education, Higher Education, and Fighter Pilot Training*. Göteborg 2004

216. MIA KARLSSON *An ITiS Teacher Team as a Community of Practice*. Göteborg 2004

217. SILWA CLAESSON *Lärares levda kunskap*. Göteborg 2004

218. GUN-BRITT WÄRVIK *Ambitioner att förändra och artefakters verkan. Gränsskapande och stabiliserande praktiker på produktionsgolvet*. Göteborg 2004

219. KARIN LUMSDEN WASS *Vuxenutbildning i omvandling. Kunskapslyftet som ett sätt att organisera förnyelse.* Göteborg 2004

220. LENA DAHL *Amningspraktikens villkor. En intervjustudie av en grupp kvinnors föreställningar på och erfarenheter av amning.* Göteborg 2004

221. ULRIC BJÖRCK *Distributed Problem-Based Learning. Studies of a Pedagogical Model in Practice.* Göteborg 2004

222. ANNEKA KNUTSSON *"To the best of your knowledge and for the good of your neighbour". A study of traditional birth attendants in Addis Ababa, Ethiopia.* Göteborg 2004

223. MARIANNE DOVEMARK *Ansvar – flexibilitet – valfrihet. En etnografisk studie om en skola i förändring.* Göteborg 2004

224. BJÖRN HAGLUND *Traditioner i möte. En kvalitativ studie av fritidspedagogers arbete med samlingar i skolan.* Göteborg 2004

225. ANN-CHARLOTTE MÅRDSJÖ *Lärandets skiftande innebörder – uttryckta av förskollärare i vidareutbildning.* Göteborg 2005

226. INGRID GRUNDÉN *Att återerövra kroppen. En studie av livet efter en ryggmärgsskada.* Göteborg 2005

227. KARIN GUSTAFSSON & ELISABETH MELLGREN *Barns skriftspråkande – att bli en skrivande och läsande person.* Göteborg 2005

228. GUNNAR NILSSON *Att äga π. Praxisnära studier av lärarstudenters arbete med geometrilaborationer.* Göteborg 2005.

229. BENGT LINDGREN *Bild, visualitet och vetande. Diskussion om bild som ett kunskapsfält inom utbildning.* Göteborg 2005

230. PETRA ANGERVALL *Jämställdhetsarbetets pedagogik. Dilemman och paradoxer i arbetet med jämställdhet på ett företag och ett universitet.* Göteborg 2005

231. LENNART MAGNUSSON *Designing a responsive support service for family carers of frail older people using ICT.* Göteborg 2005

232. MONICA REICHENBERG *Gymnasieelever samtalar kring facktexter. En studie av textsamtal med goda och svaga läsare.* Göteborg 2005

233. ULRIKA WOLFF *Characteristics and varieties of poor readers.* Göteborg 2005

234. CECILIA NIELSEN *Mellan fakticitet och projekt. Läs- och skrivsvårigheter och strävan att övervinna dem.* Göteborg 2005.

235. BERITH HEDBERG *Decision Making and Communication in Nursing Practice. Aspects of Nursing Competence.* Göteborg 2005

236. MONICA ROSÉN, EVA MYRBERG & JAN-ERIC GUSTAFSSON *Läskompetens i skolår 3 och 4. Nationell rapport från PIRLS 2001 i Sverige. The IEA Progress in International Reading Literacy Study.* Göteborg 2005

237. INGRID HENNING LOEB *Utveckling och förändring i kommunal vuxenutbildning. En yrkeshistorisk ingång med berättelser om lärarbanor.* Göteborg 2006.

238. NIKLAS PRAMLING *Minding metaphors: Using figurative language in learning to represent.* Göteborg 2006

239. KONSTANTIN KOUGIOUMTZIS *Lärarkulturer och professionskoder. En komparativ studie av idrottslärare i Sverige och Grekland.* Göteborg 2006

240. STEN BÅTH *Kvalifikation och medborgarfostran. En analys av reformtexter avseende gymnasieskolans samhällsuppdrag.* Göteborg 2006.

