Looking Beyond Scores

A Study of Rater Orientations and Ratings of Speaking

Linda Borger

UNIVERSITY OF GOTHENBURG

DEPARTMENT OF EDUCATION
AND SPECIAL EDUCATION

Licentiate thesis in Subject Matter Education at the Department of Education and Special Education, Faculty of Education, University of Gothenburg.

The licentiate thesis is available for full text download at Gothenburg University Publications Electronic Archive (GUPEA):
http://hdl.handle.net/2077/38158

This licentiate thesis has been carried out within the framework of the Graduate School in Foreign Language Education "De främmande språkens didaktik" (FRAM). The Graduate School, leading to a licentiate degree, is a collaboration between the universities of Gothenburg, Lund, Stockholm and Linnaeus University, and is funded by the Swedish Research Council (project number 729-2011-5277).

Abstract

Title: Looking Beyond Scores – A Study of Rater Orientations and Ratings of Speaking

Author: Linda Borger

Language: English with a Swedish summary

GUPEA: http://hdl.handle.net/2077/38158

Keywords: Performance assessment, paired speaking test, rater orientations, rater variability, inter-rater reliability, The Common European Framework of Reference for Languages (CEFR), Swedish national tests of English

The present study aims to examine rater behaviour and rater orientations across two groups of raters evaluating oral proficiency in a paired speaking test, part of a mandatory Swedish national test of English. Six authentic conversations were rated by (1) a group of Swedish teachers of English (*n* = 17), using national performance standards, and (2) a group of external raters (*n* = 14), using scales from the Common European Framework of Reference for Languages (CEFR), the latter to enable a tentative comparison between the Swedish foreign language syllabus for English and the CEFR.

Raters provided scores and written comments regarding features of the performances that contributed to their judgement. Statistical analyses of the Swedish raters' scores show reasonable degrees of variability and, in general, acceptable inter-rater reliabilities, albeit with obvious room for improvement. In addition, the CEFR raters judged the performances of the Swedish students to be, on average, at the intended levels of the test. Analyses of the written comments, using NVivo 10 software, show that raters took a wide array of features into account in their holistic rating decision, however with test-takers' linguistic and pragmatic competences, and interaction strategies the most salient. Raters also seemed to heed the same features, indicating considerable agreement regarding the construct. Further, a tentative comparison of the written comments and scores shows that the raters noticed fairly similar features across proficiency levels but in some cases evaluated them differently. The findings of the present study have implications for the interpretation of oral test results, and they also provide information that may be useful in the development of tasks and guidelines for different types of oral language assessment in different educational settings.

Table of contents

List of Figures

List of Tables

Acknowledgements

I am very grateful to many people who have contributed to this licentiate thesis. First, I would like to thank my supervisor, Gudrun Erickson. Thank you for your generous and insightful supervision and invaluable guidance throughout this work. I would also like to thank my co-supervisor, Liss Kerstin Sylvén. Thank you for your positive encouragement, valuable advice and constructive suggestions on the research work.

I would additionally like to thank April Ginther, Philip Shaw and Lisbeth Åberg-Bengtsson for their wise comments on draft versions of this thesis. I am particularly grateful for the valuable assistance in the research process given by Lena Börjesson. I would also like to acknowledge my gratitude to Sölve Ohlander for the helpful comments and suggestions for improvements to this thesis at the final stage.

Further, I wish to express a sincere thank you to the 31 participating raters for their contribution to this research. Without you I would not have been able to write this thesis.

I have also drawn great benefit from being part of the Swedish national graduate school for language education (FRAM). I am thankful to all my colleagues in the graduate school, both supervisors and fellow students. Thank you all for the highly constructive seminars and fruitful discussions. A special thanks goes to fellow student Lisa Källermark Haya who read and commented on my manuscript at our last seminar.

I would like to express my warm thanks to Petra Comstedt at Realgymnasiet, Linköping, Margita Edström at Lärande, and Magnus Nyström at Katedralskolan, Linköping, for the support provided.

Finally, I would like to express my gratitude to my family. I am especially grateful to my mother and my grandmother for their unconditional support and encouragement. Nils, my love, and Gustav and Fredrik, our wonderful two boys, thank you for believing in me and supporting me throughout this work.

Linköping, November, 2014
Linda Borger

Chapter One: Introduction

Language assessment[1] is a complex and important aspect of the language teaching profession. Furthermore, assessment is inherently linked to learning and teaching. Being a language teacher myself, I have come to take a special interest in language assessment, and especially issues regarding validity and reliability of performance assessment. Performance assessment involves test-takers in tasks that are designed to be as close to real-life situations as possible, and is often used to assess speaking skills, for example in the paired speaking test format. I am interested in exploring the paired speaking test format with regard to three main issues: (1) agreement between raters, (2) features that draw raters' attention when evaluating test-taker performance, and (3) whether different features are more or less salient.

A concern for foreign language (FL)[2] or second language (L2) performance tests is the potential variability of rater judgements. The terms *rater variability* and *rater effects* are used to refer to variation in scores that can be attributed to rater characteristics rather than test-takers' actual language performance or ability (McNamara, 1996). These rater effects influence the validity and reliability of scores (Messick, 1989) and are therefore important to explore.

One of the most prevalent rater effects in performance testing is *rater severity/leniency*. This is when raters award scores that are consistently too harsh or too lenient in comparison to other raters (Bachman, Lynch, & Mason, 1995; McNamara, 1996). There are several other factors that have an impact on the ratings of performance tests. For example, raters may apply and interpret assessment criteria in different ways. They may also weight specific features of the performance differently, thus awarding different scores for the same performance or conversely, the same score but for different reasons (McNamara, 1996). Secondly, rater background variables, such as their first

[1] The terminology *assessment* and *testing* is 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 the performance of the test-taker is measured and evaluated.

[2] 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.

language (Chalhoub-Deville, 1995; J. S. Johnson & Lim, 2009; Kim, 2009), their professional background (Anne Brown, 1995; Chalhoub-Deville, 1995; Hadden, 1991), and their rating experience (Cumming, 1990; Weigle, 1994, 1999), may also influence rater judgements.

Bearing in mind that rater-related variability is impossible to eliminate in performance testing, research that addresses the issue of raters' judgements of test-taker performance is crucial in order to gain a deeper understanding of the nature of rater differences. Studies that explore *rater effects*, such as severity and leniency, as well as *rater orientations*, i.e. features of the performance that raters attend to in forming their judgement, thus make an important contribution to this field. Results of such research may also have didactic implications for raters and teachers.

The present study aims to explore the rating of speaking across two groups of raters evaluating oral proficiency in a paired speaking test, part of a mandatory Swedish national test of English as a Foreign Language (EFL) at upper secondary level. Research into the paired speaking test format (or group speaking test, if there are more than two participants) can broadly be divided into three main categories: (1) features of test-taker interaction (2) effects of background variables of test-takers (so-called *interlocutor effects*) and (3) raters' and test-takers' perspectives (Galaczi, 2010). This investigation focuses on the raters' perspective. More specifically, two main areas were examined: variability of rater judgements and raters' decision-making processes. In addition, a small-scale, tentative comparison of the Swedish performance standards for English and the corresponding reference levels from the Common European Framework of Reference for Languages (Council of Europe, 2001) was made.

Background

In this section, a short background is given to the Swedish school system, in which great trust is placed on teachers' assessment of students' competences. After that, the Swedish national tests of English are described. Finally, the Common European Framework of Reference for Languages (CEFR) is briefly presented. The CEFR is explicitly related to the Swedish syllabus for foreign languages and is used by one of the rater groups in the present study.

The Swedish context

In Sweden, teachers have great responsibility with regard to assessment and grading. In the Swedish school system there are no external examinations and final grades are assigned exclusively by the students' own teachers. However, there are national tests at different levels and in different subjects to help teachers make decisions about individual students' achievements in relation to national objectives and performance standards. The national tests thus have an advisory rather than decisive function (Erickson, 2010a). Furthermore, there is no central marking of the national tests; they are marked by the students' own teachers. The main aim of the national tests is to enhance equity and comparability within the Swedish school system, but they are also regarded as a means to make the content of the national curricula and syllabuses more concrete (Erickson, 2012). The national tests are compulsory and are therefore viewed as high stakes by both teachers and students.

During a period of three years, 2009-2012, the Swedish Schools Inspectorate (SSI), commissioned by the Swedish government, has performed a re-marking of national tests in English, Swedish and Mathematics from compulsory and secondary level. Results have been published gradually, and in August 2012 a summary report was issued (The Swedish Schools Inspectorate, 2012), showing that there are considerable discrepancies between the re-marking by the SSI and the original marking by teachers. The SSI concluded that inter-rater reliability was low for those parts of the national tests with open-ended responses, such as essays, and that the teachers were generally more generous in their marking than the external raters.

Inter-rater reliability proved to be higher for the receptive skills involving English reading and listening comprehension and for the test in Mathematics, whereas the essay in the Swedish test had lower reliability (SSI, 2012). However, there is also criticism of the methodology used by the SSI; Gustafsson and Erickson (2013) for example, have discussed and questioned the re-marking procedures used and conclusions drawn.

The SSI has not re-marked the oral parts of the national tests, since recording is not mandatory and a random sample thus not possible to collect. The fact that speaking tests are not explored to the same extent as written tests is one of the reasons why it is interesting and important to examine the rating of oral proficiency in high-stakes testing.

National tests of English

The Swedish National Agency for Education (NAE) has commissioned the responsibility for national test development to different Swedish universities. The University of Gothenburg, Department of Education and Special Education, is responsible for developing the national tests and assessment materials for foreign languages – English, French, German and Spanish. In accordance with the national syllabuses, the ambition is to have a broad representation of the construct of English language proficiency. Consequently, there are different kinds of tasks in the test that are designed to be as authentic as possible.

The Swedish national tests of English focus on three broad language activities, namely reception, production and interaction. They typically comprise three subtests, involving (1) *receptive skills* in the form of listening and reading comprehension, (2) *written production and interaction* in the form of an essay, and (3) *oral production and interaction* in the form of a paired conversation. For all parts there are teacher guidelines, including test specifications, answers with comments, and authentic benchmarked examples of oral and written performance (Erickson, 2012). The speaking test is a performance-based test in which groups of two or three students discuss a given theme.[3] The speaking test focuses on both oral production and interaction (further information in Chapter Four: Material and Method).

The national tests of foreign languages are developed and designed in a collaborative process including teachers, researchers and students, as described in Erickson and Åberg-Bengtsson (2012). The collaborative approach is intended to have a positive effect on the validity of the test. The reason for this is that different stakeholders, i.e. people who are affected by the interpretation and use of the result, are involved in the design of the assessment. To sum up, the Swedish national tests of foreign languages are developed in a collaborative way that ensures that all tasks included in official tests have been reviewed by teachers, researchers and several hundred students in the relevant age group.

[3] However not the focal point of the current study, it should be mentioned that the oral component of the Swedish national tests of EFL was developed in the late 1980s and early 1990s; work documented, for example, in Erickson (1991), Lindblad (1992) and Sundh (2003).

The Common European Framework of Reference for Languages

The Common European Framework of Reference for Languages: Learning, Teaching and Assessment (CEFR) was published by the Council of Europe in 2001 and is based on over twenty years of research. It has been developed to provide help and guidance for assessment of foreign languages, as well as development of language syllabuses and curricula, and also teaching and learning materials. It is used in European countries as well as on other continents and has currently (2014) been translated into 38 languages.

One of the main purposes of the CEFR is to promote international co-operation and enable better communication between professionals who are working in the field of foreign languages and who come from different educational systems in Europe. The CEFR is intended to provide "a common basis for the explicit description of objectives, content and methods" (Council of Europe, 2001, p. 1). This common basis increases the transparency and comparability of curricula, syllabuses and qualifications, and helps to promote a shared recognition of language qualifications.

It is emphasised that in order to be comprehensive, the CEFR needs to be based on a general understanding of language learning and use. The CEFR has adopted an action-oriented approach, which means that it sees all language learners and users as 'social agents'. Language learning, including language use, is described in the following way:

> Language use, embracing language learning, comprises the actions performed by persons who as individuals and as social agents develop a range of competences, both general and in particular communicative language competences. They draw on the competences at their disposal in various contexts under various conditions and under various constraints to engage in language activities involving language processes to produce and/or receive texts in relation to themes in specific domains, activating those strategies which seem most appropriate for carrying out the tasks to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences.
>
> (Council of Europe, 2001, p. 9)

The CEFR is a comprehensive document with an ambition to encompass aspects of learning, teaching and assessment. However, it is probably best known for its common reference levels and illustrative scales. To begin with, six levels of foreign language proficiency are outlined: A1, A2, B1, B2, C1 and

C2. In addition, there are three so-called 'plus' levels: A2+, B1+ and B2+. Level A means basic user, level B independent user and level C proficient user. The first two scales in the CEFR describe the common reference levels on a global scale and a self-assessment scale (Council of Europe, 2001, pp. 24-27). The global scale "will make it easier to communicate the system to non-specialist users and will also provide teachers and curriculum planners with orientation points" (Council of Europe, 2001, p. 24). In comparison, the self-assessment scale is "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" (Council of Europe, 2001, p. 25). The self-assessment grid is used in the European Language Portfolio (ELP), developed for pedagogical purposes (Little, 2009).

In addition to the global scale and the self-assessment grid, the CEFR provides illustrative scales with "can-do" descriptors[4] for (a) communicative language activities, (b) strategies, and (c) communicative language competence. The communicative language activities include *reception* (listening and reading), *production* (spoken and written), *interaction* (spoken and written), and *mediation* (translating and interpreting). There are scales that describe, for example, oral production, written production, listening, reading, spoken interaction, written interaction, note-taking, and processing text. Furthermore, can-do descriptors are provided for *strategies*, which are used in performing communicative activities. Strategies are described as a hinge between the language learner's communicative competences and what he/she can do with these communicative activities. An example of a strategy is *monitoring and repair*, which means that the language learner can recognise his/her own mistakes and correct them, while for example speaking. Finally, scaled descriptors are provided for the communicative language competences described in the CEFR, namely pragmatic competence, linguistic competence and sociolinguistic competence (see Chapter Two: Conceptual Framework, section on Communicative competence). The levels of language proficiency are based on empirical research and consultation from experts and are intended for use in the comparison of tests and examinations in different languages and countries.

With regard to the Swedish context, the syllabuses for foreign languages are explicitly related to the CEFR. For example, just as in the CEFR descriptors, the performance standards are written as can-do statements. Furthermore, the

[4] Performance level descriptors explain the skills a test-taker should be able to demonstrate at different performance levels of the rating scale.

language activities defined in the CEFR – *reception*, *production* and *interaction* – are used in the terminology of the syllabuses of foreign languages (Börjesson, 2012). Only one of the four language activities, namely *mediation* (translating and interpreting), is not included in the Swedish syllabus for English, unlike many other countries. Finally, the action-oriented and communicative approach to language learning, teaching and assessment expressed in the CEFR also forms the basis of the Swedish foreign language curriculum and has done so since the 1980s.

Aim and research questions

Considering the potential variability of rater judgements in performance testing, it is interesting to study how raters reach their decisions. It is especially important to investigate variability due to rater characteristics in high-stakes testing situations, since these results have important consequences for test-takers. The present study thus aims to explore the rating of oral proficiency in a high-stakes paired speaking test. Six recorded paired conversations, authentic material from a Swedish national test of English for upper secondary level, were rated by (1) a group of Swedish teachers of English (n = 17), and (2) a group of external CEFR raters from Finland and Spain (n = 14). Raters provided scores and concurrent written comments to justify their rating decisions.

The first aim was to examine variability of rater judgements and consistency of rater behaviour. The second aim was to explore raters' decision-making processes by identifying and comparing rater orientations, i.e. features that attracted raters' attention as they judged the oral performances of the test-takers. In addition, these two aims were combined in an attempt to explore the relationship between scores and raters' justifications of these scores. Finally, a subordinate aim was to make a small-scale, tentative comparison of Swedish performance standards for EFL and CEFR levels.

In particular, then, the study aims to address the following research questions:

1. What can be noticed regarding variability of scores and consistency of rater behaviour?
2. What features of test-taker performance are salient to raters as they make their decisions?

3. What is the possible relationship between scores and raters' justifications of these scores?
4. At what levels in the CEFR do external raters judge the performances of the Swedish students to be?

Chapter Two: Conceptual Framework

In this chapter, a conceptual framework is outlined, comprising three parts. Firstly, theoretical considerations and descriptions of language assessment in general are given. Secondly, the development of the communicative language testing approach and the concept of communicative competence, as well as performance assessment, are described. Finally, theories of assessment of oral proficiency are presented.

Validity and reliability

According to Bachman (1990), the main concern of test development and use is not only to provide evidence that test scores are reliable, but also that interpretations and inferences made from test scores are valid. The concept of reliability refers to consistency of scores, whereas validity refers to the extent to which a test actually measures what it intends to measure.

In language testing, scores should accurately reflect a test-taker's language ability in a specific area, for example writing an argumentative essay or giving an informative speech. In order to base interpretations about language ability on a candidate's performance in a language test, language ability has to be defined in a way that is appropriate for a specific assessment situation. This is normally referred to as *construct*. In simpler terms, *construct* might be described as "the *what* of language testing" (Weir, 2005, p. 1). Consequently, the construct definition of a specific assessment task or situation governs what kinds of inferences can be made from the performance.

The assessment results must be valid indicators of the construct, and should therefore lead to adequate interpretations and conclusions. Bachman (1990) claims that validity is the most important aspect of the interpretation and use of test results. Similarly, Messick (1996) emphasises that validity "is not a property of the test or assessment as such, but rather of the meaning of test scores" (p. 245). As a result, it is not the test that should be validated but the inferences drawn from test scores and the consequences they may have.

To make sure a test score is a meaningful indicator of a test-taker's language ability, we must ascertain that it actually measures this language ability and not some other aspects. Thus, to evaluate the meaningfulness of test scores, we must provide evidence that they are not unduly affected by aspects other than the ability that the test is intended to measure. Messick (1989) described two major threats to construct validity: *construct underrepresentation* and *construct irrelevant variance*. 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 would be rater effects, i.e. variation in scores that can be attributed to rater characteristics and not to test-takers' actual language performance or ability. Both types exist in all assessments. Consequently, in all test validation, convincing arguments need to be presented in order to refute these threats.

As mentioned above, in addition to being valid, it is necessary, but not sufficient, that the test scores are reliable. Reliability has to do with the "quality of test scores themselves" (Bachman, 1990, p. 25) and whether they are consistent or not. Put more simply, this means that a test would generate similar results if it were to be given at another time. An example of this would be that if a test were to be administered to the same group of students but on two different occasions and settings, it would not make any difference to the test-taker if he/she takes the test on one occasion or in one setting rather than another. Moreover, this means that if two versions of a test are used interchangeably, it would not make any difference to the test-taker which version of these two tests he/she takes.

Bachman (1990) points out that neither reliability nor validity is absolute, since it is almost impossible to achieve measures that are free of errors in practice, and there are many factors outside the test itself that determine how appropriate the interpretation and use of a test score are in a given situation. In a perfectly reliable score, there would be no measurement errors. However, in addition to the language ability measured, there are many other factors that could affect the performance on a test and lead to possible sources of measurement errors. Such factors could be anxiety, fatigue and the conditions around the testing situation, such as the location and the time. As mentioned above, there is also the factor of rater variability. For example, two raters might assign different scores to the same language performance. It is thus easy to see that there are sources of measurement errors in all test situations.

Language assessment

Assessment of language requires (1) a clear definition of the construct, and (2) a procedure through which the language performance can be elicited, i.e. a method. Furthermore, assessment is a process that involves collecting information about something that we find interesting, using systematic and well-grounded procedures (Bachman & Palmer, 2010). The assessment is the result of this process, usually a score. In language assessment the information we are interested in collecting is, of course, students' language ability. In other words, the main purpose of language assessment is to gather information about specific aspects of the test-taker's language ability in order to make decisions about the overall language performance. The results of the assessment can then be interpreted as an indicator of the construct that is measured.

In language assessment, language skills are usually divided into different skills or abilities. For example, a distinction is made between oral and literate abilities, which can also be expressed in terms of *oracy* and *literacy* (Cumming, 2008). *Oracy* means listening and speaking and *literacy* means reading or writing. In addition, distinctions are made between *reception*, i.e. reading and listening, and *production*, i.e. writing and speaking. This model is used in the CEFR. Furthermore, each skill domain is divided into subcomponents. For example, speaking can be assessed in terms of the subcomponents of pronunciation, fluency, grammar, etc.

The convention in language assessment has been to assess the four skills reading, writing, listening and speaking separately (Purpura, 2008). Scores are then reported for each of the skills or aggregated as a total score. This tradition comes from the approach of descriptive and structural linguists such as Lado (1961) who formulated principles for the design of language testing in the 1960s. The demarcation of the four skills has been influential in language education and assessment throughout the world.

There have been challenges to the "four skills" model, especially in the 1980s when new models of communicative competence were developed (Harley, 1990). As a result, a broad set of standards in reading, writing, listening and speaking is used as the primary basis in curricula as well as testing and assessment in most educational systems today. These standards are in turn usually divided into proficiency levels (Fulcher, 2008).

Communicative language assessment

Historically, language testing and theory have followed the trends in teaching methodology. In the 1940s and 1950s, behavioural psychology and structural linguistics were the main influences on language testing and teaching. In this era, discrete-point test formats were dominant, i.e. individual or detached items without [extensive] context (Oller, 1973). Such tests are based on an analytic view of language and are developed to test separate units of language (discrete points), such as morphology, syntax, phonology, and lexicon. The focus of language assessment in those days was on issues of validity, reliability and objectivity (H. D. Brown & Abeywickrama, 2010).

In the 1970s and 1980s, however, communicative theories of language influenced both language testing and teaching. The communicative approach criticised discrete-point tests for being decontextualized and inauthentic. Instead, communication, authenticity, and context were highlighted as important features of language testing. A first step was integrative testing, mainly consisting of cloze tests[5] and dictation, which were considered to be good examples of integrated skills. A second step was taken when communicative language testing tasks were being developed after theories of communicative competence had become influential in the 1980s. Such tests were based on real-world tasks that test-takers were asked to perform.

Today, the communicative approach to language testing has become the norm. In a communicative language test, language is assessed in context and tasks should be as authentic as possible and usually involve interaction (Davies et al., 1999). Thus, the goal of communicative language tests is to measure language learners' ability to take part in acts of communication in real-life situations.

Communicative language tests cover the four skills (often tested in combination): reading, listening, writing and speaking, as well as the interaction between "speakers and listeners, texts and their readers" (Kramsch, 2006, p. 250). In tests that measure productive skills (writing and speaking), the focus is on how appropriately language learners use the language rather than how well they form linguistically correct sentences. In testing receptive skills (listening and reading), focus is on understanding the communicative intent of the speaker or writer rather than focusing on specific details, such as individual words. Very often, the two are combined so that the learner must both

[5] A cloze test consists of a text with certain words removed, i.e. gaps, which the test-taker is asked to fill.

comprehend and respond in a real-life situation. For example, students can listen to a lecture and then use the information from the lecture to write an essay.

Communicative competence

Communicative language tests are designed on the basis of communicative competence. The term was introduced in L2 and FL discussions in the early 1970s (Habermas, 1970; Hymes, 1971; Jakobovits, 1970; Savignon, 1972). The term *communicative competence* can be understood as "competence to communicate". *Competence* is a controversial term in general and applied linguistics, having its origin in both psycholinguistic and sociocultural perspectives. The introduction of this term in linguistics is usually associated with Chomsky's (1965) influential book *Aspects of the Theory of Syntax*, where he introduced his classic distinction between *competence,* defined as native speakers' tacit knowledge of their language, and *performance,* defined as the realisation of this knowledge in concrete utterances, i.e. the actual use of language in real-life situations. This is similar – although not identical – to Saussure's (1959) distinction between *la langue* (roughly corresponding to competence) and *la parole* (roughly corresponding to performance).

Chomsky's concept of linguistic competence as a theoretical basis for a methodology for learning, teaching and testing languages was soon opposed by advocates of a communicative view of language, such as Savignon (1972). An alternative to Chomsky's concept of competence was found in Dell Hymes's (1972) definition of *communicative competence*, which was considered both a broader and a more realistic notion of competence. In Hymes's definition of communicative competence, the term is viewed not only as consisting of a speaker's purely linguistic, or grammatical competence, but also as the speaker's ability to use this knowledge *appropriately* in social contexts, thus adding a sociolinguistic and pragmatic discussion to Chomsky's notion of competence. Communicative knowledge is thus divided into two components: *grammatical competence* and *sociolinguistic competence*. Furthermore, actual performance is separated from communicative competence and refers to the actual use of language in concrete situations. In Figure 1, Hymes's model of communicative competence is presented.

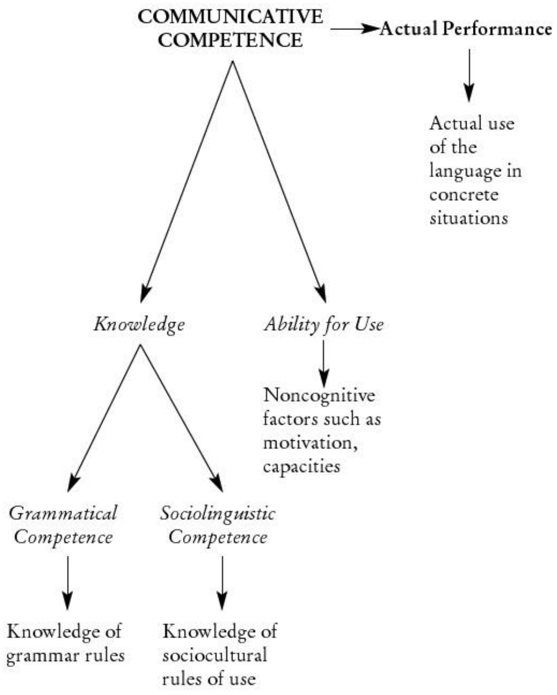

COMMUNICATIVE COMPETENCE
Actual Performance
Actual use of the language in concrete situations
Knowledge
Ability for Use
Noncognitive factors such as motivation, capacities
Grammatical Competence
Sociolinguistic Competence
Knowledge of grammar rules
Knowledge of sociocultural rules of use

Figure 1. Hymes's (1972) model of communicative competence

(Source: Johnson, 2001, p. 157)

In their landmark publication "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing", Canale and Swain (1980) provided the communicative approach with its first comprehensive model of communicative competence. It was developed for both instructional and assessment purposes and has been very influential in second language teaching and testing. Canale and Swain drew on Hymes (1972) in creating their model, which involved three components of communicative competence: (1) grammatical competence (2) sociolinguistic competence, and (3) strategic competence. Canale (1983) later expanded this model by adding a fourth component, namely discourse competence, which was part of sociolinguistic competence in the first model.