241. EVA MYRBERG *Fristående skolor i Sverige – Effekter på 9-10-åriga elevers läsförmåga.* Göteborg 2006

242. MARY-ANNE HOLFVE-SABEL *Attitudes towards Swedish comprehensive school. Comparisons over time and between classrooms in grade 6.* Göteborg 2006

243. CAROLINE BERGGREN *Entering Higher Education – Gender and Class Perspectives.* Göteborg 2006

244. CRISTINA THORNELL & CARL OLIVESTAM *Kulturmöte i centralafrikansk kontext med kyrkan som arena.* Göteborg 2006

245. ARVID TREEKREM *Att leda som man lär. En arbetsmiljöpedagogisk studie av toppledares ideologier om ledarskapets taktiska potentialer.* Göteborg 2006

246. EVA GANNERUD & KARIN RÖNNERMAN *Innehåll och innebörd i lärares arbete i förskola och skola – en fallstudie ur ett genusperspektiv.* Göteborg 2006

247. JOHANNES LUNNEBLAD *Förskolan och mångfalden – en etnografisk studie på en förskola i ett multietniskt område.* Göteborg 2006

248. LISA ASP-ONSJÖ *Åtgärdsprogram – dokument eller verktyg? En fallstudie i en kommun.* Göteborg 2006

249. EVA JOHANSSON & INGRID PRAMLING SAMUELSSON *Lek och läroplan. Möten mellan barn och lärare i förskola och skola.* Göteborg 2006

250. INGER BJÖRNELOO *Innebörder av hållbar utveckling. En studie av lärares utsagor om undervisning.* Göteborg 2006

251. EVA JOHANSSON *Etiska överenskommelser i förskolebarns världar.* Göteborg 2006

252. MONICA PETERSSON *Att genuszappa på säker eller osäker mark. Hem- och konsumentkunskap ur ett könsperspektiv.* Göteborg 2007

253. INGELA OLSSON *Handlingskompetens eller inlärd hjälplöshet? Lärandeprocesser hos verkstadsindustriarbetare.* Göteborg 2007

254. HELENA PEDERSEN *The School and the Animal Other. An Ethnography of human-animal relations in education.* Göteborg 2007

255. ELIN ERIKSEN ØDEGAARD *Meningsskaping i barnehagen. Innhold og bruk av barns og voksnes samtalefortellinger.* Göteborg 2007

256. ANNA KLERFELT *Barns multimediala berättande. En länk mellan mediakultur och pedagogisk praktik.* Göteborg 2007

257. PETER ERLANDSON *Docile bodies and imaginary minds: on Schön's reflection-in-action.* Göteborg 2007

258. SONJA SHERIDAN OCH PIA WILLIAMS *Dimensioner av konstruktiv konkurrens. Konstruktiva konkurrensformer i förskola, skola och gymnasium.* Göteborg 2007

259. INGELA ANDREASSON *Elevplanen som text - om identitet, genus, makt och styrning i skolans elevdokumentation.* Göteborg 2007

Editors: Jan-Eric Gustafsson, Annika Härenstam and
Ingrid Pramling Samuelsson

260. ANN-SOFIE HOLM *Relationer i skolan. En studie av femininiteter och maskuliniteter i år 9.* Göteborg 2008

261. LARS-ERIK NILSSON *But can't you see they are lying: Student moral positions and ethical practices in the wake of technological change.* Göteborg 2008

262. JOHAN HÄGGSTRÖM *Teaching systems of linear equations in Sweden and China: What is made possible to learn?* Göteborg 2008

263. GUNILLA GRANATH *Milda makter! Utvecklingssamtal och loggböcker som disciplineringstekniker.* Göteborg 2008

264. KARIN GRAHN *Flickor och pojkar i idrottens läromedel. Konstruktioner av genus i ungdomstränarutbildningen.* Göteborg 2008.

265. PER-OLOF BENTLEY *Mathematics Teachers and Their Conceptual Models. A New Field of Research.* Göteborg 2008

266. SUSANNE GUSTAVSSON *Motstånd och mening. Innebörd i blivande lärares seminariesamtal.* Göteborg 2008

267. ANITA MATTSSON *Flexibel utbildning i praktiken. En fallstudie av pedagogiska processer i en distansutbildning med en öppen design för samarbetslärande.* Göteborg 2008

268. ANETTE EMILSON *Det önskvärda barnet. Fostran uttryckt i vardagliga kommunikationshandlingar mellan lärare och barn i förskolan.* Göteborg 2008

269. ALLI KLAPP LEKHOLM *Grades and grade assignment: effects of student and school charachterisitcs.* Göteborg 2008

270. ELISABETH BJÖRKLUND *Att erövra litteracitet. Små barns kommunikativa möten med berättande, bilder, text och tecken i förskolan.* Göteborg 2008

271. EVA NYBERG *Om livets kontinuitet. Undervisning och lärande om växters och djurs livscykler - en fallstudie i årskurs 5.* Göteborg 2008

272. CANCELLED

273. ANITA NORLUND *Kritisk sakprosaläsning i gymnasieskolan. Didaktiska perspektiv på läroböcker, lärare och nationella prov.* Göteborg 2009

274. AGNETA SIMEONSDOTTER SVENSSON *Den pedagogiska samlingen i förskoleklasen. Barns olika sätt att erfara och hantera svårigheter.* Göteborg 2009

275. ANITA ERIKSSON *Om teori och praktik i lärarutbildningen. En etnografisk och diskursanalytisk studie.* Göteborg 2009

276. MARIA HJALMARSSON *Lärarprofessionens genusordning. En studie av lärares uppfattningar om arbetsuppgifter, kompetens och förväntningar.* Göteborg 2009.

277. ANNE DRAGEMARK OSCARSON *Self-Assessement of Writing in Learning English as a Foreign Language. A Study at the Upper Secondary School Level.* Göteborg 2009

278. ANNIKA LANTZ-ANDERSSON *Framing in Educational Practices. Learning Activity, Digital Technology and the Logic of Situated Action.* Göteborg 2009

279. RAUNI KARLSSON *Demokratiska värden i förskolebarns vardag.* Göteborg 2009

280. ELISABETH FRANK *Läsförmågan bland 9-10-åringar. Betydelsen av skolklimat, hem- och skolsamverkan, lärarkompetens och elevers hembakgrund.* Göteborg 2009

281. MONICA JOHANSSON *Anpassning och motstånd. En etnografisk studie av gymnasieelevers institutionella identitetsskapande.* Göteborg 2009

282. MONA NILSEN *Food for Thought. Communication and the transformation of work experience in web-based in-service training.* Göteborg 2009

283. INGA WERNERSSON (RED) *Genus i förskola och skola. Förändringar i policy, perspektiv och praktik.* Göteborg 2009

284. SONJA SHERIDAN, INGRID PRAMLING SAMUELSSON & EVA JOHANSSON (RED) *Barns tidiga lärande. En tvärsnittsstudie om förskolan som miljö för barns lärande.* Göteborg 2009

285. MARIE HJALMARSSON *Lojalitet och motstånd - anställdas agerande i ett föränderligt hemtjänstarbete.* Göteborg 2009.

286. ANETTE OLIN *Skolans mötespraktik - en studie om skolutveckling genom yrkesverksammas förståelse.* Göteborg 2009

287. MIRELLA FORSBERG AHLCRONA *Handdockans kommunikativa potential som medierande redskap i förskolan.* Göteborg 2009

288. CLAS OLANDER *Towards an interlanguage of biological evolution: Exploring students´ talk and writing as* an arena for sense-making. Göteborg 2010

Editors: Jan-Eric Gustafsson, Åke Ingerman and
Ingrid Pramling Samuelsson

289. PETER HASSELSKOG *Slöjdlärares förhållningssätt i undervisningen.* Göteborg 2010

290. HILLEVI PRELL *Promoting dietary change. Intervening in school and recognizing health messages in commercials.* Göteborg 2010

291. DAVOUD MASOUMI *Quality Within E-learning in a Cultural Context. The case of Iran.* Göteborg 2010

292. YLVA ODENBRING *Kramar, kategoriseringar och hjälpfröknar. Könskonstruktioner i interaktion i förskola, förskoleklass och skolår ett.* Göteborg 2010