Grammatical knowledge is mainly defined in the same way as Chomsky's definition of linguistic competence, and includes "knowledge of lexical items and of rules of morphology, syntax, sentence-grammar semantics, and phonology" (Canale & Swain, 1980, p. 29). In line with Hymes's discussion about the appropriateness of language use in different social situations, sociolinguistic competence in Canale and Swain's model comprises knowledge of "sociocultural rules of use and rules of discourse" (p. 30). Strategic competence, finally, is "made up of verbal and nonverbal communication strategies that may be called into action to compensate for breakdown in communication due to performance variables or to insufficient competence" (p. 30). In Figure 2 below, a figure of Canale and Swain's model of communicative competence, updated by Canale (1983), is presented.

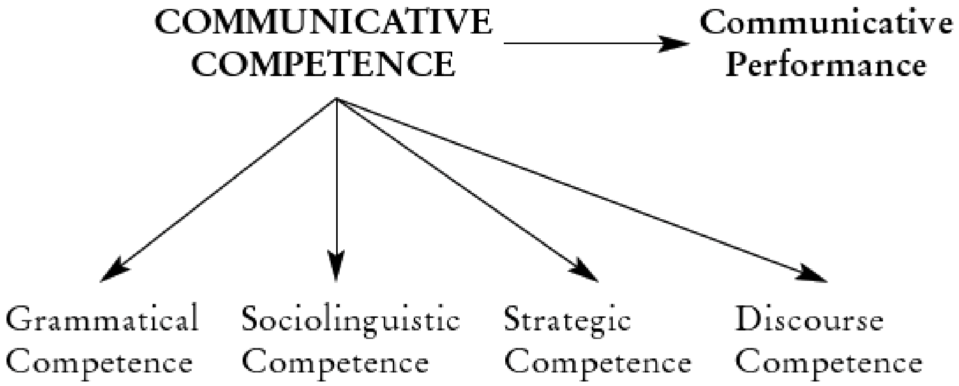


Figure 2. Canale and Swain's (1980) model of communicative competence, updated by Canale (1983)

(Source: Johnson, 2001, p. 159)

In 1990, Bachman presented an elaboration of Canale and Swain's model in his influential work *Fundamental Considerations in Language Testing*. Bachman used a wider term than communicative competence, namely *communicative language ability* (CLA), claiming that this term comprises both the meaning of language proficiency and communicative competence. The CLA model was developed further in Bachman and Palmer (1996).

In the Bachman and Palmer model, language ability comprises two main components: language knowledge and strategic competence. However, the authors stress that there are also many attributes of language users and test-takers, such as "personal attributes, topical knowledge, affective schemata, and cognitive strategies" (p. 33), that need to be taken into consideration in language assessment since they affect both language use and test-taker performance.

Language knowledge is divided into two main components: (1) organisational knowledge, and (2) pragmatic knowledge. These two components complement each other in achieving effective communication. Organisational knowledge comprises abilities involved in the control of formal language structures, i.e. grammatical and textual knowledge. Pragmatic knowledge comprises abilities that are used to create and interpret language. It is divided into two areas: functional knowledge and sociolinguistic knowledge. In Figure 3, Bachman and Palmer's model of language knowledge is presented. It should be noted that strategic competence (not included in Figure 3) refers to non-linguistic cognitive skills in language learning, which are used to achieve communicative goals, such as assessing, planning and executing. Thus, strategic competence is defined in a different way in comparison to Canale and Swain (1980).

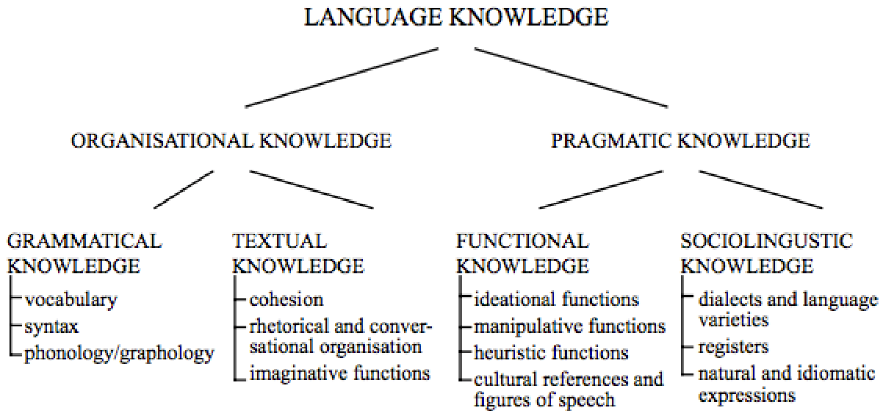


Figure 3. Areas of language knowledge (Bachman & Palmer, 1996)

(Source: Bachman and Palmer, 1996, p. 68)

The last model in this survey is the description of communicative language competence in the CEFR (Council of Europe, 2001). This model was developed for assessment as well as for learning and teaching purposes. It is also the model used by the raters in this study. In the CEFR, communicative competence is divided into three main components: *linguistic*, *sociolinguistic* and *pragmatic*. Each component of language knowledge is defined as both knowledge of and ability to use it.

Linguistic competence, for instance, applies to both knowledge of and skills to use language resources in effective communication. There are several subcategories of linguistic competence, for example lexical, grammatical, semantic, and phonological competences. Sociolinguistic competence refers to knowledge and skills of how to use language appropriately in a social context. The last component, pragmatic competence, comprises two subcategories: discourse competence, involving knowledge and skills of coherence and cohesion, and functional competence, involving knowledge and skills necessary for functional communication purposes, for example fluency.

As can be seen, strategic competence is not a componential part of this communicative model. Instead strategic competence is referred to as *production strategies*, which are used as a balance between the competences. Production strategies involve abilities such as planning, compensating, and monitoring and repair, and can thus be seen as different types of communication startegies.

In Bagarić and Mihaljević Djigunović (2007), a graphic illustration of the similarities and differences in the componential structure of the four models described above is presented (See Figure 4 below). Okvir is the Croatian name for the CEFR, which was translated into Croatian in 2005.

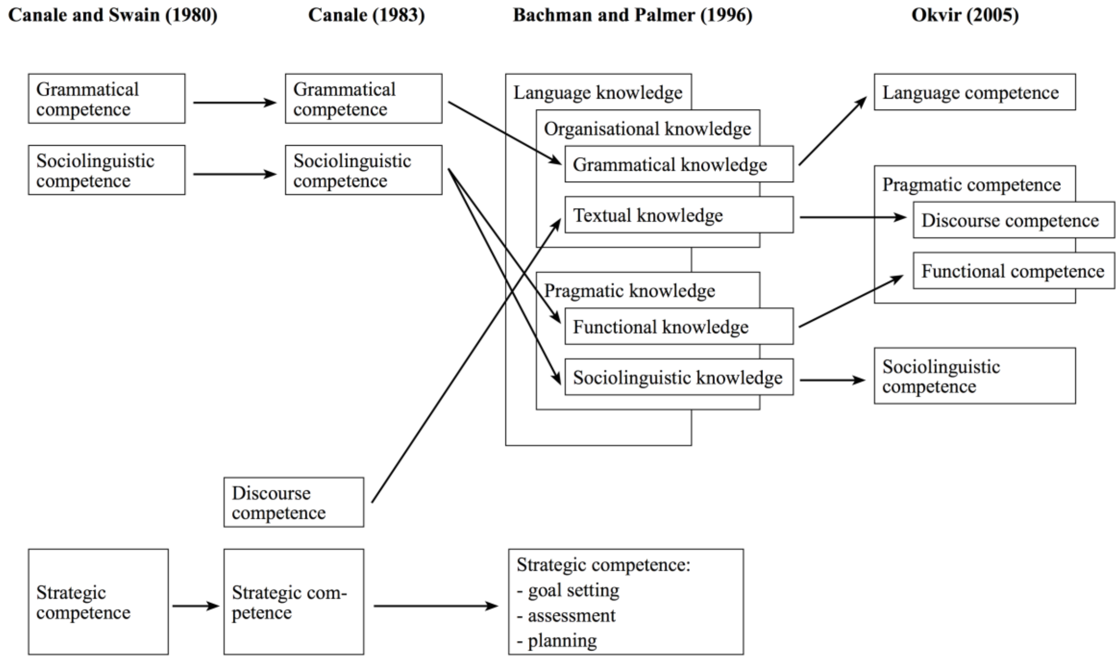


Figure 4. Similarities and differences between models of communicative competence.

(Source: Bagarić & Mihaljević Djigunović, 2007, p. 102)

To summarise, the theoretical models of communicative competence, or communicative language ability, outlined in this survey are largely based on Hymes's (1971, 1972) theory of language use in social context. As can be seen in Figure 4, the similarities between the four models are obvious, with Bachman and Palmer's model being the most highly detailed and complex one.

Challenges for communicative language testing

Despite their wide use in language testing, there are challenges to the theoretical models of communicative competence. A general question that has been posed is how, given the complexity of various models of communicative competence, test developers can make practical use of them. For instance, McNamara (1996) states that theoretical models may be difficult to apply to performance testing, because the scoring rubric is too broad and raters might find one component more important than another (e.g. grammatical competence versus pragmatic competence).

Moreover, McNamara (1995) evaluates the models by Canale and Swain and Bachman and Palmer and points to some problematic features. For example, McNamara argues that the different aspects of performance need to be expanded to include interactions that performance tests usually involve. He gives the example of speaking tests, where the candidate's performance may be affected by interaction effects, such as whom the candidate is paired up with. McNamara underlines that the potential variability is huge in "interactions between candidate and other individuals (including, of course, the judge) and non-human features of the test setting (materials, location, time, etc.)" (p. 173).

In addition, McNamara claims another weakness of the models of communicative competence is that they focus too much on the individual candidate instead of the individual in interaction. Communicative models should therefore incorporate features of social interaction as described in, for example, the discussion of co-construction by Kramsch (1986) and Jacoby and Ochs (1995), building on research from different disciplinary perspectives such as applied linguistics, conversational analysis, ethnomethodology and linguistic anthropology.

Another criticism is put forward in Harding (2014), who refers to difficulty in using the complex frameworks of communicative competence. The solution has been that language test developers "tend to be reliant on frameworks which have been designed to "unpack" existing models of communicative language ability. The CEFR is currently playing this role across many contexts as an accessible de facto theory of communicative language ability /.../" (p. 191).

Performance assessment

Performance assessment is short for the longer term “performance and product evaluation”. In brief, performance assessment requires students to show their language skills in practice by performing or producing something in an authentic or real-life situation. It has a long tradition and is used in applied linguistics as well as in other fields (McNamara, 1996). In second and foreign language testing, performance assessment has been used for about half a century both to assess language skills for a specific workplace and in educational contexts (Wigglesworth, 2008). According to the *Dictionary of Language Testing* a performance test is “a test in which the ability of candidates to perform particular tasks, usually associated with job or study requirements, is assessed” (Annie Brown & Davies, 1999, p. 144). The typical feature of performance assessment is that candidates perform relevant tasks, rather than showing more abstract knowledge as in the traditional *fixed response assessment*[6] (McNamara, 1996). In *fixed response testing*, there is interaction between only the candidate and the test instrument. In performance-based testing, on the other hand, interactions are more complex. An additional component is added: a rater who assesses test-taker performance according to a rating scale. In oral interviews and in the paired oral, a further interaction is introduced in the form of the interlocutor (the examiner in the interview and the other candidate in the paired oral). Figure 5 below illustrates these interactions in performance assessment.

[6] Fixed response assessment refers to test items where typically there is a right and wrong answer, such as the multiple-choice format, or true/false questions. Test-takers do not construct an answer. Instead, they usually choose from options already provided. The opposite test format, which incorporates performance testing, is called constructed response.

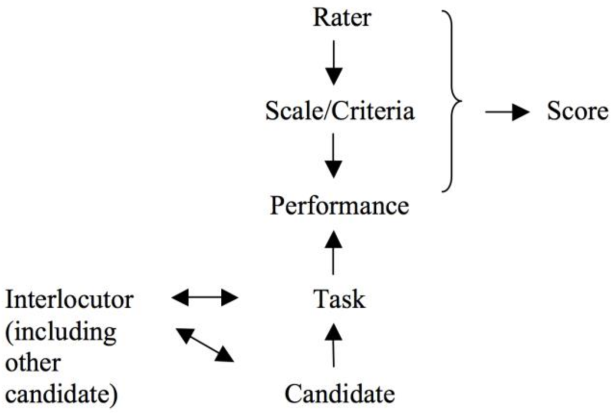


Figure 5. Interactions in performance assessment of speaking skills

(Source: McNamara, 1995, p. 173)

There are two definitions of performance tests: a *narrow*, or *strong* sense; or a *broad*, or *weak sense* (Haertel, 1992). The narrow definition is that a performance test is "any test in which the stimuli presented or the response elicited emulate some aspects of the nontest settings" (p. 984). In other words, the focus is on examinees' task completion. The new theories of communicative competence and communicative language ability presented in the 1980s and 1990s led not only to a new view of second language ability, but also changed the role of performance in language testing. The new communicative language testers supported a *broad*, or *weak sense*, of performance assessment, in which the main focus was on test-takers' language ability as opposed to task completion. This means that second language ability was measured in relation to various language components derived from the theoretical models of communicative competence and communicative language ability. One example is writing assignments, where the purpose is for the students to demonstrate their writing proficiency and where, therefore, duplicating tasks from reality may be unnecessary.

McNamara (1996) states that performance assessments always include subjective evaluations, since it is complex to evaluate human performance. Performance assessment, compared to traditional assessment, is more multifaceted and has a potential variability, which can affect fairness and reliability. This has been known for a long time and there have been various methods for establishing the extent of inter-rater disagreement and for minimizing this disagreement, for example by training raters. McNamara maintains, however, that even though measures are taken to reduce inter-rater disagreement, such as double marking, clear definitions of performance at each level of achievement, and rater training, there will still be differences between raters.

Assessment of oral proficiency

Speaking skills are an important part of the second/foreign language curriculum. However, assessing and testing oral proficiency is a challenging task. One reason for this is that speaking is in itself interactive. Furthermore, speaking is often tested in live interactions, which means that the result of the test is difficult to predict, because the conversation can take many different turns. In addition, raters need to make instantaneous decisions about different aspects of the speaking performance, even as students are speaking. A further issue is that the rating process will always, to some extent, involve variability, as discussed previously, because it is performed by human raters.

Furthermore, there are a variety of factors involved in our judgment of how well a person can speak a language. To start with, just as in writing, different aspects are tested at the same time, for example grammar, pronunciation, fluency, vocabulary, content, and coherence. These aspects sometimes correlate but may not necessarily do so in all instances. For example, a student may have poor pronunciation but can still communicate well and get the message across.

Another difficult aspect is that spoken language is transient. In the marking of an essay the examiner can always go back and read the essay several times. By contrast, the examiner of an oral test is under a lot of pressure and has to make quick and subjective judgments. Even if speaking tests are recorded and the examiner can listen to the conversation several times, this does not recollect the whole context of the communicative situation, unless it is video-recorded.

In addition, speaking is done in real-time, which means that speakers cannot plan their speech in advance. Therefore, the planning, processing and production of spoken language are done concurrently, while actually speaking. The result of this is that the structure of spoken language is different from that of written in some respects. For example, in speech sentences are often incomplete. The danger, then, is that raters do not take this difference between spoken and written language into account. For example, in assessing oral proficiency, raters might focus quite narrowly on grammatical accuracy rather than overall communicative ability, or other features of the performance being assessed.

The nature of speaking

As mentioned above, the nature of speaking is different from that of writing. In writing there is more time to plan, edit and correct. With speaking, on the other hand, planning and editing have to be done with great speed at the same time as we take part in the speech activity. This leads to some obvious differences between speaking and writing: the vocabulary in speaking is usually, but not always, less formal, the sentences are often incomplete, and there are more repetitions and repairs, as well as more conjunctions as opposed to subordination (Fulcher, 2003). These differences, as well as their bearing on language testing, will be explored further below.

With regard to vocabulary, many rating scales for speaking reward lexical richness. However, since 'simple' and 'ordinary' words are often used in spoken language, the ability to use these words naturally should also be considered a sign of advanced language proficiency (Luoma, 2004). In addition, speakers also use fixed phrases, fillers and hesitation markers to create more time to plan their speech. Fillers and hesitation markers are phrases like *kind of*, *you know*, as well as expressions like *Now, let me see*. Fixed phrases are multi-word chunks of language (Aijmer, 2004; Nattinger & DeCarrico, 1992), which either always have the same form, or constitute a formula which can be inserted in *slot-and-filler frames*, like *the bigger, the better*. Some studies indicate that there is a relationship between test-takers' use of lexical phrases (or fixed conventional phrases) and ratings of fluency (Hasselgren, 1998). In other words, raters who listen to a speaker with a wide range of fixed phrases perceive this speaker to be more fluent compared to a test-taker who does not use many fixed phrases.

As mentioned, speakers do not always use complete sentences, but rather *idea units*, which are short phrases and clauses connected with conjunctions or sometimes just spoken next to each other, perhaps with pauses in-between (Luoma, 2004, pp. 12-13). Compared to traditional written language[7], which can have quite complex sentences with subordinate clauses, the grammar in idea units of speech is simpler. The reason for this is that speakers need to communicate a message in real time, as they actually speak.

In addition, in spoken language there are usually slips and errors, for example mispronunciations. It is important, according to Luoma (2004), to train raters so that they "outgrow a possible tendency to count each 'error' that they

[7] The term traditional written language is used as opposed to newer forms of electronic or computer-mediated written language

hear" (p. 19). Moreover, there is a danger that raters may see the different components of oral proficiency, e.g. accuracy and fluency, as separate components. Fulcher (2003) gives the example that in the most extreme cases "speech is seen as accurate and disfluent (hesitant, slow, etc.) or inaccurate and fluent" (p. 27). Hence, there is a danger that raters perceive "accuracy of structure and vocabulary in speech as one component of assessment, and the quality and speed of delivery as a separate component" (p. 27).

However, it is worth noticing that some researchers stress that the difference between speaking and writing is not as big as has often been claimed, since many of the differences mentioned above only relate to casual conversation, whereas there are many conventional exchanges that speakers are engaged in on a daily basis where differences are not as big. Nevertheless, there are aspects of speech that are 'endemic': firstly, the organization of speech is arranged in specific ways, for example in turn taking; secondly, there are different kinds of interaction mainly used in speech, for example invitations and apologies; thirdly, the speaker needs to adjust his/her speech to the context and there are different 'rules' for different contexts (Fulcher, 2003, p. 24).

Speaking test formats

There are two main test formats in the assessment of speaking: direct and semi-direct (Galaczi, 2010). The direct format involves face-to-face interaction with another person, either an examiner or another test-taker, sometimes both, whereas in the semi-direct format, an automated machine, usually a computer, elicits the test-taker's speech. A characteristic feature of interaction in the face-to-face channel is that it is bi- or multidirectional and jointly constructed by the participants. In other words, the discourse is co-constructed and reciprocal in nature, which means that interlocutors are adapting their contributions as the interaction evolves. The construct measured in the direct format is thus related to spoken interaction, which is an integral part of most construct definitions of oral proficiency. In contrast, the semi-direct format is uni-directional, and lacks the component of co-construction, since the test-taker is talking to a machine. In this format, the construct is more related to spoken production and is more cognitive in nature.

Different kinds of test tasks can be used depending on which format is chosen. Semi-direct, computer-based tests, are often organised in the form of a monologue, where the test-taker responds to a prompt provided by the machine. The response can vary in length from a brief one-word response to longer responses. The direct format, in comparison, allows for a wider range of response formats with varying interlocutors and task types – both monologic and interactive. As a consequence of the more varying response formats in the direct test, a wider range of language can be elicited, thus providing stronger evidence of the underlying abilities of the test-taker. This strengthens the validity of the assessment.

Singleton and paired speaking tests

The traditional method of assessing foreign or second language oral proficiency has been the singleton direct format, in the form of one-on-one oral interviews, one of the most famous being the Oral Proficiency Interview test of the American Council on the Teaching of Foreign Languages (ACTFL:OPI). The singleton test format usually involves an examiner/rater and a test-taker participating in an open or structured question and answer session. However, due to a change in the understanding of what kind of 'speaking' construct oral proficiency tests should measure, paired tasks with peer-to-peer interaction between non-native speakers, commonly referred to as non-native speaker to non-native speaker interaction, have become increasingly common from the 1980s and onwards.

There are several reasons for the change from the singleton interview format to peer-to-peer testing. The main reason for this shift was the empirical finding that interviews resulted in test discourse or institutional talk, not representative of normal conversation. Interview discourse resulted in asymmetric interaction with a power differential between examiner and test-taker, where the structure of the test was controlled by the interviewer (Ducasse & Brown, 2009, p. 425). Turn-taking sequences usually consisted of the interviewer asking questions and the candidate answering, leaving candidates few opportunities to give examples of their own topics or have any control of the interaction (M. Johnson, 2001; Perret, 1990). The paired format, in comparison, elicited a greater variety of speech functions and a broader sample of test-taker performance (Ffrench, 2003) and also provided test-takers with better opportunities to perform conversational management skills (Brooks, 2009; Kormos, 1999).

Another reason for the spread of the paired speaking test format was the impact of theoretical models of communicative competence (Bachman & Palmer, 1996; Canale & Swain, 1980), which have influenced the design of paired oral tests. These frameworks "include a conversation management component and presuppose the need for oral tests to provide opportunities for test-takers to display a fuller range of their conversational competence" (Galaczi, 2010, p. 4)

Finally, peer-to-peer testing proved to have a positive wash-back effect on the teaching in the classroom. Teachers started using pair and group work to a greater extent, increasingly recognised as more representative of best practice. An additional reason for the growth of the paired speaking test was that peer-to-peer assessment is more cost-efficient than the oral interview, since two students are tested at the same time.

Co-construction and interactional competence as a criterion

A typical feature of any test measuring oral interaction is that performance in the test situation is co-constructed, for example between examiner/interviewer and candidate or between two candidates in a paired speaking test (Chalhoub-Deville, 2003; McNamara, 1997; Swain, 2001). This view is based to a great extent on Vygotsky (1986) and the sociocultural theory of mind (SCT). From the standpoint of SCT, "performance is jointly constructed; it is not a solo performance but rather it is a socially mediated performance with language mediating the interaction" (Brooks, 2009, p. 342). As a result, the co-constructed nature of the performance in speaking tests poses a challenge to language testers with regard to fairness, since performances are related to each other and co-constructed.

Kramsch (1986) was one of the first to draw attention to the importance of *interactional competence* as an addition to communicative competence, advocating a deeper understanding of the concept of interaction, especially when applying this construct to speaking tests. She put forward an alternative theory called Interactional Competence Theory (ICT). Kramsch criticised the existing tests of her time, and proposed that communicative tests focus "on interactional processes and discourse skills" (p. 370).

The term co-construction is central to ICT. Jacoby and Ochs (1995) define the concept of co-construction as a "range of interactional processes, including collaboration, cooperation, and coordination" (p. 171), and they also emphasize the joint responsibility needed to achieve successful interaction.

Chalhoub-Deville (2003) and Young (2000) criticised models of communicative competence, because they focus on an individual language user in a social context, and not on activities that are co-constructed by all participants taking part in the activity. Whereas communicative competence has been considered a trait that can be assessed in an individual test-taker, ICT views the same performance as co-constructed by all participants.

The understanding of spoken interaction as co-constructed by all participants has bearing on the construct definition of speaking in a second/foreign language test situation. The question, then, is how individual scores can be awarded on the basis of paired interaction considering the fact that speech is co-constructed by all the participants. How should contributions from an interlocutor be taken into account in the rating decision, since this person is co-responsible for the co-construction of speech? This question will be referred to again in the research review in Chapter Three.

Chapter Three: Previous research on second/foreign language performance tests of speaking

In this chapter, empirical research studies into performance tests of speaking are presented. The chapter starts with some general findings in the research on rating second/foreign language performance tests (both speaking and writing). After that, research studies on speaking tests are focused upon. Finally, research specifically investigating the paired speaking test format is outlined.

Test-takers' test scores on performance tests are dependent on two variables: (1) their performance on the test, and (2) raters' interpretation and summary of that performance (Papajohn, 2002). In addition to making judgements about complex linguistic performances, raters must also apply the rating criteria. This fit between raters' judgement and criteria is of great concern because of its potential negative effect on the validity of the results. This issue is addressed by McNamara (1996): "Judgements that are worthwhile will inevitably be complex and involve acts of interpretation on the part of the rater, and thus be subject to disagreement" (p. 117). Papajohn (2002) make a similar comment: "Tests of written and spoken language attempting to assess communicative competence are complex and are therefore open to raters' interpretations and to disagreement among raters. Because important decisions are often based on the results of these tests, rater biases must be identified and reduced to an acceptable level" (p. 220).

Studies of both speaking and writing performance have explored several issues of the rating process. For example, research has shown that potential sources of rater variability might be raters' linguistic background, gender of rater, and personality fit between rater and examinee (Reed & Cohen, 2001). As regards the issue of how raters weight and apply scoring criteria, researchers have, for example, shown that there are 'implicit' criteria for raters, i.e. criteria, which are not explicitly stated in the descriptors but still used by raters. Another result is that some of the stated criteria may be more salient than others to raters, and that holistic judgements may therefore be based on one or two particular features rather than the whole range. Furthermore, features of performance may be more or less salient at different proficiency levels.