293. ANGELIKA KULLBERG *What is taught and what is learned. Professional insights gained and shared by teachers of mathematics.* Göteborg 2010

294. TORGEIR ALVESTAD *Barnehagens relasjonelle verden - små barn som kompetente aktører i produktive forhandlinger.* Göteborg 2010

295. SYLVI VIGMO *New spaces for Language Learning. A study of student interaction in media production in English.* Göteborg 2010

296. CAROLINE RUNESDOTTER *I otakt med tiden? Folkhögskolorna i ett föränderligt fält.* Göteborg 2010

297. BIRGITTA KULLBERG *En etnografisk studie i en thailändsk grundskola på en ö i södra Thailand. I sökandet efter en framtid då nuet har nog av sitt.* Göteborg 2010

298. GUSTAV LYMER *The work of critique in architectural education.* Göteborg 2010

299. ANETTE HELLMAN *Kan Batman vara rosa? Förhandlingar om pojkighet och normalitet på en förskola.* Göteborg 2010

300. ANNIKA BERGVIKEN-RENSFELDT *Opening higher education. Discursive transformations of distance and higher education government.* Göteborg 2010

301. GETAHUN YACOB ABRAHAM *Education for Democracy? Life Orientation: Lessons on Leadership Qualities and Voting in South African Comprehensive Schools.* Göteborg 2010

302. LENA SJÖBERG *Bäst i klassen? Lärare och elever i svenska och europeiska policytexter.* Göteborg 2011

303. ANNA POST *Nordic stakeholders and sustainable catering.* Göteborg 2011

304. CECILIA KILHAMN *Making Sense of Negative Numbers.* Göteborg 2011

305. ALLAN SVENSSON (RED) *Utvärdering Genom Uppföljning. Longitudinell individforskning under ett halvsekel.* Göteborg 2011

306. NADJA CARLSSON *I kamp med skriftspråket. Vuxenstuderande med läs- och skrivsvårigheter i ett livsvärldsperspektiv.* Göteborg 2011

307. AUD TORILL MELAND *Ansvar for egen læring. Intensjoner og realiteter ved en norsk videregående skole.* Göteborg 2011

308. EVA NYBERG *Folkbildning för demokrati. Colombianska kvinnors perspektiv på kunskap som förändringskraft.* Göteborg 2011

309. SUSANNE THULIN *Lärares tal och barns nyfikenhet. Kommunikation om naturvetenskapliga innehåll i förskolan.* Göteborg 2011

310. LENA FRIDLUND *Interkulturell undervisning – ett pedagogiskt dilemma. Talet om undervisning i svenska som andraspråk och i förberedelseklass.* Göteborg 2011

311. TARJA ALATALO *Skicklig läs- och skrivundervisning i åk 1-3. Om lärares möjligheter och hinder.* Göteborg 2011

312. LISE-LOTTE BJERVÅS *Samtal om barn och pedagogisk dokumentation som bedömningspraktik i förskolan. En diskursanalys.* Göteborg 2011

313. ÅSE HANSSON *Ansvar för matematiklärande. Effekter av undervisningsansvar i det flerspråkiga klassrummet.* Göteborg 2011

314. MARIA REIS *Att ordna, från ordning till ordning. Yngre förskolebarns matematiserande.* Göteborg 2011

315. BENIAMIN KNUTSSON *Curriculum in the Era of Global Development – Historical Legacies and Contemporary Approaches.* Göteborg 2011

316. EVA WEST *Undervisning och lärande i naturvetenskap. Elevers lärande i relation till en forskningsbaserad undervisning om ljud, hörsel och hälsa.* Göteborg 2011

317. SIGNILD RISENFORS *Gymnasieungdomars livstolkande.* Göteborg 2011

318. EVA JOHANSSON & DONNA BERTHELSEN (Ed.) *Spaces for Solidarity and Individualism in Educational Contexts.* Göteborg 2012

319. ALASTAIR HENRY *L3 Motivation.* Göteborg 2012

320. ANN PARINDER *Ungdomars matval – erfarenheter, visioner och miljöargument i eget hushåll.* Göteborg 2012

321. ANNE KULTTI *Flerspråkiga barn i förskolan: Villkor för deltagande och lärande.* Göteborg 2012

322. BO-LENNART EKSTRÖM *Kontroversen om DAMP. En kontroversstudie av vetenskapligt gränsarbete och översättning mellan olika kunskapsparadigm.* Göteborg 2012

323. MUN LING LO *Variation Theory and the Improvement of Teaching and Learning.* Göteborg 2012

324. ULLA ANDRÉN *Self-awareness and self-knowledge in professions. Something we are or a skill we learn.* Göteborg 2012

325. KERSTIN SIGNERT *Variation och invarians i Maria Montessoris sinnestränande materiel.* Göteborg 2012

326. INGEMAR GERRBO *Idén om en skola för alla och specialpedagogisk organisering i praktiken.* Göteborg 2012

327. PATRIK LILJA *Contextualizing inquiry. Negotiations of tasks, tools and actions in an upper secondary classroom.* Göteborg 2012

328. STEFAN JOHANSSON *On the Validity of Reading Assessments: Relationships Between Teacher Judgements, External Tests and Pupil Self-assessments.* Göteborg 2013

329. STEFAN PETTERSSON *Nutrition in Olympic Combat Sports. Elite athletes' dietary intake, hydration status and experiences of weight regulation.* Göteborg 2013

330. LINDA BRADLEY *Language learning and technology – student activities in web-based environments.* Göteborg 2013

331. KALLE JONASSON *Sport Has Never Been Modern.* Göteborg 2013

332. MONICA HARALDSSON STRÄNG *Yngre elevers lärande om natur. En studie av kommunikation om modeller i institutionella kontexter.* Göteborg 2013

333. ANN VALENTIN KVIST *Immigrant Groups and Cognitive Tests – Validity Issues in Relation to Vocational Training.* Göteborg 2013

334. ULRIKA BENNERSTEDT *Knowledge at play. Studies of games as members' matters.* Göteborg 2013

335. EVA ÄRLEMALM-HAGSÉR *Engagerade i världens bästa? Lärande för hållbarhet i förskolan.* Göteborg 2013

336. ANNA-KARIN WYNDHAMN *Tänka fritt, tänka rätt. En studie om värdeöverföring och kritiskt tänkande i gymnasieskolans undervisning.* Göteborg 2013

337. LENA TYRÈN *"Vi får ju inte riktigt förutsättningarna för att genomföra det som vi vill." En studie om lärares möjligheter och hinder till förändring och förbättring i praktiken.* Göteborg 2013

Editors: Jan-Eric Gustafsson, Åke Ingerman and
Pia Williams

338. ANNIKA LILJA *Förtroendefulla relationer mellan lärare och elev.* Göteborg 2013

339. MAGNUS LEVINSSON *Evidens och existens. Evidensbaserad undervisning i ljuset av lärares erfarenheter.* Göteborg 2013

340. ANNELI SCHWARTZ *Pedagogik, plats och prestationer. En etnografisk studie om en skola i förorten.* Göteborg 2013

341. ELISABET ÖHRN och LISBETH LUNDAHL (red) *Kön och karriär i akademin. En studie inom det utbildningsvetenskapliga fältet.* Göteborg 2013

342. RICHARD BALDWIN *Changing practice by reform. The recontextualisation of the Bologna process in teacher education.* Göteborg 2013

343. AGNETA JONSSON *Att skapa läroplan för de yngsta barnen i förskolan. Barns perspektiv och nuets didaktik.* Göteborg 2013

344. MARIA MAGNUSSON *Skylta med kunskap. En studie av hur barn urskiljer grafiska symboler i hem och förskola.* Göteborg 2013

345. ANNA-LENA LILLIESTAM *Aktör och struktur i historieundervisning. Om utveckling av elevers historiska resonerande.* Göteborg 2013