Speaking tests

Research studies on test-taker performance in speaking tests focus on either the question of inter-rater reliability or on the rating process, the latter typically by analysing verbal report data to identify rater orientations. This section starts with a short overview of research studies focusing on inter-rater reliability. Then, examples of studies exploring raters' decision-making processes are given. The studies referred to in this section are usually performed within the context of a specific speaking test. Some of these tests have holistic rating scales, whereas others have analytic scales[8]. There are also examples of different speaking test formats. Further, different methodological approaches are applied in the studies. As a result of these differences, findings are not always consistent and conclusive. This has to be kept in mind throughout this review.

Inter-rater reliability

According to Fulcher (2003), who refers mainly to studies from the 1970s and 1980s of the oral proficiency interview (OPI), there is a general claim in the literature that speaking tests often achieve high inter-rater reliability. One example is Adams (1978), who examined the relationship between five factors identified in the Foreign Service Institute (FSI) oral proficiency interview (fluency, comprehension, grammar, vocabulary and accent) and the overall ratings of the students. Altogether 834 tests were used, representing 33 languages. Findings show that agreement between two raters was consistently around 0.87 or higher. This study is often referred to as justification of the reliability of the OPI.

Based on studies on rater reliability, Mullen (1980) required that two raters be used for any speaking test, as there might be individual differences between raters. Fulcher (2003) also draws the conclusion that many studies on the reliability of speaking tests recommend that at least two raters be used in order

[8] There are two main types of rating scales (also referred to as scoring rubric or proficiency scale); holistic and analytic. In the holistic rating scale, the rater will award a global score based on a range of performance features. In comparison, in the analytic scale different features of language, i.e. different criteria, are considered separately and are added up to a final score. In short, rating with an analytic scale "involves considering several aspects of language separately, whereas a holistic scale examines a number of linguistic features at the same time" (Iwashita & Grove, 2003, p. 26).

to counteract the effects that a single rater may have on scores. Studies also show that trained raters achieve higher correlation coefficients when rating speaking performance than untrained raters do.

Shohamy (1983a, 1983b) has conducted a series of studies pointing to high inter-rater reliabilities for the oral interview. Further, in Shohamy, Reves, and Bejarano (1986) four different speaking tasks were included: an oral interview, a role play exercise, a reporting task and a group discussion. Inter-rater reliabilities proved to be 0.91, 0.76, 0.81 and 0.73, respectively. In other words, somewhat lower reliability coefficients were reported for role-plays and group discussions, i.e. test formats with more than one test-taker.

Inter-rater reliability for the English national test is continuously studied in the national test development group at the University of Gothenburg. Results indicate high degrees of inter-rater reliability (0.90) for the paired speaking test, briefly commented on by Erickson (2009, p. 6).

In most rater reliability studies, a correlation coefficient is used to report inter-rater reliability. However, rater effects, such as severity or leniency, are not taken into consideration when correlation coefficients are computed. Bejar (1985) maintains that there is often agreement about the ranking of performances, even though rater severity may differ. It has been shown in Classical Test Theory (CTT) that the reliability of ratings increases as multiple raters are used in the scoring procedure. Therefore, a correction device often applied in performance testing is to award the mean score of multiple raters, or to let a third rater adjudicate when two raters fail to agree (Henning, 1996). The use of two or multiple raters is not unproblematic, however, since it may fail to give an "accurate approximation of the true ability score" (Henning, 1996, p. 54):

> It will be readily agreed, however, that in practice two raters may agree in their score assignments and both be wrong in their judgements simultaneously in the same direction, whether by overestimating or underestimating true ability. If this happens, then we have a situation in which raters agree, but assessment is not accurate or reliable because the ratings fail to provide an accurate approximation of the true ability score. Similarly, it is possible that two raters may disagree by committing counterbalancing errors in opposite directions; that is, where one rater overestimates true ability, and the other rater underestimates true ability. In this latter situation, it may happen that the average of the two raters' scores may be an accurate and reliable reflection of true ability, even though the two raters do not agree in their ratings. (p. 54)

As a result of these limitations to address rater-related variability in CTT, more complex measurements models have been introduced in language testing research, such as multifaceted Rasch analysis (Eckes, 2009). With the help of this model, variability due to different facets of the scoring procedure, such as the use of multiple raters and different tasks, can be explored.

Rater orientations

In Meiron (1998), rater behaviour in the new Speaking Proficiency English Assessment Kit (SPEAK), used by U.S. universities to screen potential international teaching assistants, was explored using verbal protocols, written retrospectives, and questionnaires with both novice and experienced raters. The test is scored holistically, but the scoring rubric is divided into four features: functional, discourse, sociolinguistic, and linguistic. Findings indicate that, in addition to using the specified rating criteria, raters also commented on self-generated features that were not explicitly mentioned in the scoring rubric. Also, when candidates had different proficiency levels, the tendency for raters was to focus on linguistic features shared by candidates, instead of salient features of the specific individual performances. Furthermore, two methodological approaches were identified: a "quasi-analytic rating" method where raters focused on specific features of the performance, such as grammar, and a more "global" or "holistic" assessment.

Pollitt and Murray (1996) examined the rating process in the Cambridge Assessment of Spoken English oral interview. They came to a similar conclusion as Meiron (1998) about rating methodologies. They found that whereas some raters had a *synthetic* process of rating, which was based more on intuition, others had a more *analytical* approach. The results also indicated that when the pairing of candidates resulted in mixed proficiency levels, raters focused mostly on the criteria for the lower-level candidate in the pair. Moreover, findings show that certain performance features were strongly related to particular levels of the rating scale. For example, raters seemed to focus more on grammatical accuracy at the lower levels and more on sociolinguistic and stylistic competence, representative of more sophisticated speech, at higher levels. A further example of a finding from the study by Pollitt and Murray (1996) was that they found that raters made inferences about candidates based on how they behaved in the language testing situation. For example, raters referred to test-takers' exam-consciousness, maturity, and willingness or reluctance to take part in the conversation.

Adams (1980) also explored differences in assessment focus in relation to proficiency levels. He studied the relationship between five features, which had been identified in the FSI oral interview, namely accent, comprehension, vocabulary, fluency and grammar, and the overall speaking score (on a scale of 1-5). The results showed that vocabulary and grammar were the main features that discriminated levels, whereas accent and fluency failed to discriminate at some levels.

The relationship between grammatical errors in transcripts of the OPI, conducted with 40 college students of French, and OPI ratings, was explored by Magnan (1988). A significant correlation between percentage of grammatical errors and OPI ratings was found. However, it was not always linear. Magnan draws two main conclusions: (1) the relationship between error and proficiency level varies depending on the kind of error, and (2) learners at higher levels try to use more complex grammatical structures and thus make more errors.

Another example of a study exploring features that are salient to raters is McNamara (1990) who used item-response theory (IRT) to investigate an Occupational English Test. Candidates participated in a role play and the rating scale included the following analytic categories: overall communicative effectiveness, intelligibility, fluency, comprehension, appropriateness and resources of grammar and expression. Findings showed that resources of grammar and expression, i.e. a candidate's grammatical and lexical accuracy, were the most significant factor for raters in determining the candidate's total score on the test. In comparison, whereas resources of grammar and expression were most harshly scored, comprehension was most leniently scored. McNamara (1996) draws the following conclusion: "It has frequently been found that raters judge aspects of performance concerned with control of the formal resources of the language, particularly grammatical structure, more severely than they rate other aspects of the performance" (p. 123).

The difficulty of using holistic rating scales was highlighted in Annie Brown (2007). Verbal protocol analysis (VPA) was used and Brown found that the largest group of rater comments (31%) related to syntax, and more than half of these comments were negative (55%). The other salient features were discourse (22%), i.e. comments about coherence, production (18%), i.e. comments referring to fluency and pronunciation; finally comprehensibility (i.e. raters' understanding of test-takers), vocabulary and strategies made up about 10%

each of the comments. Further, Brown found that different criteria seemed to be more or less noticeable at different levels to raters of the International English Language Testing System (IELTS) oral interview. One example is comprehensibility and production. These two features received most attention at the lower levels, and were, in most cases, commented on when there was a problem. Brown also found that different examiners heeded different performance features, favouring some over others. A finding similar to Meiron's (1998) was that in addition to criterion features from the rating scale (e.g. syntax and vocabulary), raters also focused on features not explicitly stated in the rating scales, such as pronunciation and fluency.

In Annie Brown, Iwashita, and McNamara (2005), two exploratory studies are reported. First, verbal report methodology was used to analyse rater orientations, and secondly, features of test-taker discourse on two task types of the Test of English as a Foreign Language test of Spoken English (TOEFL TSE) were analysed. Raters provided comments without using any rating scale, i.e. comments were unguided. Findings show that that *linguistic resources* made up a large part of the coded comments. The other categories included *phonology*, *fluency*, and *content*. However, the authors conclude that raters "take a range of performance features into account within each conceptual category and that holistic ratings are driven by all of the assessment categories rather than, as has been suggested in earlier studies, predominantly by grammar" (p. iv). Furthermore, the analysis of test-taker discourse provided empirical evidence for the comments by the raters.

The last example is Hsieh (2011), who examined rater effects and rater orientations when two rater groups judged potential international teaching assistants' oral proficiency. Data consisted of scores and raters' concurrent written comments regarding features that they paid attention to in the rating process. Findings on rater orientations show that the majority of comments were related to *phonology* and *linguistic resources*. *Fluency* was also a large category. In comparison, raters commented less on their *global impression* of the candidates and on *content*. Finally, there were very few comments pertaining to *non-linguistic factors*.

Paired speaking tests

One of the first published studies focusing on paired interactions was Iwashita (1996), who compared the impact that the pairing of candidates had on scores. Candidates, twenty adult learners of Japanese, were paired with interlocutors with similar and different proficiency levels. The results show that even though the proficiency level of the interlocutor affects the quantity of discourse, it does not significantly change the scores candidates were awarded. Another result was that test-takers were asked about their preference with regard to the two test conditions. Test-takers preferred the paired speaking test to the interview since it was less threatening. Test-taker preference for the paired speaking format has also been reported by Egyud and Glover (2001), Taylor (2000) and May (2000).

Foot (1999) criticised the paired speaking test format and questioned its fairness. One of the problems addressed by Foot was the possibility that candidates were disadvantaged because they were paired with candidates of differing proficiency levels: "unless the candidates are well-matched, their attempts to sustain a discussion are likely to be, and often are, faltering and desultory, and the outcome, for them a sense of frustration rather than of achievement" (p. 40). Moreover, Foot addressed the issue of mutual incomprehensibility, for example if both students had problems with pronunciation, or accents that were strong and difficult to understand. Finally, Foot cautioned against candidate preference of paired speaking tests to the traditional interview, claiming that this was not sufficient evidence to incorporate paired interactions in high-stakes speaking tests.

Taylor (2000) responded to Foot's criticism by reporting results from two internal studies carried out on behalf of the UCLES (University of Cambridge Local Examinations Syndicate) to compare the paired and one-on-one speaking test formats. The results of the quantitative comparisons showed that the paired format offered a more balanced interaction between participants with the examiner taking a smaller role as well. In addition, the paired format generated a larger and more varied sample of speech from the test-takers, compared to the oral interview. Furthermore, the results from qualitative comparisons showed that the paired test format elicited more communicative language functions than the traditional singleton face-to-face interview.

Swain (2001) expressed concern that there was a lack of focused research into pair and group speaking tests. Moreover, she brought up the question of individual scores in peer-to-peer interaction. Therefore, Swain proposed that paired candidate discourse be examined more closely. As a result, an increasing number of discourse-based studies have appeared, which show that peer-to-peer interaction provides the potential for a more balanced discourse with a greater variety of functions and more opportunities for interactiveness (Ducasse & Brown, 2009, p. 425).

In addition to features of test-taker interaction, Galazci (2010) divides research into the paired or group speaking test format into two other categories, namely the effect of background variables and the raters' and the test-takers' perspectives. The following section focuses on research exploring the raters' perspective.

In a study by Brooks (2009), interaction in the oral proficiency interview and the paired format was examined in relation to scores. The quantitative results show that test-takers' scores were on average higher in the paired speaking test format than in the individual. Furthermore, qualitative analysis of candidate discourse indicates that there is a substantial difference in performance in the two test formats: the interaction in paired speaking test was much more complex and linguistically demanding than the oral interview. Examples of interactive features in the paired speaking test format were: "prompting elaboration, finishing sentences, referring to partner's ideas, and paraphrasing" (p. 361). Brooks draws the conclusion that it is important that the joint construction of performance be taken into account in both the development of rating scales and in construct definition.

Galaczi (2008) is an example of a discourse-based study, in which candidate discourse in the paired speaking test format was explored in relation to scores awarded for "Interactive communication". In her analysis, Galaczi highlights three patterns of interaction: "Collaborative", "Parallell" and "Asymmetric". In collaborative interactions, the participants displayed high mutuality and high equality, for example alternating their roles as listener and speaker. In parallel interactions, partners showed high equality by initiating and developing topics, but low mutuality since they did not build on each other's ideas. Finally, in the third pattern, the two speakers showed "different discourse roles, one dominant and one passive, with moderate mutuality in topic development" (p. 106). Galaczi concludes that there is a clear relation between discourse and scores. Candidates in pairs with collaborative interaction were rated highest, whereas parallel and asymmetric dyads were rated lower.

Another study focusing on features salient to raters in their decision- making process was carried out by Orr (2002) involving the First Certificate in English

(FCE) speaking test. Verbal reports were collected from 32 raters after they had watched video recorded simulated FCE Speaking tests with two candidates in a paired interview. Findings show that raters did not pay attention to the same aspects of the rating criteria, and that they noticed non-criterion features of the performance. One consequence of this was that raters awarded different scores to the same performance, but also that they perceived different aspects of the performance when they awarded the same score. This points to the fact that raters interpret scale descriptors in different ways: "For each rater there appears to have been a unique interaction of factors which led to the awarding of a score" (p. 152). Orr concludes with an ominous remark: "The validity of the interpretations that the test users might wish to make of the results is thus brought into question" (p. 143).

Ducasse and Brown (2009) report findings from a verbal protocol study of 12 teacher raters who rated 17 videotaped paired interactions. Analysis of the verbal report data showed three main categories of interactional features that are typical of successful interaction: (1) non-verbal interpersonal communication, which includes gaze and body language; (2) interactive listening, which is about how candidates show engagement and attention while listening to each other in the conversation; and (3) interactional management, which encompasses how candidates manage the topics and turns.

The issue of individual scores based on co-constructed interaction is addressed by May (2009). She explored four raters' decision-making process when judging pairs with asymmetric patterns of interaction (see Galazci above). She analysed candidate discourse together with "rater notes, stimulated verbal recalls, rater discussions and scores awarded for interactional effectiveness" (p. 397). One of the main findings was that raters viewed key features of the interaction as mutual achievements, and May therefore suggests shared scores for interactional competence.

Finally some studies on the effect of background variables should be mentioned. A challenge for the paired speaking test is the so-called *interlocutor effects*, i.e. effects on performance that are produced by variables associated with the other participant (Galaczi, 2010, p. 6). Research has shown that there are three main variables that may have an effect: (1) proficiency level of the paired candidates, (2) their personality, and (3) their acquaintanceship. These three features will be focused upon here, but it should be noted that there are other *interlocutor effects* that have been studied, such as gender and ethnicity. Berry (1993, 2004) and Nakatsuhara (2009) have studied the effect of personality in the form of extraversion and introversion levels. Berry's two studies, examining the relationship between extraversion levels and the discourse produced, had somewhat contradictory findings. The first study reported that extroverts performed best when paired up with other extroverts. However, there were no significant differences for introverts. In the second study, she found that the degree of extraversion had no significant effect on scores for the extroverts, whereas the introverts' scores were noticeably affected. Nakatsuhara (2009) found some relation between extraversion level and test-taker performance, but it was strongly associated with task type. Both Berry and Nakatsuhuara found that extroverts favour a higher degree of freedom, as in the paired speaking test format, compared to introverts, who prefer structured and highly prompted tasks. This could have consequences if an extrovert is placed in their least favourite situation and vice versa.

The effect of peer interlocutor's proficiency level has also been researched. The main finding points to most positive effects for the paired speaking test format when proficiency levels of the test-takers in the pair differ to some extent. However, wide divergence of proficiency levels is not recommended. As mentioned above, Iwashita (1996) found that the proficiency level of the other participant could have an effect on the amount of talk (being paired with a partner of higher proficiency level usually resulted in more talk), but not so much on scores. This result is echoed in Davis (2009). In other words, talking more did not automatically render higher scores. Another study by Nakatsuhara (2006) found that conversational styles were similar in both same-proficiency and different-proficiency level pairs. Finally, Norton (2005) reported that there might be a positive effect on the quality of speech if a test-taker is paired up with a higher-proficiency partner.

The last variable that has been studied is test-taker familiarity. O'Sullivan (2002) found that there was a relationship between familiarity and scores. When working with friends, candidates received higher scores. However, the results were complex and O'Sullivan also investigated the effects of "sex-of-interlocutor". He concluded that the effect that variables such as gender and familiarity have on test scores are cultural-specific.

In this chapter, a research review focusing on performance tests of speaking has been made. Findings of previous research show that there may be differences in how raters weight and apply scoring criteria. Furthermore, raters seem to heed both criterion and non-criterion features of test-taker performance. Features may also be more or less salient to raters at different proficiency levels. As regards inter-rater reliability, results of research on the OPI indicate relatively high reliability coefficients, around 0.80. However, somewhat lower reliability coefficients were reported for speaking test formats with two or more test-takers, such as role-plays and group discussions. Finally, the paired speaking test format has been proved to have many benefits, such as eliciting a wide range of speech functions and a broad sample of test-taker performance. However, one of the main challenges is the question of the fairness and validity of this test format. For example, so-called interlocutor effects, i.e. variables associated with the other participant, may affect test scores. Also, the fact that performance is joint and co-constructed raises questions about individual marking.

Chapter Four: Material and method

In this chapter, the method of data collection and data analysis is presented. In addition, conclusions about the generalisability of the results, as well as the validity of the methods, are discussed. Finally, ethical aspects are considered.

In the present study, a mixed-methods research design was used, allowing for the collection of both quantitative and qualitative data. This was done in order to achieve a broader understanding of the data and increase the depth of the analysis (Dörnyei, 2007). In language testing, quantitative data are usually explored to examine reliability of scores or consistency/severity of rater behaviour. In comparison, qualitative data typically comprise verbal protocol analysis (VPA) to explore the rating process, which is also the case in the present study. In Table 1, an overview of the study is given, to be further explained in this section. The chapter is broadly structured in the following way: Firstly, the context of the study is outlined, and then procedures for data collection and analysis are described.

Table 1. Overview of study: sequencing of rater activity, data collection and data analysis

Sequence of rater activity (one-day seminar)	**Data collected**	**Data analysis**
Introduction with information about the research study and instructions on the rating activity, as well as a short practice session		
Raters individually listen to six paired conversations and make notes while listening	Rater notes	Not included in the analysis of the present study
After listening raters award scores/marks	Scores	(a) descriptive statistics (b) correlation statistics (c) reliability statistics
Immediately after making their judgements, raters provide features of the performance that attracted their attention, or the rating criteria they employed, as they made their judgement	Summary comments	(a) segmentation and coding (b) frequency counts of coded data
Group discussion	Filmed group discussion	Not included in the analysis of the present study

The speaking test

The oral part of the Swedish national test of English is a performance test in the *weak sense* (see section on performance assessment), the aim of which is to measure and evaluate students' general language proficiency regardless of where, when and how this ability was acquired. The test used in the present study is from the *English 6* course[9] in upper secondary school, whose minimal passing level is intended to correspond to B.2.1 in the CEFR. It is a direct paired or group (the option to use three candidates is possible) oral test with peer-to-peer interaction. In other words, a characteristic feature of the test is that discourse is co-constructed by the test-takers as the conversation evolves.

Students are divided into pairs, or sometimes groups of three. They are given 15 minutes to prepare for the test. During this time, they go through the instructions of the test as well as read a short text, which they will summarise and comment on in the test situation. The test has a theme, for example *Stress*, around which the conversation will circle. There are explicit instructions for the students that clearly state what they are expected to do. In the first part of the test, focus is on oral production, but there is also some interaction. Students are instructed to summarise the main points of the short text they have read in advance and discuss it with their partner. In the second part of the test, interaction is focused upon, and students discuss and argue about the topic based on a given set of questions or statements. In summary, the construct that the test aims to measure is mainly oral interaction, but also to some extent oral production.

The test-takers

Six audio-recorded student conversations from the pre-testing of the national test for spring 2013 were used in the present study, corresponding to twelve student performances. The students have given their consent for the use of the test material for research purposes. The candidates in the material are six pairs with one boy and one girl in each pair. The reason for this is that it makes it easier for the raters to distinguish between the two speakers if they are of different genders. The performances are quite representative of the whole rating scale. This was checked beforehand with the help of data from the raters in the

[9] In the Swedish upper secondary school, the subject English comprises three separate courses, one for each school year: English 5, English 6 and English 7. These courses are aligned to the CEFR and their minimal levels are intended to correspond to B1.2, B2.1, and B2.2, respectively.

benchmarking process (*n* = 12). As stated before, these performances were part of the benchmarking, but were never selected to be used in the material given to the teachers; consequently, the material was new to the raters in the study.

In addition, the pairing of candidates was done so that the two test-takers were well matched as regards proficiency level. This means that the candidates were on reasonably equal levels of spoken proficiency.

The Swedish raters

The Swedish raters (*n* = 17) are all practising teachers of English at the Swedish upper secondary school level. They work at different schools, both municipally and independently operated, in two different regions in Sweden. Participation in the study was voluntary, which means that they agreed to participate after receiving information about the study. They are all formally qualified teachers of English and have experience of rating the Swedish national tests of English.

I made contact by first e-mailing the head teacher with information about the study and followed up by calling the head teacher to see that he/she had received my e-mail and to ask if it would be possible for one or more of the teachers of English at the school in question to participate in the study. I either continued communicating with the head teacher or one of the English teachers that I had been advised to contact.

The teachers filled in a short background questionnaire. In Appendix 1 part of this background information, including gender and teaching experience, is presented. In summary, 24% (4/17) of the raters are men and 76% (13/17) female. Three of them speak English as their L1, whereas the others are native speakers of Swedish. Finally, teaching experience, and hence experience of rating national tests, since this is part of the teachers' job in Sweden, ranges from 1 to 29 years. There are four participants with quite little teaching experience, from one to four years. The other participants have all worked for six years or more, which means they could be categorised as quite or very experienced.

Rating criteria for Swedish raters

The speaking test in the Swedish national test of English is scored holistically using the Swedish national performance standards for course English 6 in the Swedish upper secondary school, provided in Appendix 2. In addition, there are analytic assessment factors intended to be a support for teachers in making their holistic judgement. The assessment factors are to be viewed as different aspects of qualities of spoken language, and are divided into two main categories: content and language. They are based on the communicative, and action-oriented language approach that forms the basis of the Swedish syllabuses for foreign languages, and intended to provide support for teachers in making their holistic judgement of students' oral performance. The performance standards are holistic, whereas the assessment factors are analytic. In Appendix 3, the assessment factors are shown in Swedish with translation into English.

It is important to emphasise that, in the rater instructions, it is stated that the teacher/examiner should play a minimal role in the conversation and let the students develop and advance the conversation as much as possible on their own. However, in order to make sure that both students get an equal chance to show their speaking skills the teacher may help in the conversation by, for example, asking questions.

For each national test, benchmarked examples of student conversations are selected by the developers of the test from the group of students who take part in the try-out phase of the development of the test. Teachers are instructed to listen to the benchmarked examples and read the comments on rating and marking as preparation for their own assessment.

The external CEFR raters

In addition to the raters from the Swedish school system, external raters from Finland (*n* = 7) and Spain (*n* = 7) participated in the present study. They rated the same six conversations as the Swedish raters. The reason for including external examiners was to make a small-scale comparison between the Swedish performance standards for EFL and the CEFR-levels. In addition, the CEFR raters were also part of the analysis of rater orientations, i.e. features that raters pay attention to when awarding scores. As mentioned previously, the foreign language syllabuses in the Swedish school system are adapted and aligned to the CEFR-levels. The minimal passing level of the course English 6, in the present study, for example, is intended to correspond to the B2.1-level in the CEFR. It is worth noting that only the minimal passing level has been textually aligned. Hence, no maximum achievement level is specified. There have been some textual analyses and continuous, empirical observations for validating the foreign language courses in the Swedish school system in relation to the

CEFR levels, but, so far, no large-scale, systematic studies have been performed (Erickson, 2010b). Therefore, the opportunity to compare the Swedish performance standards for EFL to the CEFR levels with the help of external raters is valuable.

Yet another reason for involving external CEFR raters was that the Swedish teachers are not used to working with scaled CEFR descriptors. The Swedish national performance standards for EFL, in the form of national goals and grading criteria, are aligned to the CEFR but this is not explicitly stated in the grading criteria for different proficiency levels. In comparison, the Finnish and Spanish raters all had previous experience working with the CEFR scales in assessment contexts.

It is worth emphasising that, whereas the external CEFR raters had previous experience of using CEFR scales in testing, as opposed to the Swedish raters, they were familiar neither with the specific speaking test, nor with Swedish oral tests in general. There is no equivalent speaking test of EFL in their countries, at least not during the time of the study. In other words, rating this specific model of a paired speaking test, focusing on interaction, was a new experience to them. Therefore, the focus in the analysis of the external raters' scores was not on rater variability but rather on their ranking of the performances, as well as at what levels in the CEFR they assessed the Swedish students' performances to be. Moreover, the CEFR raters were also included in the analysis of features that raters paid attention to while making their holistic judgements. There was no background questionnaire for the CEFR raters. In this group, there were two men and twelve women. None of them had English as L1. The common denominator was their previous experience of rating with CEFR-based scales.

Rating criteria for the external CEFR raters

The criteria used by the external raters are taken from *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment – A Manual* (Council of Europe, 2009). According to the Council of Europe webpage, the Manual aims "to help the providers of examinations to develop, apply and report transparent, practical procedures in a cumulative process of continuing improvement in order to situate their examination(s) in relation to the CEFR". The Manual provides forms and related tables for all the communicative language activities, described in chapter 4 of the CEFR, and for the various aspects of communicative language competence, described in chapter 5. In the present study, the tables below from the Manual were used. These tables are included in Appendix 4.