346. KRISTOFFER LARSSON *Kritiskt tänkande i grundskolans samhällskunskap. En fenomenografisk studie om manifesterat kritiskt tänkande i samhällskunskap hos elever i årskurs 9.* Göteborg 2013

347. INGA WERNERSSON och INGEMAR GERRBO (red) *Differentieringens janusansikte. En antologi från Institutionen för pedagogik och specialpedagogik vid Göteborgs universitet.* Göteborg 2013

348. LILL LANGELOTZ *Vad gör en skicklig lärare? En studie om kollegial handledning som utvecklingspraktik.* Göteborg 2014

349. STEINGERDUR OLAFSDOTTIR *Television and food in the lives of young children.* Göteborg 2014

350. ANNA-CARIN RAMSTEN *Kunskaper som byggde folkhemmet. En fallstudie av förutsättningar för lärande vid teknikskiften inom processindustrin.* Göteborg 2014

351. ANNA-CARIN BREDMAR *Lärares arbetsglädje. Betydelsen av emotionell närvaro i det pedagogiska arbetet.* Göteborg 2014

352. ZAHRA BAYATI *"den Andre" i lärarutbildningen. En studie om den rasifierade svenska studentens villkor i globaliseringens tid.* Göteborg 2014

353 ANDERS EKLÖF *Project work, independence and critical thinking.* Göteborg 2014

354 EVA WENNÅS BRANTE *Möte med multimodalt material. Vilken roll spelar dyslexi för uppfattandet av text och bild?* Göteborg 2014

355 MAGNUS FERRY *Idrottsprofilerad utbildning – i spåren av en avreglerad skola.* Göteborg 2014

356 CECILIA THORSEN *Dimensionality and Predictive validity of school grades: The relative influence of cognitive and socialbehavioral aspects.* Göteborg 2014

357 ANN-MARIE ERIKSSON *Formulating knowledge. Engaging with issues of sustainable development through academic writing in engineering education.* Göteborg 2014

358 PÄR RYLANDER *Tränares makt över spelare i lagidrotter: Sett ur French och Ravens maktbasteori.* Göteborg 2014

359 PERNILLA ANDERSSON VARGA *Skrivundervisning i gymnasieskolan. Svenskämnets roll i den sociala reproduktionen.* Göteborg 2014

360 GUNNAR HYLTEGREN *Vaghet och vanmakt - 20 år med kunskapskrav i den svenska skolan.* Göteborg 2014

361 MARIE HEDBERG *Idrotten sätter agendan. En studie av Riksidrottsgymnasietränares handlande utifrån sitt dubbla uppdrag.* Göteborg 2014

362 KARI-ANNE JØRGENSEN *What is going on out there? - What does it mean for children's experiences when the kindergarten is moving their everyday activities into the nature - landscapes and its places?* Göteborg 2014

363 ELISABET ÖHRN och ANN-SOFIE HOLM (red) *Att lyckas i skolan. Om skolprestationer och kön i olika undervisningspraktiker.* Göteborg 2014

364 ILONA RINNE *Pedagogisk takt i betygssamtal. En fenomenologisk hermeneutisk studie av gymnasielärares och elevers förståelse av betyg.* Göteborg 2014

365 MIRANDA ROCKSÉN *Reasoning in a Science Classroom.* Göteborg 2015

366 ANN-CHARLOTTE BIVALL *Helpdesking: Knowing and learning in IT support practices.* Göteborg 2015

367 BIRGITTA BERNE *Naturvetenskap möter etik. En klassrumsstudie av elevers diskussioner om samhällsfrågor relaterade till bioteknik.* Göteborg 2015

368 AIRI BIGSTEN *Fostran i förskolan.* Göteborg 2015

369 MARITA CRONQVIST *Yrkesetik i lärarutbildning - en balanskonst.* Göteborg 2015

370 MARITA LUNDSTRÖM *Förskolebarns strävanden att kommunicera matematik.* Göteborg 2015

371 KRISTINA LANÅ *Makt, kön och diskurser. En etnografisk studie om elevers aktörsskap och positioneringar i undervisningen.* Göteborg 2015

372 MONICA NYVALLER *Pedagogisk utveckling genom kollegial granskning: Fallet Lärande Besök utifrån aktör-nätverksteori.* Göteborg 2015

373 GLENN ØVREVIK KJERLAND *Å lære å undervise i kroppsøving. Design for utvikling av teoribasert undervisning og kritisk refleksjon i kroppsøvingslærerutdanningen.* Göteborg 2015

374 CATARINA ECONOMOU *"I svenska två vågar jag prata mer och så". En didaktisk studie om skolämnet svenska som andraspråk.* Göteborg 2015

375 ANDREAS OTTEMO *Kön, kropp, begär och teknik: Passion och instrumentalitet på två tekniska högskoleprogram.* Göteborg 2015

376 SHRUTI TANEJA JOHANSSON *Autism-in-context. An investigation of schooling of children with a diagnosis of autism in urban India.* Göteborg 2015

377 JAANA NEHEZ *Rektorers praktiker i möte med utvecklingsarbete. Möjligheter och hinder för planerad förändring.* Göteborg 2015

378 OSA LUNDBERG *Mind the Gap – Ethnography about cultural reproduction of difference and disadvantage in urban education.* Göteborg 2015

379 KARIN LAGER *I spänningsfältet mellan kontroll och utveckling. En policystudie av systematiskt kvalitetsarbete i kommunen, förskolan och fritidshemmet.* Göteborg 2015

380 MIKAELA ÅBERG *Doing Project Work. The Interactional Organization of Tasks, Resources, and Instructions.* Göteborg 2015

381 ANN-LOUISE LJUNGBLAD *Takt och hållning - en relationell studie om det oberäkneliga i matematikundervisningen.* Göteborg 2016

382 LINN HÅMAN *Extrem jakt på hälsa. En explorativ studie om ortorexia nervosa.* Göteborg 2016

383 EVA OLSSON *On the impact of extramural English and CLIL on productive vocabulary.* Göteborg 2016

384 JENNIE SIVENBRING *I den betraktades ögon. Ungdomar om bedömning i skolan.* Göteborg 2016

385 PERNILLA LAGERLÖF *Musical play. Children interacting with and around music technology.* Göteborg 2016

386 SUSANNE MECKBACH *Mästarcoacherna. Att bli, vara och utvecklas som tränare inom svensk elitfotboll.* Göteborg 2016

387 LISBETH GYLLANDER TORKILDSEN *Bedömning som gemensam angelägenhet – enkelt i retoriken, svårare i praktiken. Elevers och lärares förståelse och erfarenheter.* Göteborg 2016

388 cancelled

389 PERNILLA HEDSTRÖM *Hälsocoach i skolan. En utvärderande fallstudie av en hälsofrämjande intervention.* Göteborg 2016

Editors: Åke Ingerman, Pia Williams and
Elisabet Öhrn

390 JONNA LARSSON *När fysik blir lärområde i förskolan.* Göteborg 2016

391 EVA M JOHANSSON *Det motsägelsefulla bedömningsuppdraget. En etnografisk studie om bedömning i förskolekontext.* Göteborg 2016

392 MADELEINE LÖWING *Diamant – diagnoser i matematik. Ett kartläggningsmaterial baserat på didaktisk ämnesanalys.* Göteborg 2016

393 JAN BLOMGREN *Den svårfångade motivationen: elever i en digitaliserad lärmiljö.* Göteborg 2016

394 DAVID CARLSSON *Vad är religionslärarkunskap? En diskursanalys av trepartssamtal i lärarutbildningen.* Göteborg 2017

395 EMMA EDSTRAND *Learning to reason in environmental education: Digital tools, access points to knowledge and science literacy.* Göteborg 2017

396 KATHARINA DAHLBÄCK *Svenskämnets estetiska dimensioner - - i klassrum, kursplaner och lärares uppfattningar.* Göteborg 2017

397 K GABRIELLA THORELL *Framåt marsch! – Ridlärarrollen från dåtid till samtid med perspektiv på framtid.* Göteborg 2017