(a) Table C1: GLOBAL ORAL ASSESSMENT SCALE (p. 184)

(b) Table C2: ORAL ASSESSMENT CRITERIA GRID (p. 185)

(c) Table C3: SUPPLEMENTARY CRITERIA GRID: "Plus levels" (p. 186)

The first of these scales, Table C1, is a global scale, supposed to be used in the first 2-3 minutes of a speaking sample to decide approximately at what level the speaker is. After this, the rater/examiner should change to table C2, and assess the performance in more detail. Table C2 is divided into five analytic criteria, based on components of communicative language competences as well as on interaction and production strategies described in the CEFR: accuracy, coherence, fluency, interaction, and range. For each criterion, descriptors are provided for the different performance levels. Table C3 comprises supplementary criteria with descriptors for the ”Plus levels” in the CEFR (B2+, B1+, A2+).

The rating scales

Since the Swedish and European raters used different – although related – criteria, two different scales are employed in the present study. The main aim of examining the Swedish raters' scores is to analyse variability of scores and inter-rater consistency, i.e. the consistency of scoring between raters in order to see how well the teachers agree on the rating. As mentioned in Chapter One: Introduction, this is an area where the Swedish Schools Inspectorate has expressed criticism, since according to their studies scoring reliability is too low for parts of the national tests. However, the oral parts of the national tests have not been scrutinised, which makes it interesting to examine inter-rater consistency in the present study.

In the case of the external European raters, the aim is to see at what levels in the CEFR they judge the Swedish students' performances to be. This is interesting from a validation point of view, since the minimal passing level of the test is intended to measure B.2.1-level, but there has been very little empirical validation of the alignment claimed.

Below, in Tables 2 and 3, the rating scales for the Swedish raters and the external European raters are provided.

Table 2. Ten-point scale used by the Swedish raters

F-	F+	E-	E+	D-	D+	C-	C+	B	A
1	2	3	4	5	6	7	8	9	10

Table 3. Nine-point scale used by the CEFR raters

A1	A2	A2+	B1	B1+	B2	B2+	C1	C2
1	2	3	4	5	6	7	8	9

As can be seen, the Swedish raters used a ten-point scale, with a passing level, E–, intended to correspond to a minimal B2 in the CEFR, B2.1. The European raters used a nine-point scale, which covers the full CEFR range of levels, including plus levels (A1-C2). It is important to stress the difference between these two scales and also that they serve two different purposes. The Swedish scale was used to examine inter-rater consistency and rater behaviour, whereas the CEFR scale was used in a more general way to examine judgements of levels in the CEFR.

Data collection procedure

The data from the raters were collected during a one-day rating seminar. The structure and organisation of this seminar were identical for both the Swedish and the external raters. The Swedish raters participated in the rating seminar in June 2013, on two different occasions, since they were divided into two groups and came from different parts of Sweden. The Finnish raters participated in September 2013, the Spanish in November 2013.
The one-day seminar was structured as follows:

1. Introduction with information about the research study and instructions on the rating activity, as well as a short practice session
2. Individual rating of six audio-recorded conversations from a Swedish national test of English
3. Group discussion about the rating activity

In the introduction, information about the research study was presented and procedures for the rating activity were explained. In order for the raters to try the rating procedure once before starting the individual rating, we listened to one benchmarked conversation together. The raters made notes while listening and decided on a mark for each of the two students. We then had a short discussion and compared views and opinions on important features of the oral performances of the two students. Finally, some of the comments on the same two performances from the benchmarking material were presented.

As a second step in the rating seminar, the raters were asked to listen to each conversation individually, using headphones, with stops and repetition where needed, and to take notes by hand while listening (on separate sheets for each test-taker). They were instructed to take notes freely, like recording a stream of consciousness, and write down everything that came to their mind about each oral performance in focus. In other words, they did not have to worry about writing full sentences or being correct, but rather "jot down" as many aspects as possible, including verbatim quotes, to which they paid attention while listening and forming their judgements. The notes were used in two ways: firstly, they were of help to the raters as they made their individual rating decision and filled in their assessment forms, and secondly, they were used in the group discussion to help remember when talking about individual performances and comparing rating decisions. Because of the limited scope of the present study, rater notes are not included in the analysis. However, it would be interesting to examine the relationship between rater notes and the summary comments in a future study.

After listening to each conversation, the participants filled in an assessment form for each student performance with a mark/score and a summary comment about the performance. For the summary comment they were asked to explain what features of the oral performance they paid attention to in making their decisions. Since we only had one day, the time constraint for the raters was to spend a maximum of 30 minutes on each conversation. The conversations were on average 15 minutes long, which left 15 minutes for deciding on a score and writing summary comments.

The Swedish raters were allowed to use either Swedish or English as they wrote their comments (and notes). 12 of the Swedish raters decided to write their summary comments in English, and 5 in Swedish. However, the external CEFR raters were asked to write in English, because of the difficulty of translating their texts had they written in their first language. Rater comments in Swedish have been translated into English. These translated quotations are marked with the abbreviation Tr. A list of the corresponding verbatim Swedish quotations is provided in Appendix 5.

To end the day, the raters had a group discussion, which was filmed and recorded, where they talked about assessing English oral proficiency in a paired speaking test and also compared their ratings and comments. The group discussion was filmed in case I might use this material in the analysis of the data. However, it was decided quite early on that the data consisting of summary comments and scores would suffice for the present study. Appendix 6 contains the written instructions to the CEFR raters.

Data analysis

The data in the current study thus consist of two parts: scores and summary comments. The first category, scores, is more 'quantitative' in nature, as compared to the summary comments, which are mainly 'qualitative'. Hence, the analysis was divided into two parts as a result of the mixed-methods design. It should be noted here that the description of the analysis of the qualitative data is longer than that of the analysis of quantitative data. The reason for this is that the qualitative data analysis process needs to be explained in a transparent and explicit way, due to its interpretative nature.

Analysis of quantitative data

To answer the first research question dealing with inter-rater variability and rater severity among the Swedish raters, the scores were analysed using (a) descriptive statistics, (b) correlation statistics and (c) reliability statistics. Data were entered into SPSS version 21, a software package used for statistical analyses. Firstly, descriptive statistics were run. Then, Spearman rank order correlations and Kendall's Tau correlations were performed for the pair-wise ratings of the Swedish raters, in order to measure inter-rater reliability. Finally, Cronbach's alpha, which measures internal consistency for the whole rater group, was calculated. Information and comments on these measures will be given in connection with the presentation of the results.

To answer the fourth research question, concerning the levels in the CEFR that external raters judge the performances of the Swedish students to be at, CEFR raters' scores were analysed using descriptive statistics.

Analysis of qualitative data

To answer the second research question regarding features that drew raters' attention as they judged the oral performances, the written summary comments that raters provided were analysed following procedures for verbal protocol analysis (VPA), as suggested by Green (1998). VPA is used to explore cognitive processes (Green, 1998). The verbal protocol typically consists of utterances made by the informant as he/she is asked to either 'talk aloud' or to 'think aloud' while carrying out a task. In the case of rating speaking performance, the verbal reports are usually performed retrospectively, in the form of stimulated recall, since it is impossible to comment on a speaking performance while listening to it. In several studies using this method (Ducasse & Brown, 2009; May, 2006, 2009, 2011a, 2011b; Orr, 2002), raters are asked to listen to the performance once and then record an oral summary statement. After that they listen once again, but this time stopping from time to time to record comments (a 'stimulated recall') when they notice interesting features of the performances. These recorded statements constitute the verbal reports.

In the present study, verbal reports were collected in the form of written summary comments. In comparison with previous studies that have used VPA, there are some differences. In the current study, no stimulated recall was carried out. In other words, the raters did not listen to the conversation again, stopping at intervals to record comments. Instead they just listened once, but could stop and go back and forward as they wished.

Another major difference is that the verbal reports in the present study are written and not oral as in the studies mentioned above. The main reason for using written material was authenticity. Since the raters were only available for the one-day seminar, it was important to make this rating situation as authentic as possible for the participants. In the case of VPA, informants are usually trained in giving verbal reports before taking part in the real study. There was no time for such training in the present study. However, writing notes during listening and summarising the impression afterwards in a written comment is part of many raters' normal rating procedure. From my point of view, the raters seemed at ease with this rating procedure.

Another reason for using written reports is that the number of raters in the present study is quite large compared to previous VPA studies. In May (2011) and Ducasse and Brown (2009), 4-12 raters are used. Finally, a comparison can be made between this study and that of Hsieh (2011), who also used only written comments to analyse rater orientations (the study is described in Chapter Three: Previous research on second/foreign language performance tests of speaking).

Analysis of written comments

Each rater thus gave a verbal report in the form of a summary comment for each test-taker's performance. This amounts to 372 summary comments – 31 raters commented on 12 student performances. The summary comments were written digitally in a Word document by the raters during the rating seminar, and so no transcription was necessary. In accordance with Green (1998), the verbal reports were divided into segments that each represents a different process. In the present study, this means that each segment comprises one main idea that the rater paid attention to. An example of a segmented summary comment, with segments indicated by backslash, for the girl in conversation 1, is given below (Example 1):

> Example 1: General communication skills are good/
> she has fluency/
> and structure./
> She listens to what the male is saying and as the conversation develops she acknowledges his thoughts and even adds her own opinion to the subject at hand. She even puts the question back to him for further discussion./
> Her vocabulary, phraseology and idiomatic expression are good./
> She has a problem with some words, cult, busy, essay, symbol and students but this is only a disruption in communication./
> This is weighed up by her depth and breadth of the content of what she is saying./
> There is some complex explanation of her opinion in a couple of places, for example "chat" and "facebook"/

Development of a coding scheme

The next step in the qualitative data analysis was to develop a coding scheme that would describe the raters' summary comments and answer the relevant research questions in an adequate way. The difficulty of developing and using a coding scheme is highlighted in Green (1998). The crucial point is that there may be a lack of agreement as to what features exactly constitute the "precise nature of the coding categories that may be used for the analysis of verbal report data", which could lead to the consequence that "[t]wo researchers may independently develop different schemes for the analysis of the body of data" (p. 68). This does not invalidate the technique according to Green, but she nevertheless cautions that this inherent variability has an effect on the inferences that can be drawn from the results.

Moreover, a balance needs to be drawn between the wish to cover the minutiae of the verbal reports and the necessity of identifying broader coding categories. This has to do with aspects of feasibility as well as reliability and is an essential step in the analysis.

To start with, I read through some of the summary comments and rater notes in order to see what features raters commented on. In addition to this, I studied the criteria used by the raters, both the Swedish performance standards and assessment factors, as well as the CEFR tables from the Manual. To complement this, I also reviewed the illustrative scales in the CEFR for communicative competence, which is divided into linguistic, sociolinguistic and pragmatic competences, (Council of Europe, 2001, pp. 108-130), as well as the scales for interaction strategies (pp. 85-87) and production strategies (pp. 64-65). In summary, the coding scheme was thus developed on the basis of the criteria the raters used, the illustrative scales in the CEFR, and the written rater comments. In Table 4, the main categories and subcategories are presented; the complete coding scheme is provided in Appendix 7.

Table 4. Coding categories

Main categories	Subcategories
Accuracy	Grammatical accuracy
	Phonological control
	Vocabulary control
Coherence	Coherence and cohesion
	Flexibility to circumstances
	Topic development
Fluency	Fluency mentioned in general
	Hesitation and pauses
	Speed of delivery fast or slow
Intelligibility	
Interaction	Cooperating
	Dominates discussion (usually negative)
	Has a passive role in discussion
	Manages or controls discussion (usually positive)
	Turntaking
Other	
Production strategies	Monitoring and repair
	Compensating
Range	General linguistic range
	Vocabulary range
	Ability to express viewpoints
Sociolinguistic appropriateness	
Task realisation	Completing and understanding task requirements
	Length of response - brief or extended discourse by candidate
	Overall comments
	Summary of text

As can be seen in Table 4, the coding scheme consists of ten main categories and 23 subcategories. It is worth noting that not all main categories have subcategories. Each segment was coded in terms of the main category, and then the subcategory. In other words, the segment was first coded in relation to one or more of the ten main categories. Then, when applicable, the segment was further coded as being related to one or more of the subcategories.

Below, some examples of issues that came up in the development of the final coding scheme are outlined. Firstly, in order to avoid making too many main categories, a category called *task realisation* was created. In this main category, the following subcategories were included: *length of response by candidate* (either extended or very brief), *completing and understanding the task requirements*, *comments on the overall performance of the candidate*, and *candidate's ability to summarise* *the text* (which was part of the tasks in the test). As it turned out, these subcategories were quite small, which means that if they had been used as main categories, they would have been even smaller. The argument to use them in the same main category is that they all refer to how the candidate, in one way or another, fulfilled the task requirements.

Another category requiring explanation is the *other* category. This category was used for all instances where a coded comment did not fit into any of the other main categories.

It is also worth explaining the difference between the subcategories *topic development*, which is part of the main category *coherence*, and the subcategory *ability to express viewpoints*, which is part of the main category *range*. First of all *topic development* is based on the illustrative scale in the CEFR called *thematic development*, which is an aspect of discourse competence. In this scale, a candidate's ability to develop a description or a narrative in a clear way, "expanding and supporting his/her main points with relevant supporting detail and examples" (Council of Europe, 2001, p. 125) is described. In my material, raters did not comment on candidates' ability to give clear descriptions and narratives, but more on their ability to develop and elaborate on their topics with supporting examples and details (*topic development*). The other category that is somewhat related and similar to *topic development* is *ability to express viewpoints*. In the CEFR, there are illustrative scales for different aspects of linguistic competence. Two of the main categories in the coding scheme, *accuracy* and *range*, are based on the scales for linguistic competence. As for *range*, two illustrative scales from the CEFR were useful, namely *vocabulary range* and *general linguistic range* (Council of Europe, p. 110-112). However, in my material there were also comments on candidates' ability to express viewpoints and ideas. This aspect could not be found in a separate illustrative scale in the CEFR, but was embedded in *general linguistic range* (Council of Europe, p. 110). Moreover, the rating scale used by the CEFR raters from the Manual (Appendix 4) also includes the test-taker's ability to express viewpoints as a part of *range*. For the B2 level, for example, it is stated in the descriptor that the test-taker should have "a sufficient range of language to be able to give clear descriptions, *express viewpoints on most general topics* without too much conspicuous searching for words, using some complex sentence forms to do so" (my italics). Hence the subcategory *ability to express viewpoints* under the main category *range* was created. When I did the coding, it was shown that it was not always easy to separate *topic* *development* from *ability to express viewpoints*, and in some cases the comments were therefore double coded.

Inspired by, for example, May (2011), I decided to use an additional coding layer, by also coding the evaluative response of the rater, i.e. if the comment is Positive, Negative or Mixed. I considered having a fourth subcategory in the evaluative response category, called Neutral, but after going through the material I decided that it would be most practical just to have the three. In cases where it is not quite clear whether the comment is Positive or Negative, Mixed is used. Also, several cases are clearly both Positive and Negative, in which case Mixed is used as well (see examples in the results section).

Further, also in line with May's (2011) coding scheme, I decided to code the *focus* of the comment (*Focus of response* in the coding scheme in Appendix 7), because I soon realised the raters had made many comments where they did not refer to the individual candidate but were rather comparing the candidates in the pair and how they interacted. This is interesting to code, since one of the challenges of rating paired interaction is the difficulty of separating scores when the discourse is joint and co-constructed by the candidates. The *inter-candidate comparisons* consisted of comments on (1) similarities between the two candidates, (2) differences between the two candidates, (3) candidates' proficiency levels, and (4) the interaction between the two candidates. Moreover, there were some comments comparing the candidate's development during the test, and these were coded as intra-candidate comparisons. Comments in this main category often referred to other categories as well, and were thus double-coded. They were also coded for evaluative response, where applicable. In Example 2, the comment refers to similarities between the speakers' personalities (shy) and is thus coded as *inter-candidate comparison*. There is also a reference to lack of *topic development* (not having much to say about the topics), which is coded under the main category *coherence*. Also, the reference is coded as Negative. In Example 3, a comparison is made between the speakers, pointing to a difference between them as regards *vocabulary*. The comment is thus coded as *vocabulary range* (vocabulary is limited), in addition to *inter-candidate comparison*. It is also coded as Mixed (limited vocabulary, but slightly broader than partner's).

Example 2 Both speakers gave the same impression: not terribly talkative, a bit shy perhaps, but positive towards each other and the texts. They didn't have much to say about the topics.

Example 3 The speaker's vocabulary is limited but slightly broader than that of the female speaker's.

A third additional coding category was named *rater reflection*. When reading through the comments, I realised that some of them had the character of inferences or discussion. These instances were named *rater reflection*, and were divided into three main groups, referring to (1) *the matching of candidates*, (2) *the rating decision*, and (3) *rater reflections in general*. In this category, evaluative responses were only coded for when the reflection referred to another category (for example interaction). In all other cases, rater reflection comments were not coded for evaluative response (or any other categories). Two examples below are given to illustrate. In Example 4, no reference is made to a specific category, and hence no evaluative response is coded. In Example 5 (translated in parenthesis), on the other hand, there are references to *language* (coded as *general linguistic range* under the main category *range*) and *pronunciation* (coded as *phonological control* under the main category *accuracy*). Furthermore, these two references are coded as Negative. The rest of the segment is the rater's reflection on the grade, which follows from the linguistic aspects.

Example 4 I feel he could have performed better with a more collaborative partner with better contributions.

Example 5 Hans språk är inte det bästa, och inte heller hans uttal. Men han förtjänar ett högre betyg med tanke på innehållet. (His language is not the best, and neither is his pronunciation. But he deserves a higher grade considering the content.)

As already mentioned, the complete coding scheme is provided in Appendix 7.

Inter-coder agreement

To check the reliability of the coding, an assistant researcher, not connected to the study, who has long experience working with the CEFR scales, as well as with the Swedish performance standards for course *English 6*, co-coded about 10% of the raters' summary comments.

The inter-coder agreement achieved was about 85% on main categories indicating a satisfactory level of agreement. For subcategories, the agreement rate was naturally a little lower. Segments on which there was disagreement were carefully considered and discussed with the co-coder. In some cases this led to changes in the coding scheme.

One of the categories that there was disagreement about was *Sociolinguistic appropriateness* and *Flexibility to circumstances*, which we had interpreted in different ways, and perhaps not even used in a consistent way. Therefore I once again studied the scales in the CEFR for *Sociolinguistic appropriateness* and *Flexibility* (categorised as part of discourse competence, which in turn is a subcategory of pragmatic competence) and found that they are very similar. The two scales are provided in Appendix 8.

One example of a segment that we had coded differently is found in Example 6 from a Swedish rater (translated in parenthesis):

> Example 6 Talaren anpassar samtalet till syfte, mottagare och situation. (The speaker adapts the conversation to purpose, recipient and situation)

It was mainly Swedish raters who commented on test-takers' ability to adjust what he/she says to purpose, recipient and situation, since this is mentioned in the assessment factors (Appendix 3). I had coded this segment as *Flexibility to circumstances*, whereas the co-coder had coded it as *Sociolinguistic appropriateness*. I decided to keep this and all similar comments where raters commented on adaptation to purpose, recipient and situation as *Flexibility to circumstances*. I argue that this comment is in line with the B2+ descriptor in the scale for *Flexibility to circumstances*:

> **Flexibility**
> Can adjust what he/she says and the means of expressing it to the situation and the recipient and adopt a level of formality appropriate to the circumstances. (Council of Europe, 2001, p. 124)

Example 7 provides another illustration of a similar comment:

> Example 7 Appropriate language? Sucks.

In Example 7, the rater is asking whether the candidate is using appropriate language, meaning that "sucks" might be too informal in this situation. The co-coder coded this instance as *Sociolinguistic appropriateness*, whereas I had coded it as *Flexibility to circumstances*. However, on closer inspection both the descriptor for *Flexibility to circumstances* for the B2+ level, shown above, and the same descriptor for *Sociolinguistic appropriateness* could be relevant in this situation:

> **Sociolinguistic appropriateness**
>
> Can express him or herself confidently, clearly and politely in a formal or informal register, appropriate to the situation and person(s) concerned. (Council of Europe, 2001, p. 122)

In this case, I therefore chose to code it as both *Sociolinguistic appropriateness* and *Flexibility to circumstances*. This has also been done in similar comments, where it is difficult to say whether the rater is talking about *Sociolinguistic appropriateness* or *Flexibility to circumstances*.

Example 8 provides a last example of a segment that was double-coded as both *Sociolinguistic appropriateness* and *Flexibility to circumstances*, since the comment refers to both appropriateness of language use and level of formality appropriate to circumstances.

Example 8 well-adapted and appropriate "comfort zone, inappropriate, twisted role models". (apart from CRAP ☹ – but he is aware of it and apologizes! ☺)

Coding of summary comments

After segmentation, the summary comments were coded using the coding scheme. In Example 9, the coding of one rater's summary comment is provided (Coding scheme with a key to the codes is provided in Appendix 7). It is from a Swedish rater and his/her comment on candidate 1, female student. It is the same comment that was shown above in the section on segmentation of summary comments.

Example 9

General communication skills are good/**TR:OV/Pos**
she has fluency/**FL:FLU/Pos**
and structure./**CO:CC/Pos**
She listens to what the male is saying and as the conversation develops she acknowledges his thoughts and even adds her own opinion to the subject at hand. She even puts the question back to him for further discussion./**IN: COOP/Pos, RA: EXP/Pos**
Her vocabulary, phraseology and idiomatic expression are good./**RA:VOC/Pos**
She has a problem with some words, cult, busy, essay, symbol and students but this is only a disruption in communication./**AC:PC/Mix**
This is weighed up by her depth and breadth of the content of what she is saying./**CO:TOD/Pos**

> There is some complex explanation of her opinion in a couple of places, for example "chat" and "facebook"/**RA:EXP/Pos**

As can be seen, the fourth segment in Example 9 is double-coded, because the rater talks about two categories/aspects in the same segment, and it is not possible to split the segment into two, since it represents one main idea. This was done consistently when relevant.

Use of computer-assisted qualitative data analysis software

The software program NVivo 10, a software package designed to assist in qualitative data analysis, was used to organise the data. Summary comments were entered into NVivo 10, then segmented and coded. Bringer, Johnston, and Brackenridge (2004) emphasise that use of this kind of software, sometimes referred to as Computer Assisted Qualitative Data Analysis Software (CAQDAS), is a research tool and not a methodology. The authors stress that this kind of research tool does not do the analysis for the researcher, but it helps in quantifying qualitative data. However, "[t]he researcher must still interpret, conceptualize, examine relationships, document decisions, and develop theory. The computer can assist in these tasks but by no means does the computer analyse qualitative data" (p. 249). This danger of using the research tool in the wrong way is important to keep in mind. NVivo 10 provides many possibilities for exploring different links between the data, which is very intriguing but must never become a substitute for the actual analysis.

Methodological considerations

This section briefly touches upon issues of reliability and validity in relation to the research design of the present study. First, the quantitative methods are discussed, then the qualitative. Finally, some general remarks are made.

Validity and reliability of the quantitative method

Because of the small sample size – 17 Swedish raters and 14 external ones – it is not possible to generalize the results over populations and settings. In other words, external validity, and hence generalisability, is limited. Moreover, the selection of participants is not random or representative, which is an additional threat to validity. However, even though participants are not randomly selected, care was taken to invite potential participants from different upper secondary schools in two different cities in Sweden. Hence, the Swedish raters were from different schools within two different cities. This by no means makes the sample representative but at least there is some geographical distribution in the material. For practical reasons it was not possible to select participants for the two external groups of raters in the same way as for the Swedish raters. The external raters were invited to participate in the research study by personal contacts in Spain and Finland, respectively. My contacts were responsible for inviting potential participants and made sure the group consisted of raters with experience of the CEFR, which was the main requirement.

Scores are analysed using descriptive, correlation and reliability statistics to examine rater profiles and issues of variability. As mentioned above, validity will be limited, because of the small sample investigated. Nonetheless, using statistical analyses provides a useful complement to the qualitative analysis of rater comments.

The reliability of the scores that the raters produced could have been affected by the fact that they had limited time to make their decisions and write comments. However, in all rating situations time is an issue that can affect reliability. The main aim was to make the data collection procedure as authentic as possible, thus resembling a rating situation that raters were used to. Moreover, the raters sat at a stretch with breaks for coffee and lunch; of course, this could also affect the reliability of scores. Still, once again, this is similar to a "real" rating situation.

Validity and reliability of the qualitative method

In this study, only audio-recorded material was used. In previous studies, video-recordings have been used as well. Findings indicate that raters find body language to be an important feature of interactional competence, even though it was not stated in the rating criteria that body language was to be observed. In the present study, it is not possible to draw any such conclusions, since raters cannot see candidates' body language. However, it is not stated in the rating criteria or descriptors that body language should be considered; in other words, this is a matter of the definition of the construct. Moreover, body language also introduces another element of interpretation, with obvious effects on validity as well as reliability.

As the number of raters in the study is small, the generalisability of the results from the analysis of the verbal reports will obviously be limited. However, using a qualitative method like verbal protocol analysis, in which verbal reports are segmented and coded, takes much time. Therefore, it was not possible to include more raters in the study. The results from the verbal report data can be seen as an illustration of *some* Swedish, Finnish and Spanish raters' decision-making in reaching a judgement on paired oral discussions. The findings can then be related to previous studies of rater orientations to see if they coincide with or differ from the results of the present study.

The reliability of coding and analysing verbal reports can also be questioned, as mentioned above, since it involves subjective judgements on the part of the researcher. However, reliability can be strengthened if a second opinion is used for parts of the material, i.e. a co-coder. In the present study, as described above, an external coder co-coded 10% of all the rater comments. Overall, inter-rater agreement was about 85%, which is satisfactory.

Closing remarks on validity and reliability

As for construct validity, the present study aims to analyse the rating process, i.e. what features of communicative language ability raters pay attention to while forming their judgements, and the rating product, i.e. the scores. Consequently, the study includes both 'qualitative' data in the form of summary comments, and 'quantitative' data in the form of scores. The fact that there are data reflecting both rating process and rating product strengthens validity, since rating is approached from a broader perspective – not just scores, but also what lies behind the scores.

Moreover, data collection procedures are standardised and controlled as far as possible. Data were collected during one day for each group (two groups of Swedish raters and two groups of European raters). The groups had exactly the same set-up for this day with (1) introduction, (2) individual rating, and (3) short group discussion and conclusion.

Ethical concerns

Informed consent and confidentiality

The participants, i.e. the raters, were informed about the conditions of the research project. Participation was voluntary. Further, the participants could withdraw from the project at any time.

The material consists of audio-recordings of paired conversations from the development phase of the Swedish national test of English from spring 2011. The students have given their consent to the material being used for research purposes.

The data collected from the participants, i.e. rater notes, summary comments, and scores, have been used in accordance with the guidelines by the Swedish Research Council. Schools and raters are kept anonymous throughout the process, including the presentation.

Chapter Five: Results

In this chapter, the results are accounted for. First, results from the analysis of quantitative data are presented, and then results from the analysis of the qualitative data.

Descriptive statistics for Swedish raters

The question relevant to the Swedish raters' judgements is research question 1: What can be noticed regarding variability of scores and consistency of rater behaviour?

To start pursuing this question, descriptive statistics for the Swedish raters' scores (*n* = 17) were explored. In Table 5 below, descriptive statistics for ratings per candidate (*N* = 12) are given. Each candidate has a code, for example C1F and C1M. This is to be understood as candidate one, female student and candidate one, male student, etc.

Table 5. Descriptive statistics: ratings per candidate (*N* = 12) for Swedish raters (*n* = 17)

Candidate	Mean	SD	Median	Mode	Range
C1F	5.9	1.5	6.0	6	(4–8)
C1M	7.4	1.5	7.0	9	(5–9)
C2F	9.1	0.8	9.0	9	(7–10)
C2M	8.0	1.5	8.0	8	(4.5–10)
C3F	4.9	1.7	5.0	3	(3–8)
C3M	6.4	1.5	7.0	7	(3–9)
C4F	3.4	1.0	3.0	3	(1–5)
C4M	2.9	1.0	3.0	3	(1–5)
C5F	9.4	0.5	9.0	9	(9–10)
C5M	7.1	1.1	7.0	7	(5–9)
C6F	8.2	1.1	9.0	9	(6–10)
C6M	7.3	1.3	7.0	7	(4–10)

As can be seen in Table 5, the mean, median and mode range 3-9 for the twelve performances. It is interesting to note that there are no performances with a mean, median or mode below 3, which is a fail grade in the Swedish rating scale. When it comes to range, there are some clear instances of variability. The performances with most variability are C3M, who displays a range of scores

3-9, and C6M, with a range 4-10. These two instances will be particularly interesting to examine in relation to the qualitative data in the form of summary comments. How does the rater who judges this performance to be a 3 (corresponding to E-) justify this score, compared to the rater who judges the same performance to be a 9 (corresponding to a B)? An opposite example would be C5F, where all the raters seem to agree that she is either a 9 or a 10 (corresponding to B or A). This is also an interesting example to examine in the qualitative data analysis and see whether raters notice the same features of the performance, since they seem to agree on the mark, or whether they notice different features but still award the candidate the same score.

To further illustrate the issue of variability, a graph showing median and range of scores for the twelve candidates is presented in Figure 6.

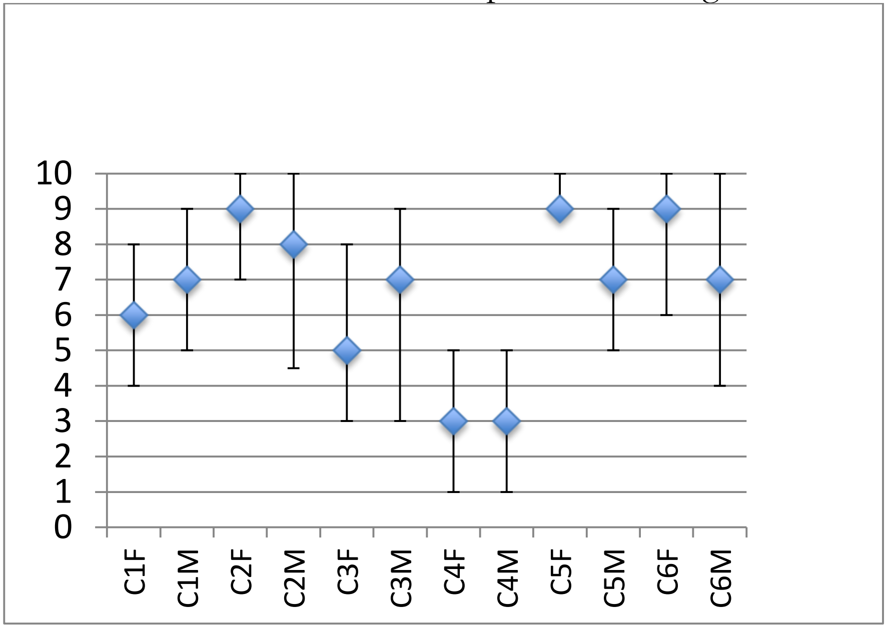


Figure 6. Median and range per candidate (*N* = 12)

As can be seen in Figure 6, there is a considerable degree of variability. As mentioned above, C5F has the smallest range, whereas C3M and C6M have the largest range. The problematic issue about range, however, is that it is determined by two scores in the distribution, namely the highest and the lowest. If there are outliers in the material, for example one rater with an extreme score, this will influence range. In other words, we need to examine all the raters' scores for each performance in order to see whether there is a case of extreme scores. To give one example of this, Figure 7 shows the distribution of scores for C3M:

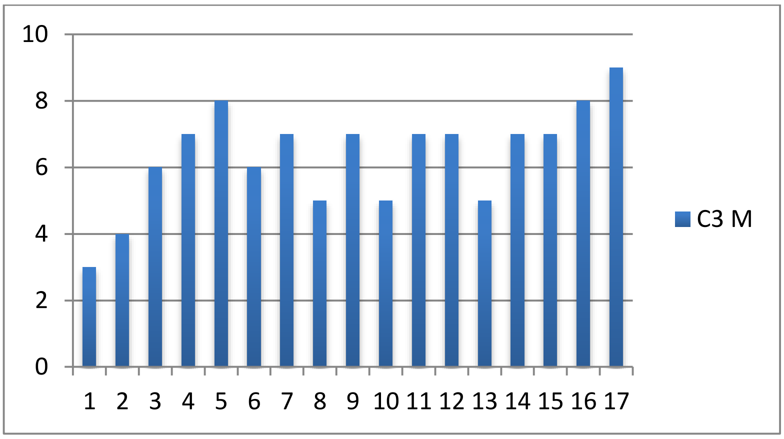


Figure 7. Distribution of scores (*n* = 17) for C3M

As can be seen in Figure 7, raters 1 and 2 have awarded C3M quite low scores (3 and 4, respectively), compared to the rest of the rater group. In other words, raters 1 and 2 seem to be more severe in their rating of C3M than the other raters. In Appendix 9, the distribution of scores per candidate is provided for all test-takers.

As a complement to descriptive statistics for ratings per candidate, descriptive statistics for the Swedish raters (*n* = 17) are shown in Table 6[10].

[10] In three instances, two raters could not decide on a mark and thus awarded a "double" score for the same performance (for example B/A). This was done once by one rater and two times by a second rater. In these three instances, half a point was used, for example 9.5, if the rater awarded a double score of 9-10.

Table 6. Descriptive statistics for Swedish raters (*n* = 17)

Rater	Mean	SD	Median	Mode	Range
R1	5.9	2.7	7.0	3	(2–9)
R2	5.6	2.6	5.5	4	(1–9)
R3	6.3	2.1	6.5	5	(3–9)
R4	7.4	2.2	7.5	7	(3–10)
R5	8.0	2.3	8.5	8	(3–10)
R6	6.7	1.7	6.5	6	(4–9)
R7	6.3	2.0	6.5	7	(3–9)
R8	6.0	2.7	6.5	5	(1–9)
R9	6.3	2.5	7.0	7	(2–9.5)
R10	7.3	2.5	8.0	8	(3–10)
R11	6.8	2.4	7.0	7	(3–10)
R12	6.3	2.3	7.0	9	(3–9)
R13	6.9	2.9	8.5	9	(2–10)
R14	6.3	2.5	6.5	6	(2–9)
R15	6.3	2.4	6.5	7	(2–10)
R16	7.8	1.8	8.0	8	(4–10)
R17	7.2	2.4	7.0	7	(3–10)

In Table 6, we can observe rater profiles with differences in severity/leniency. Rater 2 seems to be the harshest rater, with a mean of 5.6, whereas rater 5 is the most lenient with a mean of 8.0. However, Figure 8 shows that the Swedish rater group as a whole have a fairly even distribution of their mean. In other words, variability exists, but does not seem to be excessively large.

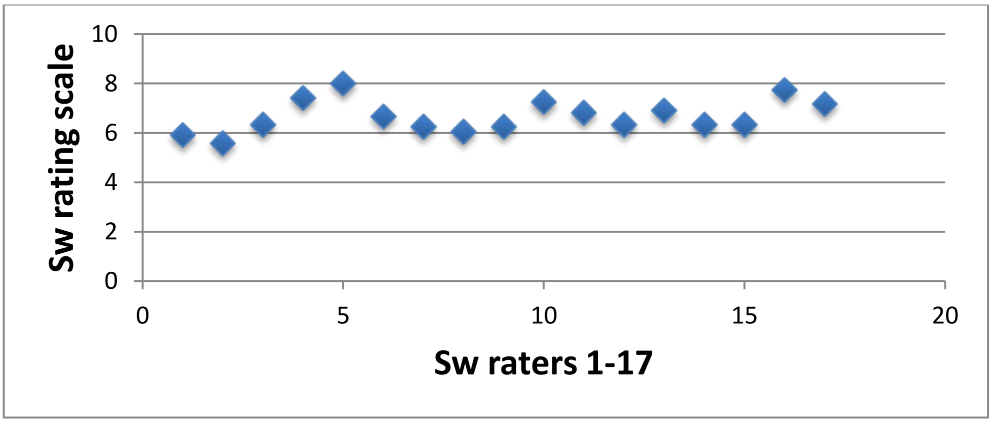


Figure 8. Means of Swedish raters' scores

A final illustration of variability of the ratings is provided in a box plot for the Swedish raters (*n* = 17) in Figure 9.

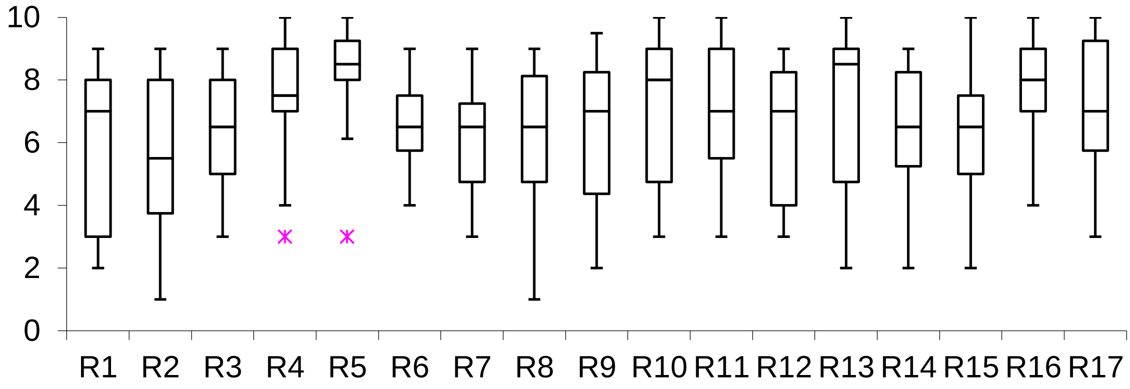


Figure 9. Box plot for Swedish raters (*n* = 17)

Medians are denoted by solid black lines while the top and bottom box edges denote the first and third quartile. Whiskers denote the largest and smallest data within 1.5 times the interquartile range

As illustrated in Figure 9, most raters use the range of the scale (1-10), but there are some exceptions. Most notably, raters 4 and 5 have boxes that are rather small and high up on the rating scale (indicating a general tendency to award high scores). There are also outliers in the scores of rater 4 and 5, showing that they have awarded one score each that is extreme compared to the rest of their scores. In this case the outliers are at the lower end of the scale.

Finally, four examples of rater profiles are shown in histograms in Figure 10. These four raters were chosen as examples based on the results of the box-plot above. A type of rater effect identified in performance testing is *restriction of range* (Wilson & Case, 2000), which refers to overuse of certain categories, for example if raters concentrate their scores to the lower or higher end of the scale. *Central tendency* is the most common type of restriction of range and applies to raters who use predominantly the middle categories, thus avoiding extreme categories. Rater 5 is an example of a fairly lenient rater with most of the scores at the higher end of the scale, thus displaying signs of restriction of range. Rater 15, in comparison, has a fairly even distribution of scores, however with a slight tendency to award scores in the middle of the scale (central tendency). Raters 2 and 1 are also interesting to compare. Whereas both users use the range of the scale, they have different peaks in the histogram.

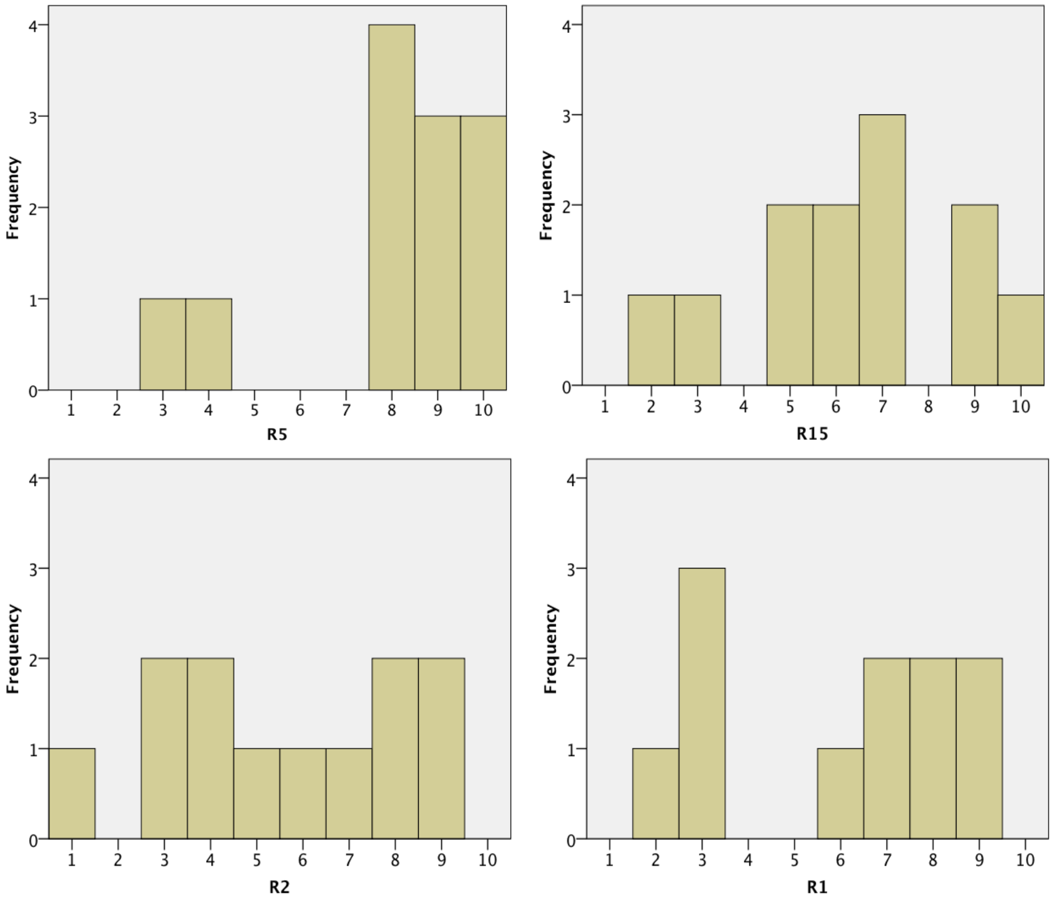


Figure 10. Examples of rater profiles based on score distribution

In summary, some general remarks may be made regarding the variability of ratings and severity of the Swedish raters. Based on the descriptive statistics, it can be seen that there are clear rater profiles with differences in leniency and severity. For example, the means of the scores vary between 5.6 and 8.0 on the ten-point scale. It is also obvious that some performances are more difficult to agree on than others. This will be especially interesting to explore in the qualitative analysis of raters comments.

Inter-rater reliability of Swedish raters

In order to compute correlations, the data were entered into SPSS. Correlation analyses measure the strength of the relationship between two variables. In this case the variables are pairs of Swedish raters' scores. Two main measures of non-parametric rank correlations are presented, namely Spearman's (rho) rank correlation coefficient and Kendall's Tau rank correlation coefficient. Both measures assess statistical relationships based on the rank order of the data.

Spearman's rank correlation coefficient is the most widely used measure. The main differences between the two measurements are that Kendall's Tau usually generates lower values than Spearman's rho. Furthermore, calculations are based on concordant and discordant pairs, whereas for Spearman's rho, calculations are based on deviations. Finally, and importantly, Kendall's Tau is more insensitive to error as compared to Spearman's rho, and *p*-values are more accurate even with small sample sizes.[11] It was decided that both measurements should be used in the present study, Spearman's rho being the most common, while Kendall's Tau is a stricter measurement, taking more parameters into account and so being regarded as superior to the Spearman values.

In the first step, Spearman's rank order correlation coefficients were computed for the 17 Swedish raters. A table with the correlations is provided in Appendix 10. Significant pair-wise correlations between the Swedish raters have a range between .59 and .95 ($p < .05$). The lowest correlation is between rater 9 and 16 (.59), whereas the highest can be found between raters 9 and rater 10 (.95). There were also a few non-significant correlations ranging from .39 to .56. In order to get an overview of the correlations, the median was computed with a value of .77, indicating reasonably satisfactory inter-rater consistency.

In addition, Kendall's *tau-b* coefficients were computed for the 17 Swedish raters' scores (see Appendix 10). As expected, these correlations were somewhat lower than the Spearman correlations, ranging from .47 to .89 ($p < .05$). As with the Spearman rank order correlations, there were a few instances of non-significant correlations, ranging from .30 to .44. The median was calculated at .66.

As a final step, Cronbach's Alpha, measuring internal consistency of the whole group, was calculated. Cronbach's Alpha was .98 for the whole group of Swedish raters, which indicates stable consistency.

In summary, the calculated Spearman rank order correlation coefficients for the Swedish raters with a median of .77, and the calculated Kendall's Tau rank order correlation coefficients with a median of .66, indicate reasonably satisfactory rater agreement, although no set figures can be given to define 'good agreement'. In addition, Cronbach's alpha was very high, .98 for the whole group, implying stable internal consistency.

[11] Information retrieved from www.statisticssolutions.com, "Kendall's Tau and Spearman's rank correlation coefficient", 2014.

Descriptive statistics for external CEFR raters

The research question relevant to the external CEFR raters' scores is number 4: At what levels in the CEFR do external raters judge the performances of the Swedish students to be? To answer this question, the statistical presentation in this section focuses on the external raters' (*n* = 14) scores in relation to the CEFR scale. As a result, this section is structured in a different way as compared to the previous section about Swedish raters' statistics. First, descriptive statistics for ratings per candidate are given, followed by a comparison of the rank order of performances for the European and Swedish raters.

To start with, descriptive statistics for the CEFR raters' (*n* = 14) scores per candidate (*N* = 12) are shown in Table 7. Since the research question concerns at what levels in the CEFR the external raters judge the performances to be, mean and median scores are the focus of the analysis and interpretation. It needs to be borne in mind that the CEFR scale is a nine-point scale, whereas the Swedish scale is a ten-point one; consequently, no direct comparison can be made between the values of the two.

Table 7. Descriptive statistics: ratings per candidate (*N* = 12) for CEFR raters (*n* = 14)

Candidates	Mean	SD	Median	Mode	Range
C1F	5.6	1.2	5.5	5	(4–8)
C1M	6.3	1.1	6.0	6	(4–8)
C2F	7.9	1.1	8.0	9	(6–9)
C2M	7.3	1.1	7.0	7	(6–9)
C3F	4.9	1.4	5.0	5	(3–8)
C3M	5.6	1.2	5.0	5	(4–8)
C4F	3.9	1.1	4.0	4	(2–6)
C4M	3.5	1.1	4.0	4	(2–5)
C5F	7.6	1.2	7.5	7	(5–9)
C5M	5.6	1.5	6.0	6	(4–8)
C6F	5.8	1.3	6.0	6	(4–8)
C6M	5.3	1.4	5.5	4	(4–8)

The speaking test used in the present study is intended to correspond to level B2 in the CEFR, with a minimal pass corresponding to B2.1. In the rating scale that the European raters used, plus levels were included, which means that B1+, corresponding to a five on the nine-point scale, may be an acceptable cut-off point for the absolute minimum passing level of the test. In Table 7, it is shown that C4M was awarded the lowest scores by the external CEFR raters with a mean of 3.5. The candidate with the highest mean, 7.9, was C2F. In other words, the candidate with the lowest mean was, based on the average ratings, at the B1 level (a four in the rating scale if 3.5 is rounded off). In comparison, the candidate with the highest mean score was at the C1 level (7.9 rounded off to 8). Consequently, there is a range in the performances from B1 to C1.

If we count B1+, corresponding to a five on the nine-point scale, as an acceptable minimum level to pass the test, in line with the reasoning above, we can see that ten of the twelve performances were on average rated at or above B1+. There are two performances, whose means are below the intended level of the test, namely C4F and C4M. C3F is a borderline case. Her mean is 4.9, which, if rounded off is 5, i.e. B1+. When looking at the Swedish raters' statistics, we can see that C4F and C4M have a mean of 3, which corresponds to a low passing grade (E-). The range of the ratings, however, is 1-5, which shows that there are some Swedish raters who awarded a Fail to these two candidates. As for C3F, the range for the Swedish raters' scores is 3-8, which suggests that they value this performance as a pass and thus at the B2.1-level. C3F, then, is an interesting case to follow up in the qualitative analysis of rater comments, to see whether the Swedish and the CEFR raters comment on this performance in different ways.

Considering the median, we can see that two performances are at B1 (4), two performances at B1+ (5), five performances at B2 (6), one performance at B2+ (7), and two performances at C1 (8). These results may suggest that the CEFR raters are somewhat harsher around the cut-off point, or minimal level, since as many as four candidates were rated as B1 or B1+ when looking at the median. For the Swedish raters there were no performances with a mean or median below the minimum passing grade.

In summary, the results from the analysis of the external CEFR raters' scores show that the rank ordering of performances is fairly similar between the Swedish and the CEFR raters. In addition, the means of the CEFR raters' scores are between B1+ and C1 for all performances but two, which is well in line with the intentions of the test. Also, the two performances rated lower than B1+, were rated as a Fail by some of the Swedish raters, suggesting that these two candidates' performances were borderline cases.

Finally, the Swedish and external CEFR raters were compared with regard to how they had ranked the performances. This aspect is interesting to investigate, since it allows for a comparison between the Swedish and the CEFR raters, even though they use different rating scales. The results are shown in Figure 11.

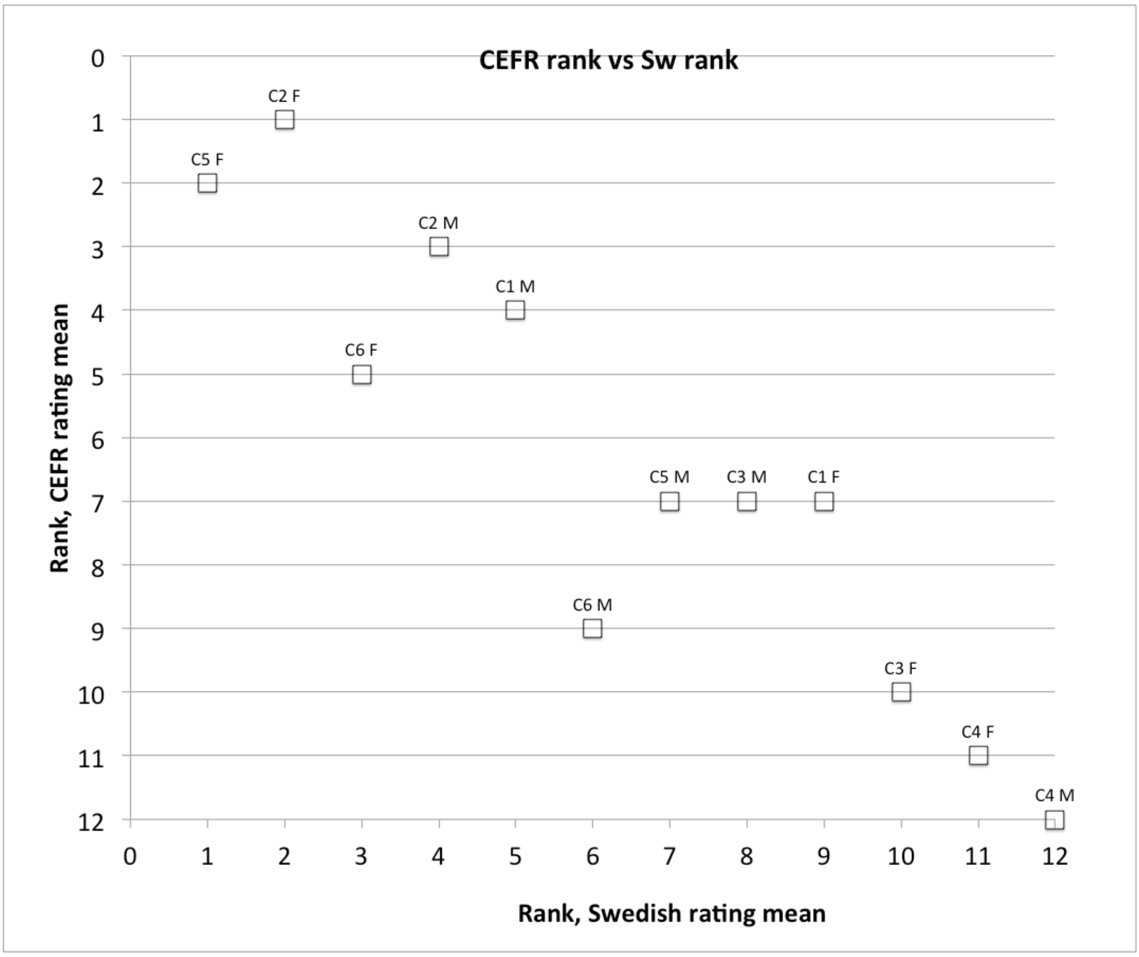


Figure 11. Comparison of rank orderings (CEFR vs Swedish raters)

As can be seen in Figure 11, the rank ordering, based on means, is fairly similar between the Swedish and the external CEFR raters. It is worth noting that C2F is ranked highest among the CEFR raters, but second among the Swedish raters. The opposite relationship applies to C5F, who is ranked number one by the Swedish raters and number two by the CEFR raters. We can also see that three performances were ranked equally high by the CEFR raters, namely C1F, C3M and C5M. The Swedish raters ranked these performances as seven, eight and nine, respectively. A considerable difference is that C6F is ranked high among the Swedish raters, in position three, whereas the CEFR raters rank this performance as number five. This example will be examined further in the qualitative analysis to see how the CEFR raters comment on this performance in comparison with the Swedish raters.