398 RIMMA NYMAN *Interest and Engagement: Perspectives on Mathematics in the Classroom.* Göteborg 2017

399 ANNIKA HELLMAN *Visuella möjlighetsrum. Gymnasieelevers subjektsskapande i bild och medieundervisning.* Göteborg 2017

400 OLA STRANDLER *Performativa lärarpraktiker.* Göteborg 2017

401 AIMEE HALEY *Geographical Mobility of the Tertiary Educated – Perspectives from Education and Social Space.* Göteborg 2017

402 MALIN SVENSSON *Hoppet om en framtidsplats. Asylsökande barn i den svenska skolan.* Göteborg 2017

403 CATARINA ANDISHMAND *Fritidshem eller servicehem? En etnografisk studie av fritidshem i tre socioekonomiskt skilda områden.* Göteborg 2017

404 MONICA VIKNER STAFBERG *Om lärarblivande. En livsvärldsfenomenologisk studie av bildningsgångar in i läraryrket.* Göteborg 2017

405 ANGELICA SIMONSSON *Sexualitet i klassrummet. Språkundervisning, elevsubjektivitet och heteronormativitet.* Göteborg 2017

406 ELIAS JOHANNESSON *The Dynamic Development of Cognitive and Socioemotional Traits and Their Effects on School Grades and Risk of Unemployment.* Göteborg 2017

407 EVA BORGFELDT *"Det kan vara svårt att förklara på rader". Perspektiv på analys och bedömning av multimodal textproduktion i årskurs 3.* Göteborg 2017

408 GÉRALDINE FAUVILLE *Digital technologies as support for learning about the marine environment. Steps toward ocean literacy.* Göteborg 2018

409 CHARLOTT SELLBERG *Training to become a master mariner in a simulator-based environment: The instructors' contributions to professional learning.* Göteborg 2018

410 TUULA MAUNULA *Students' and Teachers' Jointly Constituted Learning Opportunities. The Case of Linear Equations.* Göteborg 2018

411 EMMALEE GISSLEVIK *Education for Sustainable Food Consumption in Home and Consumer Studies.* Göteborg 2018

412 FREDRIK ZIMMERMAN *Det tillåtande och det begränsande. En studie om pojkars syn på studier och ungdomars normer kring maskulinitet.* Göteborg 2018

413 CHRISTER MATTSSON *Extremisten i klassrummet. Perspektiv på skolans förväntade ansvar att förhindra framtida terrorism.* Göteborg 2018

414 HELENA WALLSTRÖM *Gymnasielärares mentorshandlingar. En verksamhetsteoretisk studie om lärararbete i förändring.* Göteborg 2018

415 LENA ECKERHOLM *Lärarperspektiv på läsförståelse. En intervjustudie om undervisning i årskurs 4-6.* Göteborg 2018

416 CHRISTOPHER HOLMBERG *Food, body weight, and health among adolescents in the digital age: An explorative study from a health promotion perspective.* Göteborg 2018

417 MAGNUS KARLSSON *Moraliskt arbete i förskolan. Regler och moralisk ordning i barn-barn och vuxen-barn interaktion.* Göteborg 2018

418 ANDREAS FRÖBERG *Physical Activity among Adolescents in a Swedish Multicultural Area. An Empowerment-Based Health Promotion School Intervention.* Göteborg 2018

419 EWA SKANTZ ÅBERG *Children´s collaborative technology-mediated story making. Instructional challenges in early childhood education.* Göteborg 2018

420 PER NORDÉN *Regnbågsungar: Familj, utbildning, fritid.* Göteborg 2018

421 JENNY RENDAHL *Vem och vad kan man lita på? Ungdomars förhållningssätt till budskap om mat och ätande utifrån ett forskarinitierat rollspel.* Göteborg 2018

422 MARTINA WYSZYNSKA JOHANSSON *Student experience of vocational becoming in upper secondary vocational education and training. Navigating by feedback.* Göteborg 2018

423 MALIN NILSEN *Barns och lärares aktiviteter med datorplattor och appar i förskolan.* Göteborg 2018

424 LINDA BORGER *Investigating and Validating Spoken Interactional Competence: Rater Perspectives on a Swedish National Test of English.* Göteborg 2018