In summary, the results show that both rater groups, from different educational systems and backgrounds, rank the performances in a fairly similar way.

Analyses of written rater comments

To answer the second research question regarding what features of candidates' performance are salient to raters as they make their rating decision, the written summary comments were analysed both qualitatively and quantitatively. The qualitative analysis involves segmentation and coding of the written comments. The coded data were then tallied and percentages were computed for each coded category and for the two groups of raters (i.e. Swedish raters and external CEFR raters). When tallying the frequency of coded comments within the main and subcategories, each coded comment was only counted once for both main category and subcategory. This section starts with quantitative results, after which examples from the qualitative analysis of categories are given.

Quantitative results are presented by means of statistics for the coded categories in Table 8.

Table 8. Frequency counts and percentage of coded comments across rater groups

	Acc*	Coh	Flu	Intell	Inter	Other	Strat	Range	Soc.li	Task	Total
Swe **(*n* = 17)**											
Freq.	385	261	157	39	219	18	78	289	18	130	1594
%	24%	16%	10%	2%	14%	1%	5%	18%	1%	8%	100%
CEFR **(*n* = 14)**											
Freq.	159	97	166	29	154	5	21	174	1	55	861
%	18%	11%	19%	3%	18%	1%	2%	20%	0%	6%	100%
Total **(*N* = 31)**											
Freq.	544	358	323	68	373	23	99	463	19	185	2455
%	22%	15%	13%	3%	15%	1%	4%	19%	1%	8%	100%

** Categories in the following order: Accuracy, Coherence, Fluency, Intelligibility , Interaction, Other, Production strategies, Range, Sociolinguistic appropriateness, Task realisation*

As shown in Table 8, the largest groups of comments relate to *accuracy* (22%), *range* (19%), *coherence* and *interaction* (15% respectively), and *fluency* (13%). *Task realisation* comprises about 8% of the comments. Finally, the three last categories are very small: *intelligibility* (3%), *other* (1%) and *sociolinguistic appropriateness* (1%).

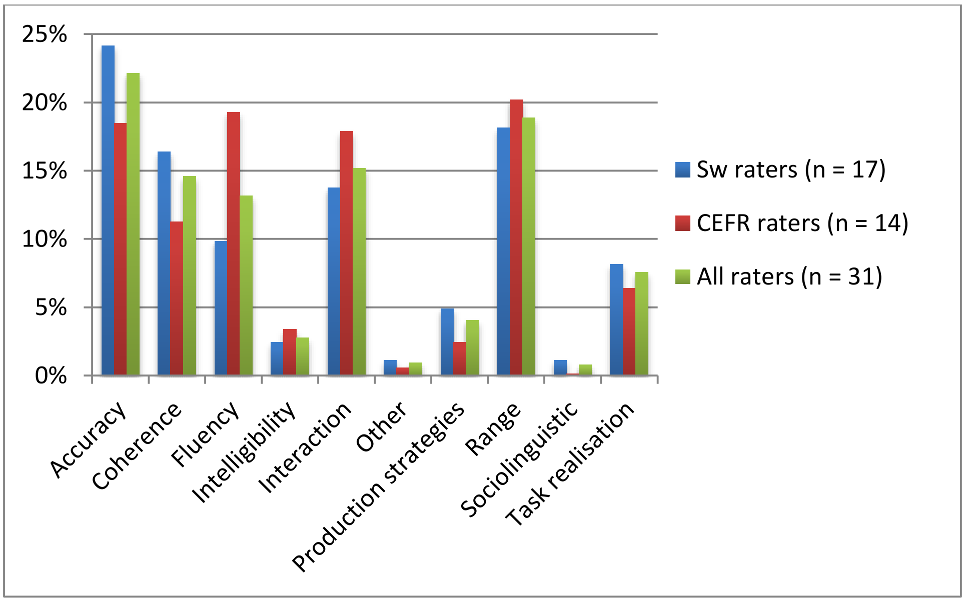


Figure 12. Distribution of comments coded for the main categories

The same data as in Table 8 are graphically illustrated in Figure 12, showing the distribution of comments coded for the main categories. The three bars show the Swedish and the CEFR raters, as well as the total, i.e. both groups considered together. When looking at the two groups separately – Swedish raters and external CEFR raters – somewhat different rater orientations can be noticed, as illustrated in Figure 12.

To further highlight the differences in rater orientations, a comparison of the Swedish and CEFR raters' distribution of coded comments is presented in Table 9.

Table 9. Comparison of rater orientations between Swedish and CEFR raters

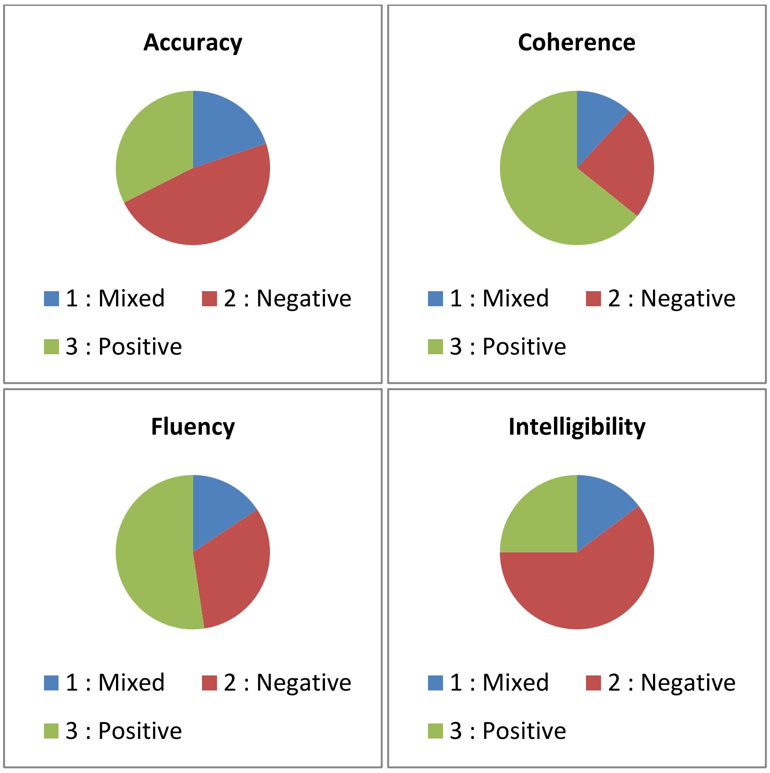
Swedish raters	**CEFR raters**	**Total**
1. Accuracy (24%)	1. Range (20%)	1. Accuracy (22%)
2. Range (18%)	2. Fluency (19%)	2. Range (19%)
3. Coherence (16%)	3. Accuracy (18%)	3-4. Coherence (15%)
4. Interaction (14%)	4. Interaction (18%)	3-4. Interaction (15%)
5. Fluency (10%)	5. Coherence (11%)	5. Fluency (13%)
6. Task realisation (8%)	6. Task realisation (6%)	6. Task realisation (8%)
7. Strategies (5%)	7. Intelligibility (3%)	7. Intelligibility (3%)
8. Intelligibility (2%)	8. Strategies (2%)	8. Strategies (4%)
9-10. Other (1%)	9. Other (1%)	9. Other (1%)
9-10. So-li (1%)	10. So-li (0%)	10. So-li (1%)

Table 9 shows that *accuracy* is the largest category for the Swedish raters, indicating that *accuracy* has an important role in the decision making process. However, *accuracy* is not the most salient feature for the external CEFR raters. Instead, *range*, *fluency*, *accuracy* and *interaction* have a fairly similar proportion of the comments, making them appear almost equally important in the decision-making process of the CEFR raters. *Coherence* plays a slightly smaller role (11%), but is still a major category. Another observation about the CEFR raters is that the analytic criteria in their rating scale, Table C2 from the Manual (Appendix 4), namely *accuracy*, *fluency*, *coherence*, *interaction* and *range*, together with *production strategies* and *sociolinguistic appropriateness* (which are more or less embedded in the descriptors), comprise a vast majority of the coded comments (about 90%). The rest of the categories, which can be considered to be non-criterion features, i.e. not explicitly mentioned in the descriptors, are a small group (about 10%) and thus seem less salient.

If we return to the results of the coded comments for the Swedish raters, it is clear that *accuracy* seems to be the most salient feature (24%). An interesting difference between the rater groups is that *fluency* was not commented upon as much by the Swedish raters (10%) as it was by the external raters (19%), suggesting that this feature is less important to the Swedish raters. On the other hand, *coherence* seems to be slightly more salient to the Swedish raters (16%) than to the CEFR raters (11%). Further, *range* (18%), *coherence* (16%) and *interaction* (14%) seem to play an equally important role in the rating decision of the Swedish raters, while that of *fluency* seems to be slightly smaller (10%). A result similar to that of the CEFR raters is that the categories that are not explicitly mentioned in the criteria have the lowest proportion of comments (about 11%).

In sum, there seem to be many different performance features taken into account as raters make their holistic decisions. Even though *accuracy* does play an important part when looking at the total number of comments, the other criterion categories contribute a substantial part as well. Moreover, comments that belong to features not explicitly stated in the descriptors were rather few. It was also clear that the Swedish and CEFR raters differed somewhat in rater orientation. In other words, they seemed to favour slightly different performance features over others.

As the next step, the number of Positive, Negative or Mixed comments (i.e. evaluative response) in the main categories was checked. The results, calculated for all raters (*N* = 31), are shown in Figure 13 below.



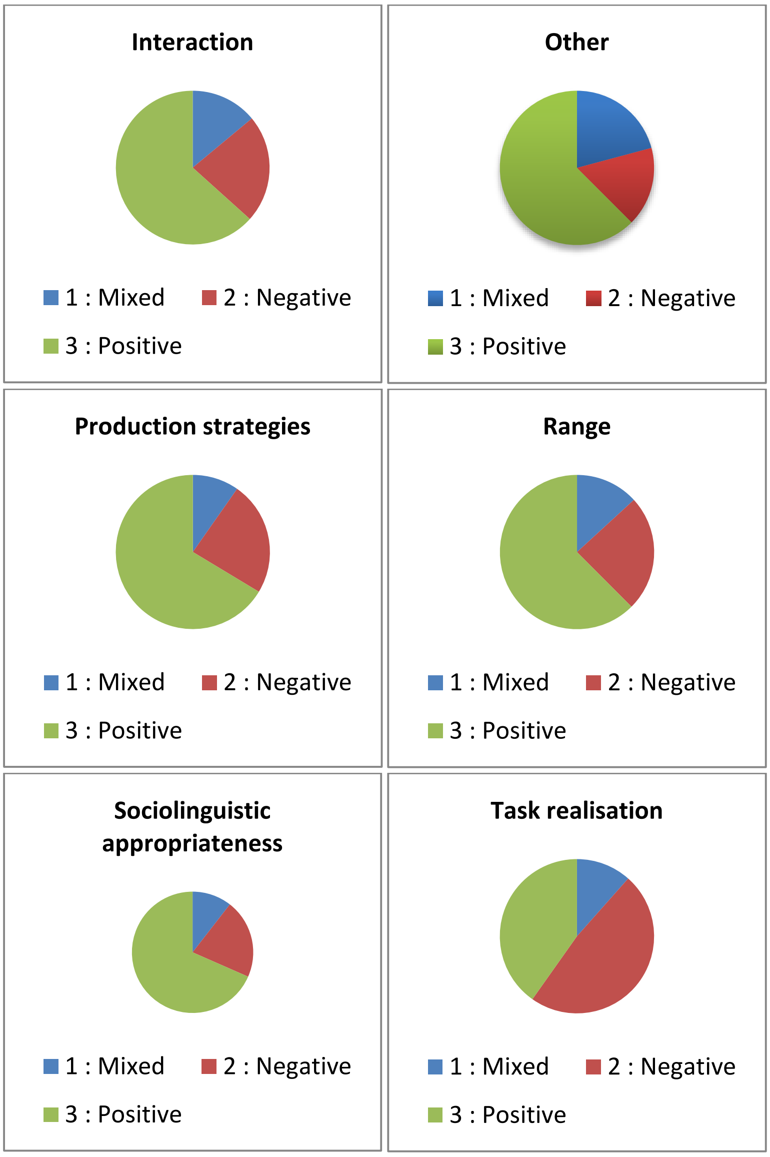


Figure 13. Evaluative responses per category

In Figure 13, it is shown that there are three categories where the results differ most from the general pattern, namely *accuracy*, *intelligibility*, and *task realisation*, for which a majority of the comments are Negative. For the other categories, the pattern looks similar with Positive as the largest category, Negative as the second largest and Mixed as the smallest. *Fluency* stands out somewhat with a quite large proportion of Negative comments (32%). For the other groups, Negative comments vary between 17% and 24%.

Comments per category

In this section, examples will be given from the rater comments to illustrate and explore the categories. For every main category with subcategories, a graph is provided, showing the distribution between the subcategories as well as evaluative responses (Positive, Negative and Mixed) per subcategory. For some categories, however, there are no subcategories. In those cases, reference is made to Figure 13 showing evaluative responses per category. At the end of the presentation for each category, short reference is made to the rating criteria employed by the raters: (1) analytic performance level descriptors from the CEFR scales used by the CEFR raters, and (2) holistic performance level descriptors from the Swedish performance standards for course English 6, as well as analytic assessment factors, used by the Swedish raters. The term *descriptors* is used in a general sense to refer to both CEFR scales and Swedish national performance standards for EFL.

Accuracy

Figure 14 indicates that *phonological control* had the largest number of comments, closely followed by *grammatical accuracy*, while *vocabulary control* had fewer references. Further, as regards evaluative responses, it was shown that *vocabulary control* differs from the other categories. For this category, there was a clear majority of Negative comments, whereas the other two categories had a majority of Negative comments but also many Positive and Mixed ones.

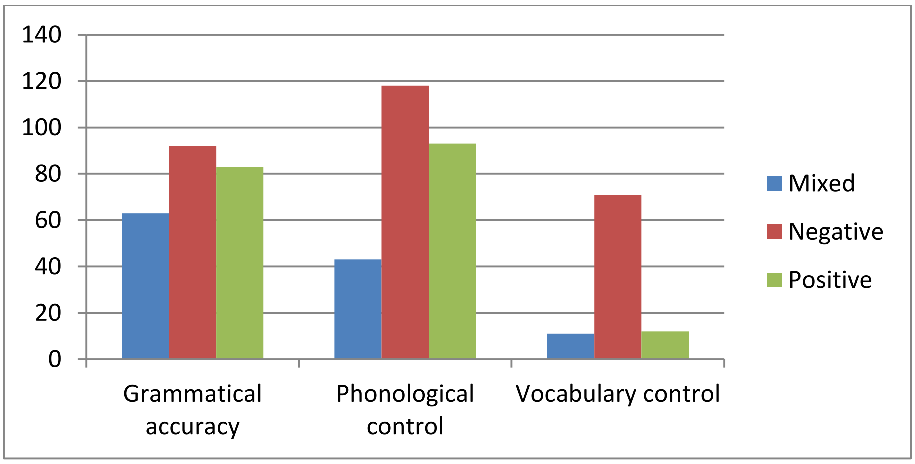


Figure 14. Evaluative responses per subcategory for accuracy

Raters commented on candidates' grammar, both in general terms (Extract 1) and with more specific examples, pointing to particular types of errors (Extract 2). The most common types of errors noted by the raters were use of verb forms, subject-verb agreement and singular/plural marking. More specifically, the comments on *grammatical accuracy* were conceptualised in terms of frequency of errors, i.e. many or few (Extract 3), and the ability to produce correct syntax (Extract 4).

> Extract 1: On the whole his grammar is correct but not very advanced (general mixed) /Sw
>
> Extract 2: Grammar seems to be an improvement area; ing-form/no ing-form is mixed at will, subject-verb agreement is a problem area. (specific negative) /Sw
>
> Extract 3: Just a few slips in grammar: Many people has…, You have lots of interest (mixed) /Sw
>
> Extract 4: Her syntax is spotless. (positive) /CEFR

Further, comments on *grammatical accuracy*, as well as *vocabulary control* and *phonological control*, were often related to their impact on *intelligibility* (Extract 5), and *interaction* (Extract 6).

> Extract 5: when he presents his card a few grammar problems make it a bit difficult to understand what he is trying to say. (negative) /Sw
>
> Extract 6: /His pronunciation is very good/ and his use of language accurate. He makes the very odd mistake, which in no case hinders communication. (mixed) /CEFR

When referring to candidates' *pronunciation*, comments addressed its *accuracy* (Extract 7) and *nativeness* (Extract 8). Comments also referred to pronunciation and its impact on *intelligibility* (Extract 9) and *interaction* (Extract 10). In addition, raters referred to test-takers' *intonation* and *accent* (Extract 11).

> Extract 7: Some mispronunciations, especially /z/. Other examples are "age", "students", "essay"… (negative) /CEFR
>
> Extract 8: Her pronunciation is near native and sounds fresh and natural at all times (positive) /CEFR
>
> Extract 9: and his pronunciation makes it hard to understand what he is trying to say sometimes. (mixed) /Sw

Extract 10: pronunciation at times hard to follow, which at times leads to communication breakdown (negative) /CEFR

Extract 11: The speaker has a strong accent which affects pronunciation at times (negative) /Sw

As shown above, *vocabulary control* is a category with a large majority of Negative comments. In this category, raters referred to *accuracy* and *precision* of vocabulary (Extracts 12-14). There are also comments on the *adequacy* of the vocabulary for the task (Extract 15), as well as the *appropriateness* of different lexical choices (Extract 16). Just as for the other subcategories with regard to *accuracy*, raters were concerned with incorrect vocabulary use and its impact on *intelligibility* or *clarity* of content (Extract 16).

Extract 12: but made mistakes with some very common words. (negative) /CEFR

Extract 13: She also uses some expressions/words incorrectly (learn to handle with money). (Tr.) (negative) /Sw

Extract 14: Her use of words is not always precise. (negative)/CEFR

Extract 15: Rather correct, reasonably good vocabulary – appropriate for the task. (positive) /Sw

Extract 16: Inappropriate or incorrect choice of words – may cause misunderstanding. (negative) /CEFR

Furthermore, raters frequently commented on candidates' use of *non-idiomatic* vocabulary and expressions (Extract 17).

Extract 17: Several non-English phrases: "It tells that", "Trying to fake us" (negative) /Sw

Accuracy is described in general terms in the CEFR descriptors, either as high degree of *grammatical accuracy* or as systematic basic mistakes. *Phonological control* and *vocabulary control* are not mentioned at all. In the holistic descriptors that the Swedish raters used, *accuracy* is not mentioned. However, in the assessment factors it is stated that raters should take grammatical structures, vocabulary and pronunciation into account. Considering this, it seems like both the Swedish and CEFR raters seem to judge all three subcategories as important even though some of them are not explicitly referred to in the criteria.

Coherence

Figure 15 shows that, in the *coherence* category, the two subcategories *topic development* and *coherence and cohesion* had about the same proportion of comments, whereas *flexibility to circumstances* was commented on less. Furthermore, for all three subcategories, the largest proportion of comments was Positive.

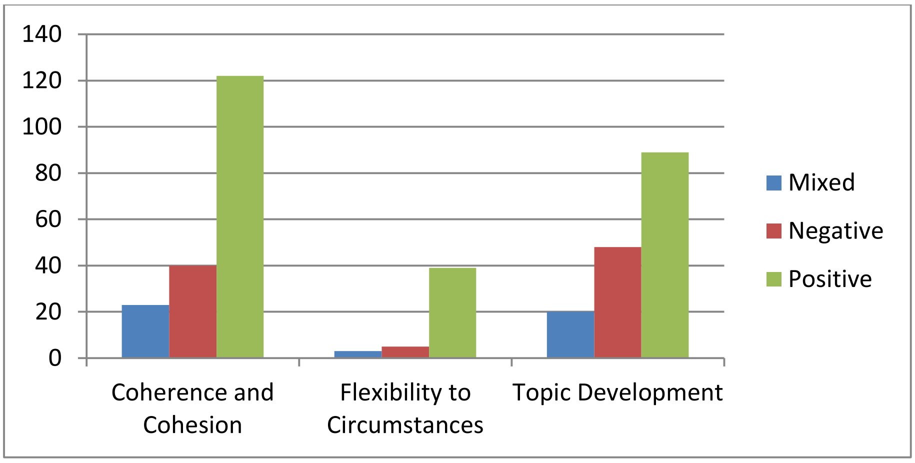


Figure 15. Evaluative responses per subcategory for coherence

Comments in this category included references to the structure and organisation of candidates' speech, *coherence and cohesion*, as well as to the development of *content* (*Topic development*). In addition, a smaller proportion of the comments referred to how well candidates can use language flexibly and adapted to the situation (*Flexibility to circumstances*).

Starting with *topic development*, raters commented on the amount of elaboration or detail in responses. There were many comments on candidates' ability to develop, give examples and new perspectives, as well as cover many aspects of the topic (Extracts 18-19). Also, comments referred to candidates' ability to argue his/her point (Extract 20). There were also some Positive examples of *topic development* that pointed to good interactional skills (Extract 21).

> Extract 18: Adds widened perspective to topic on television – How it affects young people. Nuanced. (positive) /Sw

> Extract 19: WHAT he says is interesting and good, the level of accuracy is mostly good, but he never develops the topics into any depth; not very many examples, perspectives, not very complex. (mixed) /Sw

> Extract 20: she finds it difficult to develop her arguments and opinions, maybe because she has little to say, but maybe because her English does not allow her to elaborate….. (negative) /CEFR

> Extract 21: She makes several good observations and uses examples to develop her thoughts, which moves the topics along, (positive) /Sw

While some of the comments were readily identifiable as referring to the development of content of test-taker speech, e.g. elaboration of ideas, some of the raters' comments did not appear to distinguish between content of discourse and means of expression (Extract 22). This is only natural since content development and ability to express content go hand in hand and it is not always possible to distinguish between them. As mentioned in Chapter Four: Material and method, these instances were double-coded as both *coherence* (*topic development*) and range (*ability to express viewpoints*).

> Extract 22: seems to be more solid when she begins to explain the topic herself. Nice reflections and uses personal experiences to strengthen her point. (positive) /Sw

Comments on *flexibility to circumstances* referred to how appropriate or adequate the language was in the given situation (Extract 23). There were also comments that referred to candidates' ability to adapt to speaker, situation and purpose (Extract 24). Adaptation to circumstances is explicitly stated in the Swedish criteria, which is why mainly Swedish raters commented on this.

> Extract 23: Formal, well-adapted level of English mostly but also some (*VERY*) informal expressions (*sucks*, *kind of*) too. (mixed) /Sw

> Extract 24: and with adaptation to purpose, recipient and situation. (positive)/Sw

Finally, there were some comments about candidates' ability to rephrase ideas in alternative linguistic forms, and vary formulations of what he/she wants to say (Extract 25).

> Extract 25: and rephrases what she says for her partner to understand. (positive)/CEFR

The final subcategory, *coherence and cohesion*, referred to comments about general *structure* and *clarity* of content (Extracts 26-27), *referencing* (Extract 28), and use of *cohesive devices* (Extract 29). Just as for *accuracy*, references were made to *coherence* in relation to its impact on *intelligibility* (Extract 30).

Extract 26: Coherent, structured and relaxed language is what she used. (positive) /Sw

Extract 27: In part 2, however, she sometimes has difficulty in putting her point across. On the other hand, she's good at rounding points off. (mixed) /CEFR

Extract 28: Refers back to previous discussions regularly (positive) /Sw

Extract 29: She has enough language repertoire to make her herself clear and keep going comprehensibly, although there is a lack of cohesive elements. (mixed) /CEFR

Extract 30: I lost his meaning early on and his partner requests that he explain his summary as it's hard to understand. (negative) /Sw

In the CEFR descriptors, *coherence* is described using terms and expressions like *organisational patterns*, *connectors*, *cohesive devices*, *smoothly flowing, well-structured speech* and linking of elements into a *connected, linear sequence of points*. In the holistic descriptors used by the Swedish raters, terms like *structured* and *coherent* are used. In addition, it is also stated that candidate speech should be *adapted to purpose, situation, recipient and genre*. Finally, the Swedish assessment factors stress that the content of candidates' speech should be assessed in relation to how elaborate it is. The raters seemed to incorporate many of these features in their comments.

Fluency

General references to the *fluency* of test-takers' speech were most common (Figure 16). However, there were also some comments on specific aspects, namely *hesitation and pauses*, and *speed of delivery*. The majority of general comments were Positive, whereas the comments on *hesitation and pauses* were predominantly Negative, with only a few Positive comments. The category *speed of delivery* was insignificant with only four comments.

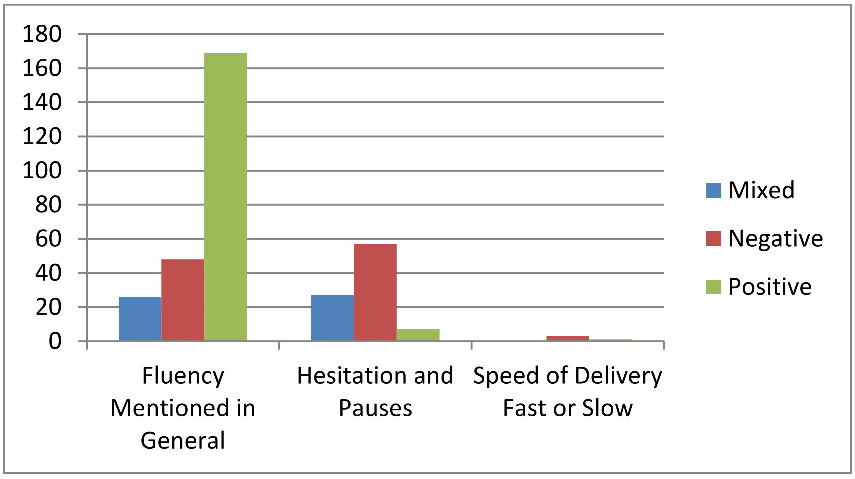


Figure 16. Evaluative responses per subcategory for fluency

Raters made frequent references to the *fluency* of candidates' speech. A majority were non-specific, relating to overall evaluations of *fluency* (Extracts 31-32).

Extract 31: She manages to put her message across all along, though she's clearly finding it hard to show consistent fluency. (negative) /CEFR

Extract 32: The speaker produces fluent, natural speech (positive) /CEFR

The second subcategory is more specific and refers to *hesitations and pauses* in speech (Extract 33). As noticed above, this subcategory had more Negative than Positive comments. There were also comments referring to the impact of *hesitations and pauses* on *intelligibility* (Extract 34).

Extract 33: The student makes frequent pauses and this, in combination with pronunciation, leads to loss of fluency (Tr.) (negative) /Sw

Extract 34: She speaks with several pauses and hesitation, which impairs the understanding. (negative) /Sw

Furthermore, raters made frequent inferences about the reasons for *hesitation and pauses*. At times the cause of hesitation/pauses was attributed to linguistic limitations, for example searching for the right words or grammatical structures (Extract 35), and at other times it was attributed to more pragmatic reasons, such as candidates' thinking about or planning the content of their response (Extract 36), or candidates' inability to express views or elaborate on a topic (Extract 37).

Extract 35: There are quite a lot of pauses – also longer ones – and hesitation when the speaker is looking for correct words and expressions. This is why the speech is not very fluent or coherent. (negative) /CEFR

Extract 36: /The range of vocabulary and structures is good enough/ though she sometimes pauses for planning or rephrasing. She gets lost ? but succeeds in continuing the conversation. (mixed) /CEFR

Extract 37: It is sometimes difficult for the speaker to come up with things to say → pauses. (negative) /CEFR

Fluency in the CEFR descriptors is conceptualised in terms such as *fluently and spontaneously*, *fairly even tempo*, *hesitant as he or she searches for expressions*, *pauses*. In the holistic descriptors used by the Swedish raters, *fluency* is used as a term, but is not explained or exemplified. In the assessment factors, the expression *fluency and ease* is used. Despite this somewhat vague description in their criteria, Swedish raters made the same kinds of comments as CEFR raters on both general *fluency* and *pauses and hesitations*.

Intelligibility

As can be seen in Figure 13, comments on *intelligibility* were mainly Negative, but there were some Positive and Mixed references as well. Quite naturally, perhaps, *intelligibility* was mainly taken into account in the rating decision when it caused problems. Usually the cause of *intelligibility* was clear. In some cases, raters referred to *accuracy*, more specifically to *pronunciation* (Extract 38), *grammatical accuracy* (Extract 39) and *vocabulary control* (Extract 40*)*. Lack of *intelligibility* was also related to *coherence* (Extract 41) and *fluency* (Extract 42). Sometimes, however, it was not clear what the cause of the intelligibility problem was (Extract 43).

Extract 38: Difficult to understand at times – sounds tend to become muddled, inaccuracies (nothing (to?) fat lose), pronunciation errors (e.g. [jast] i st f [dzast]), struggles to form utterances. (negative) /CEFR

Extract 39: when he presents his card a few grammar problems make it a bit difficult to understand what he is trying to say. (negative) /Sw

Extract 40: Unidiomatic and sometimes difficult to understand: They just is on the way you will get boring on later in your life…, (negative) /Sw

Extract 41: The parts about personal brands and smart phones were very tricky to understand. The content is not coherent. (negative) /Sw

Extract 42: She speaks with several pauses and hesitation, which impairs the understanding. /Sw

> Extract 43: I lost his meaning early on and his partner requests that he explain his summary as it's hard to understand. (negative) /Sw

Finally, in Figure 17, the proportion of comments on intelligibility per candidate is shown. The hypothesis is that performances at the lower levels might have more comments on (lack of) *intelligibility*.

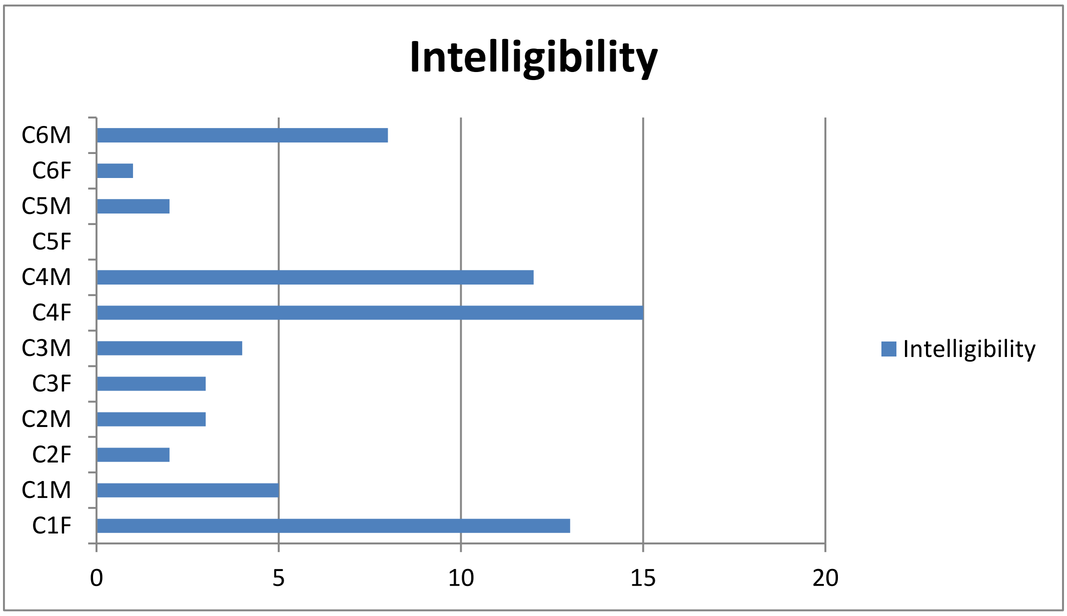


Figure 17. Proportion of comments per candidate coded as intelligibility

We can see from Figure 17 that C4F, C1F and C4M have the largest number of comments on *intelligibility*. C4F and C4M were also rated lowest by both the European and Swedish raters, suggesting that *intelligibility* plays an important part in the rating decision for lower levels. C1F, on the other hand, was ranked quite low by the Swedish raters, but somewhat higher by the CEFR raters (See Figure 11).

When looking at the rater comments for C1F, raters often commented on her lack of *accuracy* (especially pronunciation problems), but they were positive towards her *fluency*, *coherence* and *interaction*. In other words, her problems with pronunciation seem to have rendered many comments on (lack of) *intelligibility*, but her other skills seem to have compensated for this in her final grade. There was one more student with low grades, C3F, whom both the Swedish and CEFR raters had ranked as the third lowest candidate. She does not seem to confirm the hypothesis that *intelligibility* plays an important part at the lower levels. One more candidate is prominent in Figure 17, namely C6M, who has many comments on *intelligibility*. Both CEFR and Swedish raters ranked him in the middle section. However, he was one of the candidates for whom there was a large range among the Swedish raters, indicating that raters did not agree on how to rate this performance.

To summarise, *intelligibility* was mostly taken into consideration by raters when it caused problems. However, when looking at the distribution of comments on *intelligibility* across all candidates, there is no clear answer as to whether *intelligibility* is most salient at lower proficiency levels. In this material, candidates with the lowest average ratings, but also some candidates with average ratings, received most comments on *intelligibility*.

Intelligibility was not explicitly mentioned in the rating criteria for either the Swedish or the CEFR raters, however clarity of expression was. Clarity of expression is a general concept that can be related to different aspects of communicative competence, such as *pronunciation* and *coherence*, which is what the raters in the current study seem to have done.

Interaction

Figure 18 shows that the largest subcategory for *interaction* was *cooperating*, which is essentially about how speech is co-constructed by the participants. The other subcategories were quite small: *turntaking*, *manages or controls discussion*, *dominates discussion* and *has a passive role in discussion*. Furthermore, comments on *cooperating* were mainly Positive. References to *turntaking* had a majority of Positive comments but there were also some examples of Negative and Mixed comments. Finally, the two subcategories *dominates discussion* and *has a passive role in discussion* are both Negative in nature. Hence, the results showed mainly Negative evaluative comments. In comparison, *manages or controls discussion* was mainly used in a positive context.

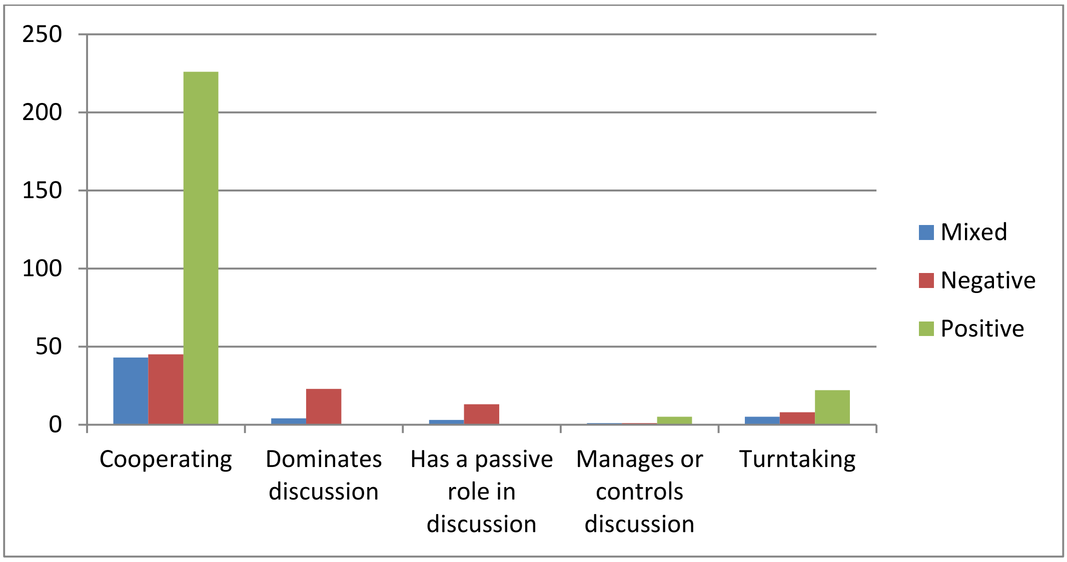


Figure 18. Evaluative responses per subcategory for interaction

Comments referring to the category *cooperating* were predominantly about how one of the candidates, or both of them, helped to advance the conversation. In the CEFR descriptors for *cooperating* (Council of Europe, 2001, p. 86), it is stated in the B2 descriptors that candidates should be able to "give feedback on and follow up statements and inferences and so help the development of the discussion". In addition, candidates should "help the discussion along on familiar ground, confirming comprehension, inviting others in etc." The raters seemed to have noticed these features, as will be illustrated in the examples below. First of all, candidates' ability to help the conversation along was achieved, for example, by asking questions or agreeing, (Extracts 44-45), as well as asking for or giving clarification (Extract 46). There were also comments on active listening skills (Extract 47).

Extract 44: He makes consistently very good contributions to the discussion, asking questions and introducing new topics. (positive) /CEFR

Extract 45: She does not contribute much to the conversation. She keeps asking "What do you think?" as she struggles to find something to say. (negative) /CEFR

Extract 46: and was prepared to engage with her partner and explain points he may not have understood. (positive) /Sw

Extract 47: She shows she's been listening to male partner and makes good contributions to the conversation. (positive) /CEFR

The three subcategories, *dominates*, *manages/controls* or *has a passive role* are examples of *interlocutor effects*; i.e. how the pairing of candidates affected test-taker performance. There were two students who were categorised as dominant by some raters: C5F and C2F. As regards C5F, comments clearly pointed to her dominance in the discussion, but raters perceived this in slightly different ways: either as purely negative for *interaction* (Extract 48), or as a feature that might be somewhat positive (Extract 49). There were also discussions of how her dominant behaviour might affect the grade (Extracts 50-51). The majority of the comments about this specific candidate's dominant behaviour were Negative, and it may therefore be assumed that this would affect the score in a negative way. However, when looking at the results of the ranking, this candidate was ranked highest among the Swedish raters and second highest among the CEFR raters, indicating that her interactional skills did not affect the rating decision in a negative way. There were also some raters who perceived this trait as positive since the candidate controls rather than dominates the discussion. In other words, this is an example of a comment coded as *manages conversation* (Extract 52).

Extract 48: She dominates conversation totally, which is not great for the purpose of interaction but her partner is slow and she jumps in not realising, perhaps, that he needs more time to think and find the right words than her. (negative) /CEFR

Extract 49: She is helpful, a bit bossy though, dominates in interaction, explains on behalf of her counterpart, overpowers rather than collaborates with him to achieve a conversation. (mixed) /CEFR

Extract 50: However, the student has a tendency to take over the conversation and does not let her partner join the conversation. She does not give her partner time to think when he, for example, cannot find the right words, which makes him feel stressed, causing her to take over even more – this is something that lowers the grade somewhat, since the conversation turns into a monologue rather than a dialogue. (Tr.) (negative) /Sw

Extract 51: NB. This should be a discussion! The student takes over quite a bit! Touches the grade A but behaves somewhat rudely to her partner. Let him speak! Work on turn-taking! (negative) /Sw

Extract 52: Actually, she controls discussion, asks the questions, asks him for clarification of what he says. Makes and helps both conversation and test flow.--> B2+-C1 (positive) /CEFR

Next, comments on the male candidate in the same conversation are shown, before returning to C2F. Many raters saw C5M, quite naturally considering the comments above, as a *passive speaker*. Once again, raters commented on how this passive behaviour might affect the grade (Extracts 53-54). Some raters also seemed to infer that the boy's personality could be a reason for his passive behaviour (Extract 55).

> Extract 53: Let's his partner take command too often and is not as involved in the discussions as he maybe could be, which reduces the grade as it's harder to get a full picture. (negative) /Sw

> Extract 54: During the rest – he is repeatedly interrupted by the female student, who speaks too much. It is hard to hear his full range, since he does not "fight" her verbally, he lets her take over. (negative) /Sw

> Extract 55: He is the silent partner of the bubbly personality ;) He does not seem to mind that his partner does almost all the talking. When he really wants to say something, he manages to do it, but this doesn't happen often. (mixed) /CEFR

As for C2F, raters seemed to be divided in their opinion about her interactional skills (Extracts 56-57). In an example from another conversation with two other candidates, comments also showed that a more proficient partner can be beneficial for the other candidate in the pair (Extract 58). More comments on the pairing of candidates are given under the section *rater reflection*.

> Extract 56: She interacts with ease and skill with natural turntaking, referencing, … She is engaged in keeping the conversation going on. (positive) /CEFR

> Extract 57: The pattern seems to be either she starts a subject – or she lets the male student start it – and the she "kills it off', by a very smart comment which is hard to counter for him. She confirms his comments "true, true" and BAM, she takes over again. /…/ But, her weakness is "interaction" – she's great at "production". (negative) /Sw

> Extract 58: If it hadn't been for her, he would have had a difficult time managing this task, but she asked him good questions, which made him think. (Tr.) /Sw

The final subcategory for *interaction* is *turntaking*. Comments referred to turntaking rules in general (Extract 59), and more specifically, candidates' ability to initiate and maintain discourse (Extract 60). When candidates did not initiate discourse, raters often commented on the (negative) impact this had on the interaction (Extract 61).

Extract 59: She takes her turn when appropriate (positive) /CEFR

Extract 60: He can initiate and maintain a simple conversation, (positive) /CEFR

Extract 61: He also does little to keep the discussion going but mostly just waits for his partner to respond to his comments. (negative) /Sw

Interaction is conceptualized in the descriptors used by the CEFR raters in terms of *getting and keeping the floor*; *initiate, maintain and close discourse*; *can help discussion along*; *can repeat back to confirm mutual understanding*. In the Swedish holistic descriptors, interaction is used as a term, but not explained. In the assessment factors, however, *communicative strategies* are mentioned and exemplified (referring both to ability to develop and advance the conversation, as well as production strategies). Both the Swedish and CEFR raters seemed to employ many aspects of candidates' interactional strategies in their rating decision.

Other

As can be seen in Figure 13, there was a majority of Positive comments for the category *Other*. Many comments in this category were about degree of confidence (Extracts 62-63), degree of relaxation (Extract 64), or use of "safe" language (Extract 65)

Extract 62: She is an unafraid speaker who takes risks while interacting with the male speaker, it works. (positive) /Sw

Extract 63: She is really in a hurry and repeats herself a lot, which makes her seem insecure. (Tr.) (negative) /Sw

Extract 64: Seems to be enjoying the conversation; (positive) /Sw

Extract 65: He uses however a safe language. (negative) /Sw

Production strategies

As can be seen in Figure 19, *monitoring and repair* was the largest group within the main category *production strategies* and it refers to candidates' ability to backtrack and correct slips and errors, or reformulate what he/she wants to say. *Compensating* was slightly smaller. This category refers to candidate's ability to use "circumlocution and paraphrase to cover gaps in vocabulary and structure"

(Council of Europe, 2001, p. 64). *Monitoring and repair* had a majority of Positive references, whereas *compensating* was more Mixed.

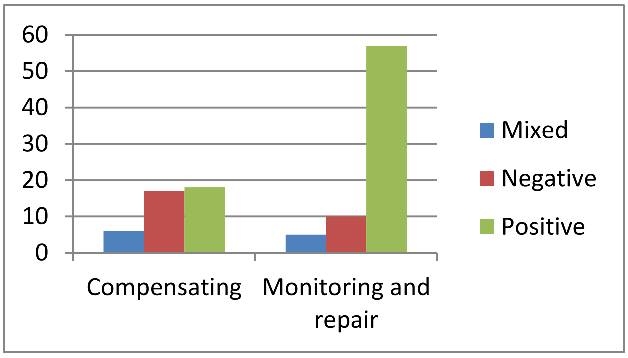


Figure 19. Evaluative responses per subcategory for production strategies

In many comments, the category *monitoring and repair* was related to linguistic awareness and control on the candidate's part (Extract 66). In other words, *monitoring and repair* was seen as a communication strategy. There were also comments on candidates' ability to backtrack and correct mistakes (Extract 67).

Extract 66: Corrects himself often, showing an awareness of the mistakes he is making. (positive) /Sw

Extract 67: Corrects herself when she, on some rare occasion, makes a grammatical error. If she starts a sentence incorrectly, she starts over and makes sure that she produces correct language and content. (Tr.) (positive) /Sw

Comments from the category *compensating* referred to candidates' ability to paraphrase content and use circumlocution (Extracts 68-69)

Extract 68: He tries work out any problems that may arise in the conversation, he struggles with explaining how some students might feel when they are not receiving top grades in school and he finally manages to work it out in the end. (positive) /Sw

Extract 69: and he uses strategies when he can't find the right words, for example, he explains what he means. (Tr.) (positive) /Sw

In some cases, lack of *compensating strategies* was manifested in seeking help from the partner in the conversation (Extract 70), which was coded as Negative since the candidate is not using his/her own *production strategies*. However, from the rater's perspective, the partner who offers help shows skills that can be rewarded in the rating.

Extract 70: She even needs help from her partner in a couple of cases. (negative) /CEFR

Production strategies are mentioned both in the CEFR descriptors (e.g. can correct most of his/her mistakes) and in the assessment factors that the Swedish raters use (communicative strategies to solve linguistic problems). It is clear that both rater groups employ this criterion in their rating decision.

Range

Figure 20 indicates that *vocabulary range* was the largest category, *general linguistic range* the second largest, and *ability to express viewpoints* the smallest. All three categories had a majority of Positive comments.

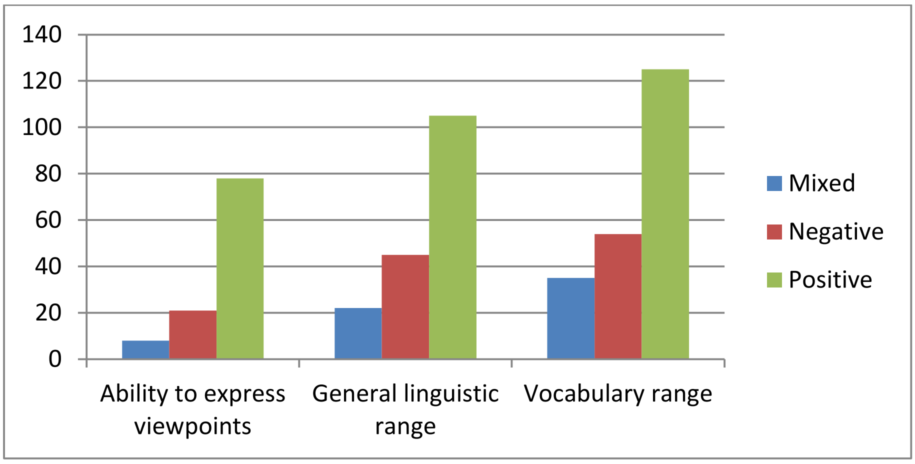


Figure 20. Evaluative responses per subcategory for range

Comments on vocabulary *range* were either general or specific. General references referred to *variation, richness* and *sophistication* (basic or advanced vocabulary) of the lexical repertoire, including use of *idiomatic expressions* (Extracts 71-71). Other comments were more specific, drawing attention to specific lexical choices (Extracts 73-74)

Extract 71: Examples of idiomacy and variation to vocabulary. Not advanced, but extensive and varied. (mixed) /Sw

Extract 72: The language is simple but varied and contains a few idiomatic phrases/expressions. (mixed) /Sw

Extract 73: The student has a relatively good vocabulary and uses some really good expressions (comfort zone, interpret, appreciate the little things, life experience, hard to settle down). (Tr.) (positive) /Sw

Extract 74: but VERY repetitive markers (marker! = “exactly”). Further examples of her NOT being very varied: ”it depends on, kind of” (negative) /Sw

In comparison, comments on *general linguistic range* referred to linguistic resources in general, rather than distinguishing between grammar and vocabulary. Raters used terms such as *language*, *linguistic repertoire*, *linguistic usage*, and *sentence structure*. Comments referred to *sophistication* and *richness* of language (Extracts 75-76), and *control/command* of language (Extracts 77-78). Raters also commented on candidates’ ability to express him/herself with ease and *fluency* (Extract 79)

Extract 75: and her language is nuanced in many occasions. (positive) /Sw

Extract 76: She also seemed to be able to use a range of structures (positive) /CEFR

Extract 77: She has a very good command of language structures and lexical items. (positive) /CEFR

Extract 78: but does not seem to be able to tackle issues and topics which are predictable, using simple language, (negative) /CEFR

Extract 79: In part 2 produces longer sentences with ease. /CEFR

Finally, examples from the category *ability to express viewpoints* are given (Extracts 80-81). In many cases, comments within this category were related to interactional effectiveness (Extracts 82-83).

Extract 80: and gives her viewpoint. She exemplifies and gives her thoughts throughout the test. (positive) /Sw

Extract 81: She finds some problems to describe her point of view, but she ends up finding the way to do it without help. (mixed) /CEFR

Extract 82: He does take part in the discussion, however, and gives his opinion on what his partner talked about (Tr.) (positive) /Sw

Extract 83: and she explains what she means when agreeing or disagreeing with the male speaker. (positive) /Sw

Range is conceptualized in the CEFR descriptors as the ability to "express him/herself with sufficient vocabulary and language on general topics and sufficient range of language to express viewpoints". In the holistic descriptors and analytic assessment factors that the Swedish raters used, variation and range of vocabulary, phraseology and idiomatic expressions, as well as richness and elaboration of content, are mentioned. The raters seem to have expanded on these criteria in their comments to a very large extent.

Sociolinguistic appropriateness

As can be seen in Figure 13, the majority of comments on *sociolinguistic appropriateness* were Positive. It should be kept in mind, however, that this category was the smallest in relation to the total number of comments. It referred to candidate's ability to express him/herself in a formal or informal register appropriate to the situation (Extracts 84-86).

> Extract 84: Uses the word "crap" which is not appropriate in this context – he apologises however, which shows that he is aware of this. (Tr.) (mixed) /Sw
>
> Extract 85: Says "stuff" a bit too often. (Perhaps a bit influenced by spoken, informal English and jargon). (negative) /Sw
>
> Extract 86: No bad language or too colloquial terms are used. (positive) /Sw

In the CEFR descriptors, it is stated that test-takers should be able to express him/herself clearly in an *appropriate style*. However, appropriateness is only mentioned at the B2 level, appearing to indicate that this is a skill acquired at the higher levels. In the holistic descriptors used by the Swedish raters there is no explicit reference to sociolinguistic appropriateness. As mentioned in the section on *coherence*, adaptation to purpose, situation, recipient and genre is stated in the Swedish rating descriptors. However, in the current study these instances are coded as *flexibility to circumstances*, which is a subcategory of *coherence*. In summary, then, the Swedish and CEFR raters seemed to refer to sociolinguistic appropriateness only to a limited extent.

Task realisation

Figure 21 shows that the category *task realisation* consisted mainly of comments on how candidates summarised the short text they had read in advance. The majority of comments on *summary of text* were Negative, but Positive comments were frequent, too. For the *overall* comments, Positive evaluations were dominant. In contrast, *completing and understanding task requirements* and *length of response* were mainly Negative.

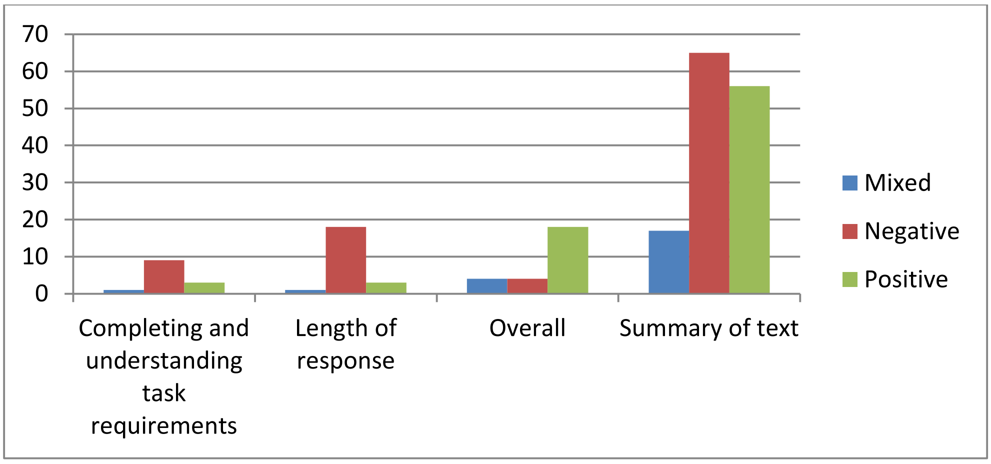


Figure 21. Evaluative responses per subcategory for task realisation

The category *summary of text* was quite straightforward with comments on how well the candidates summarised the text. Often, the comments are about whether the student uses his/her own words or reads straight from/uses many words from the text (Extracts 87-88).

> Extract 87: Summarizes the card well, in her own words (positive) /Sw

> Extract 88: She stuck to the text a bit too much when summarizing her text. (negative) /CEFR

Some comments on *summary of text* also point to consequences for the paired interaction if one of the candidates cannot summarise his/her text in a satisfactory way. In these cases, the discussion in the pair, which is meant to be about the text, may suffer (Extracts 89-91). There were also comments on the *overall* skills or production of the candidates (Extracts 92-93).

> Extract 89: She talks very briefly about her card, which is why the discussion is also brief (Tr). (negative) /Sw

Extract 90: Short about card (skips the brand part) – could be better – has to develop more since his partner doesn't understand what he means (negative) /Sw

Extract 91: Eleven gör en alltför kort sammanfattning av det som star på hans kort. För den som lyssnar blir informationen inte tillräcklig helt enkelt. (negative) /Sw

Extract 92: The production was overall superb (positive) /CEFR

Extract 93: Student lacks basic skills (negative) /Sw

Raters also commented on the length of response; whether there was brief or extended discourse by candidates (Extracts 94-95).

Extract 94: She has not got much to say or if she has something to say, then her comments are short. (negative) /CEFR

Extract 95: Long, sustained presentation (positive) /CEFR

Finally, comments in the category *completing and understanding the task requirements* were about whether candidates had fully grasped the instructions (Extract 96-97)

Extract 96: Does not fully get the statements in the instructions. (negative) /Sw

Extract 97: He follows the instructions of the task and it seems like he has a clear picture of what he wants to say (even if there were pauses in the beginning). (Tr.) (positive) /Sw

Comments coded as rater reflection

Raters made many inferences about test-takers based on their performance in the test, but also reflections on the rating decision, i.e. the grades. *Rater reflections* constituted about 5% of the total number of coded comments (excluding evaluative response). This category was divided into three main groups, referring to: (1) *matching of candidates*, (2) *rating decision*, and (3) *rater reflection in general.* In Figure 22, the distribution of comments per subcategory for rater reflection is shown. It is indicated that *rater reflections* regarding *rating decision* were most common, followed by *rater reflection in general*; lastly, a minor proportion of comments pertained to *matching of candidates*.

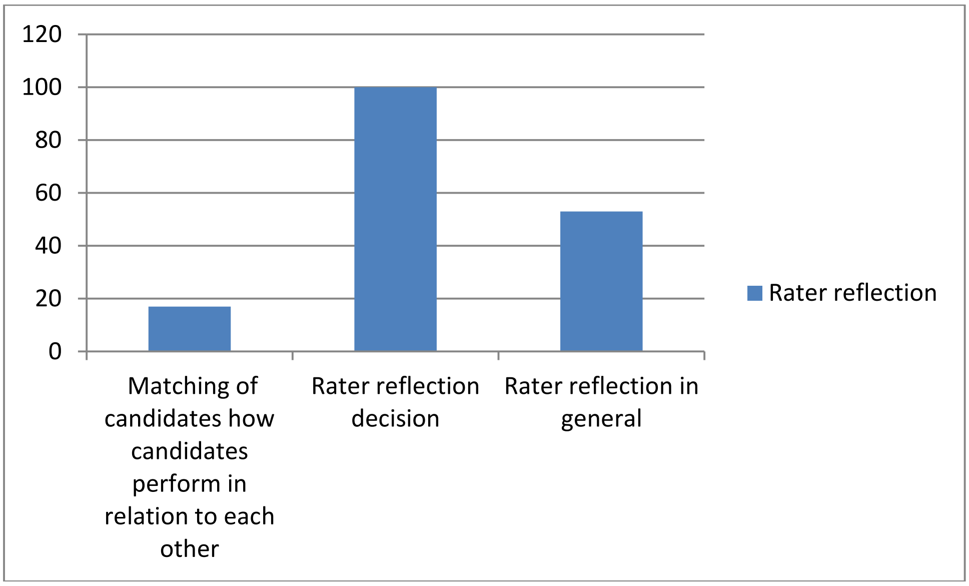


Figure 22. Comments per subcategory for rater reflection

With regard to matching of candidates, negative and positive consequences as a result of the candidates' various proficiency levels were mentioned (Extracts 98-99). In a few cases, raters commented that they thought the examiner should intervene to make the discussion more equal (Extract 100). Raters also speculated quite openly about different aspects of test-taker performance, and seemed to be aware of the fact that they were making *inferences* (Extract 101).

Extract 98: I feel he could have performed better with a more collaborative partner with better contributions. /CEFR

Extract 99: I think she helps her partner achieve a higher grade than he has achieved before because she adapts her language and asks good questions. (Tr.) /Sw

Extract 100: In part two, she follows conversation well. If anything, as I have said before, she takes over in a way which does not allow her partner to show his full potential. Maybe the examiners should have intervened (?). /CEFR

Extract 101: Also it is sometimes difficult for him to join the conversation, since he is interrupted several times Maybe he could have participated more with another interlocutor but we don't know that. (Tr.) /Sw

As regards *rating decision*, there were comments on both specific features that affected the rating decision (Extracts 102-103), as well as justifications of marks, (Extract 104). Further, references were made to descriptors in the rating scale (Extracts 105-106).

Extract 102: It is unfortunate that the non-English accent is so strong. This student does a good job at completing the task! /Sw

Extract 103: His language is not the best, and neither is his pronunciation. But he deserves a higher grade considering the content. (Tr.) /Sw

Extract 104: The reason why she gets an E grade and he doesn't is because she has better ideas and follows the instructions better. (Tr.) /Sw

Extract 105: Her fluency may be B1+, but the rest of the elements are B1 /CEFR

Extract 106: Interaction is high, range not so much, but fluent speaking and ok grammar takes this one to B2. /CEFR

It was mostly the CEFR raters who made reference to the descriptors/rating criteria. This is quite natural, since they had scaled descriptors for all the five analytic criteria (*accuracy*, *coherence*, *fluency*, *interaction* and *range*). The Swedish raters, in comparison, used broad, holistic descriptors (i.e. the national performance standards) for the different levels of proficiency, such as:

> students can express themselves clearly with fluency, and with some adaptation to purpose, recipient and situation. In addition, students can choose and use essentially functional strategies which to some extent solve problems and improve their interaction.

Let us now move on to general reflections, which included examples of inferences of different kinds. For example, raters speculated about reason for lack of *topic development* (Extracts 107-108), or about certain general *behaviours* (Extracts 109-110), as well as *personality* (Extracts 111-112). There were even inferences about *body language* (Extract 113).

Extract 107: We cannot be sure if it's for lack of ideas or lack of language, but I'm inclined to think it's the latter as they're discussing on a subject which should be quite relevant to their generation and interests. Still, it's only my perception… /CEFR

Extract 108: She repeats back of what he has said to confirm mutual understanding or maybe she has not got much to say. /CEFR

Extract 109: In the beginning I was under the impression that she was listening actively and was interested in what he said, but I noticed after a while that she repeated everything he said, and that she didn't have many thoughts of her own on the topics discussed. To some extent, she interrupts the conversations with her "yes", "I think so" and "yeah". (Tr.) /Sw

Extract 110: She uses laughing to cover her lack of vocabulary. /CEFR

Extract 111: His (apparent) personal shyness probably does not help him to take the initiative in the conversation as often as he should. /CEFR

Extract 112: He really is more of a listener than a leader or even a real partner in a conversation. Is this his personality? Maybe. /CEFR

Extract 113: Based on hearing the conversation, I can also read good use of non-verbal gestures. /CEFR

Finally, there were also general comments on the examiner's role or involvement in the test (Extracts 114-115).

Extract 114: It's the examiner that makes them move from one part into the next. Is Examiner's intervention necessary? /CEFR

Extract 115: Dealt well with an examiner that was a little too involved. /Sw

Comments coded as inter- or intra-candidate comparison

About 11% of the total number of coded comments (excluding evaluative response) were categorised as *inter-* or *intra-candidate comparisons*. The *inter-candidate comparisons* consisted of comments on (1) comparisons with other pairs, (2) similarities between the two candidates, (3) differences between the two candidates, (4) candidates' proficiency levels, (5) the interaction between the two candidates. Finally, there was a fifth subcategory referring to *intra-candidate comparisons*, comparing an aspect of a candidate's performance over time in the conversation. What was special about these comments was that they referred to the pair, and not to the individual test-taker. In other words, the question of separate scores when the performance is co-constructed, is focused upon here.

To get an overview of the way different raters used these comments, and to illustrate rater orientations for this category, Figure 23 is provided.

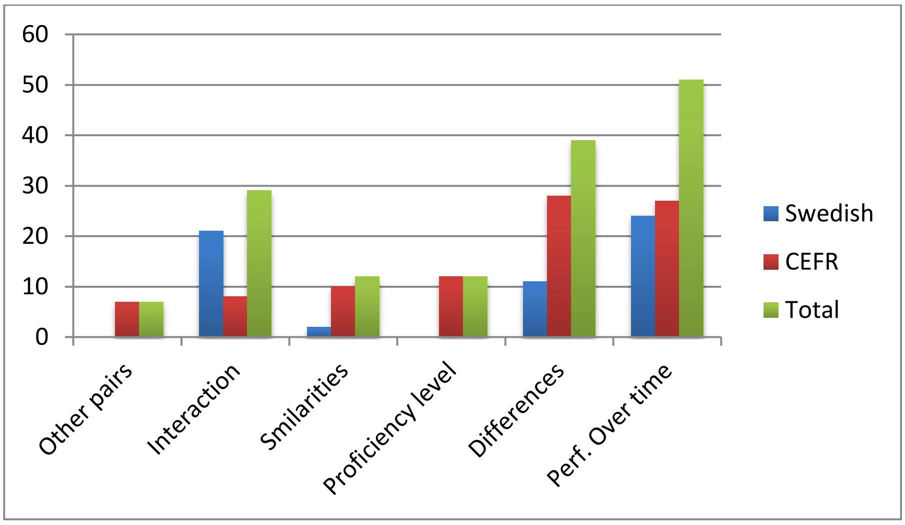


Figure 23. Comments coded as inter- or intra-candidate comparisons

Figure 23 indicates that the CEFR raters seemed to make proportionally more *inter-* and *intra-candidate comparisons*, compared to the Swedish raters (interaction being the exception). It was also shown in the coded comments that whereas all CEFR raters included some sort of comparison in their comments, not all of the Swedish raters did, confirming the picture that the CEFR raters seemed to make more *inter-candidate comparisons* in general. In addition, both the Swedish and the CEFR raters made many *intra-candidate comparisons* (intra-candidate comparisons (referred to as ”Performance over time” in Figure 23). Below, examples of the categories are provided.

As mentioned, raters noted *similarities* between candidates (Extracts 116-117). There were also examples of rater comments referring to *differences* between candidates (Extracts 118-119).

Extract 116: Both of them jump from one topic to the other and make a few comments but there is not a real discussion. (negative) /CEFR

Extract 117: The speakers help each other well here, they give and take, ask for clarifications, examples (positive) /Sw

Extract 118: not quite as comprehensive as the male speaker's, also simpler. (negative) /CEFR

Extract 119: The speaker pauses and hesitates more than the female speaker, also speaks more briefly and in a simpler way (negative) /CEFR

Furthermore, there were comparisons of candidates' *proficiency levels* (Extracts 120-121). Also, candidates were compared in terms of their *interaction* (Extracts 122-124). In some of these comments there was a strong *individual* focus (Extract 122), whereas other comments referred to the interaction in the *pair* (Extract 123), or sometimes both (Extract 124).

> Extract 120: She seems to me to be at about the same levels as her interlocutor. /CEFR
>
> Extract 121: This pair seems quite well matched in terms of competence. /CEFR
>
> Extract 122: She asks good and relevant questions to her interlocutor. This moves the conversation forward and contributes to interesting discussions. There is good interaction in the pair, and she contributes to this to a great extent. (Tr.) (positive) /Sw
>
> Extract 123: No real interaction in terms of posing questions to one another, but agreeing/disagreeing mutually on the text. (mixed) /CEFR
>
> Extract 124: Very good interaction most of the time, which creates a conversation between the two. It but comes to a halt at some occasions when they become silent. But she takes initiative to move on in the conversation. (positive) /Sw

There were a few instances of comments that compared the pair with other candidates/pairs in the test (Extract 125). Finally, raters frequently commented on candidates' performance over time in the conversation, typically noting candidates' development, or lack thereof, through the conversation (Extracts 126-127). As can be seen, many candidates seem to function better in part two of the test, judging from rater comments. In part two, focus is on oral interaction, as opposed to the first part, which also involves oral production (summary of short text).

> Extract 125: Fluent, but not that much compared to participants in other conversations. The conversation does not flow smoothly. /CEFR
>
> Extract 126: The further we come, the more relaxed he seems; high level of fluency and ease. (positive) /Sw
>
> Extract 127: Her fluency is sometimes disturbed because she can't find the words. However, this is better in part two and when she is thinking freely. (Tr.) (mixed) /Sw

To briefly summarise this section on analytic categories, raters took a wide range of performance features into account in their judgements. Examples of comments from the different categories have been given in this section. Most of the comments raters made were related to the criteria and descriptors in their respective rating scales. However, there were exceptions. For example pronunciation was not mentioned specifically in the CEFR descriptors, and only marginally in the Swedish assessment factors. Also, a small proportion of comments (about 12%) did not to pertain to specific criteria, thus being "self-generated". Further, there were comments that were categorised as *rater reflection* (5%) and *inter- and intra-candidate comparisons* (11%).

Relationship between rater comments and scores

Distribution of comments per candidate

This section relates to research question number three: "What is the possible relationship between scores and raters' justifications of these scores? Distributions of comments per candidate can be seen in Table 10. This table was inspired by a similar one in Brown (2007, p. 131). For each candidate, the total number of coded comments in each category across all raters was calculated as a percentage of the total number of comments for this candidate. This was then compared to the mean for this category. For each category, percentages that are more than one standard deviation higher than the mean are shown in bold type. For these particular cases, this specific feature seems to be more salient to raters than the average for this category.

Table 10. Comments by category for each candidate (%)

	Acc*	Coh	Flu	Intell	Inter	Other	Strat	Range	Soli	Task
C1F	**27**	15	12	5	15	2	4	15	0	4
C1M	21	**18**	13	2	15	0	3	17	**4**	6
C2F	23	14	14	1	18	1	1	**23**	1	4
C2M	23	14	14	2	14	2	1	19	1	**10**
C3F	22	14	10	1	17	1	7	17	0	8
C3M	19	**18**	10	2	15	1	**11**	16	1	6
C4F	22	12	11	**7**	12	0	**10**	19	0	7
C4M	25	12	11	**6**	11	1	2	22	0	**11**
C5F	20	13	13	0	**24**	0	2	20	0	7
C5M	23	11	12	1	16	1	1	**24**	1	**11**
C6F	17	15	**22**	1	12	1	5	18	0	9
C6M	22	16	**17**	4	12	**3**	1	17	0	9
Mean %	22	15	13	3	15	1	4	19	1	8
S.D.	3	2	3	2	4	1	4	3	1	2

* Categories in the following order: Accuracy, Coherence, Fluency, Intelligibility, Interaction Other, Production strategies, Range, Sociolinguistic appropriateness, Task realisation

Table 10 was produced to see if there were any obvious differences in distribution of comments, which could indicate that the focus of comments was different for different candidates. However, Table 10 shows that the distribution of comments for each category was, in general, very similar among all candidates, with a few exceptions. Moreover, it can be seen that, for each candidate, one or two categories were one standard deviation above the mean and thus seemed more salient. We can see, for example, that C1F received proportionally more comments on *accuracy*, indicating that this could be a more salient feature for her. The reason for this was examined in the section on *intelligibility*. Furthermore, C1M had proportionally more comments on *coherence* and *sociolinguistic appropriateness*. The reason why this candidate stands out when it comes to *sociolinguistic appropriateness* is that he was the candidate who happened to say "crap" and then apologised for his bad language, which was commented on by many raters. With regard to *coherence*, he has a large majority of Positive comments (81%), indicating that *coherence* is a strong feature for this candidate, noticed by many raters. It is beyond the scope of this investigation to explore each candidate for highly salient features, but Table 10 gives at least some indications of individual rater focus for different candidates. It also shows that individual candidates seem to have at least some feature each that is more salient to raters than others, and these features seem to be highly individual.

To explore the issue of the relationship between comments and scores somewhat further, a comparison between the two students with the highest scores (C2F and C5F) and those with the lowest scores (C4F and C4M) was made, to see if there were any clear differences pertaining to high and low proficiency levels. As can be seen in Table 10, the distribution of comments was fairly similar between the four candidates, despite the fact that they had been ranked lowest and highest. There were a few exceptions, however. C5F had the highest proportion of comments on *interaction* (24%) across all candidates, more than one standard deviation above the mean. C2F also had a large proportion of comments (18%) on *interaction* compared to the other candidates. C4F and C4M, on the other hand, had a lower proportion of comments on *interaction* (12% and 11% respectively), possibly indicating that *interaction* is a more salient feature at higher proficiency levels.

As can also be seen in Table 10, C4F had a large proportion of comments on *production strategies* (10%). This was not the case for C4M, however. There was one other candidate who had a large proportion of comments (more than one standard deviation above the mean) on *production strategies*; C3M. This candidate was ranked as number eight among the Swedish raters and number seven among the CEFR raters. Thus, it seems that there is no obvious link between proficiency level and use of *production strategies*.

Finally, as was explored in the section on *intelligibility*, C4F and C4M, who had the lowest marks, also had proportionally more comments on *intelligibility*, suggesting that this feature might be more salient at the lower proficiency levels. C1F, who was ranked quite low by the Swedish but somewhat higher by the CEFR raters, also had a large proportion of comments on *intelligibility* (5%). However, as stated before, the other candidate with low ranking by both CEFR and Swedish raters, C3F, did not have a large proportion of comments on *intelligibility*.

As a final step of this analysis, evaluative comments per candidate were checked, the results of which can be found in Figure 24. The three candidates with the lowest scores and ranking, both by the Swedish and the CEFR raters, C3F, C4F, and C4M, had a majority of Negative comments. All other candidates had a majority of Positive (and Mixed) comments. In addition, the two candidates with the highest scores, C2F and C5F, had a large majority of Positive comments and rather few Mixed and Negative ones.

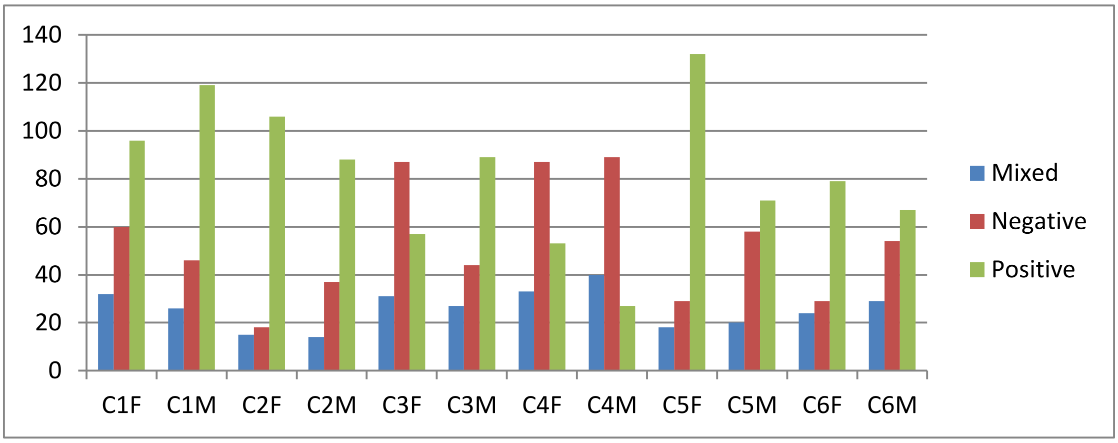


Figure 24. Evaluative comments per candidate

Examples of relationship between comments and scores

In the quantitative results, some interesting examples emerged that could be explored in the qualitative section. Relating to this section, some illustrative examples are given in Appendix 11. This appendix consists of four tables, which will be referred to in the text below. First of all, among the Swedish raters, C3M and C6M had the largest range, whereas C5F had the smallest. In other words, these performances are interesting to compare in two respects: (1) for performances with a large range, comments at the lower end of the scale can be compared with comments at the higher end; and (2) for performances with a small range, comments can be compared to see whether raters notice the same features in the performance or not, since they have awarded the same marks.

For C3M, a performance with a large range, a comparison of rater comments at the lower and higher end of the scale is provided in Table 1 in Appendix 11. To the left, the comments of two raters who awarded low scores are shown. Conversely, to the right two raters who awarded high scores to the same performance are shown. The comments are divided into features the raters claim to pay attention to. It is clear that the raters notice roughly the same features of the performance. As expected, however, the raters who have awarded higher scores see these features as positive whereas the raters who awarded low scores see them in a more negative light.

C6M has two extreme scores among the Swedish raters: one rater awarded this candidate a four (E+) and another one gave a ten (A). The rest of the raters awarded this performance a six, seven or an eight. In Appendix 11, Table 2, comparisons for this performance are given between the two raters with extreme scores. Here we can see that the two raters noticed roughly the same features. Sometimes they evaluated them in a similar way ("good examples" vs. "complex"; "interacts well" vs. "the speakers help each other well") and sometimes in a different way ("unclear" vs. "structured"). However, there are also differences. Whereas the rater with the high mark noticed broad and varied vocabulary, and did not make any remarks at all on *accuracy*, the rater who awarded a low score, noticed grammar and phrasal errors. This might suggest that, here, the raters actually award scores based on partly different performance features. From a research point of view, this could be followed up by analysing test-taker discourse to see how the features raters comment on are correlated with the actual performance.

Another interesting aspect is how raters comment on the same performance when they agree on the mark. Appendix 11, Table 3 shows two Swedish raters' comments on C5F, a performance on which all Swedish raters agreed that it was either a nine or a ten. It is shown that the raters noticed the same features to a very large extent. Moreover, the comments they made were very similar, indicating that the raters agreed on both the mark and the reasons for the mark.

Finally, a last comparison between scores and comments is made for C6F, whom the Swedish and the CEFR raters had ranked somewhat differently. The Swedish raters ranked this performance as number three, whereas the CEFR raters ranked it as number five. In Table 4 in Appendix 11, two Swedish raters' comments and two CEFR raters' comments are compared. It is once again clear that both the CEFR and the Swedish raters take the same performance features into account. The two Swedish raters commented more on *accuracy* than the CEFR raters. For this performance, it is worth noting that none of the raters commented on interactional skills to a very large extent. The raters seemed to be fairly much in agreement about the candidate's problems in *fluency*. However, there seemed to be slightly different opinions on whether *coherence* and *range* needed to be improved (CEFR raters) or were satisfactory (Swedish raters). It seems that these two features make up the main differences, which led to a slightly lower ranking for this performance by the CEFR raters compared to the Swedish raters.

In sum, the distribution of comments across candidates seemed to be fairly similar regardless of proficiency level. However, *interaction* was commented on to a greater extent for the two highest-scoring candidates, whereas *intelligibility* was more salient for the two lowest-scoring candidates. In addition, each candidate seemed to have one, or sometimes two features, which were proportionally more salient. Furthermore, examples of comments from raters who awarded a candidate a high grade were compared to comments by raters who awarded the same candidate a low grade. Results showed that raters noticed fairly similar features but there were some differences in how they evaluated them, and in some cases they actually seemed to base their decision on partly different performance features. In the example where raters had awarded the same score for the same performance, it was clear that raters noticed the same features to a fairly large extent and also evaluated them in the same way. Finally, rater comments on a performance that the Swedish and the CEFR raters had ranked differently were compared. It was found that two features were evaluated differently: the CEFR raters viewed them as improvement areas, whereas the Swedish raters found them satisfactory.

Chapter Six: Discussion

In this chapter, the main findings of the study are reviewed and further comments and interpretations are offered. This chapter follows the same structure as the results section, with the 'quantitative' results being discussed first, followed by the more 'qualitative' ones.

Rater variability and reliability

Swedish raters

In the present study, 17 Swedish raters, and 14 European CEFR raters, rated six paired speaking tests from the Swedish national test of English for upper secondary school. The raters used two different rating scales. The Swedish raters had a ten-point rating scale based on the Swedish performance standards, whereas the CEFR raters had a nine-point scale based on the common reference levels in the CEFR. The intention was not to compare the two rater groups, since they used different scales. Instead, two separate analyses were made to answer two of the four research questions. For the Swedish raters the main research question was: What can be noticed regarding variability of scores and consistency of rater behaviour? For the CEFR raters, the relevant research question was: At what levels in the CEFR do external raters judge the performances of the Swedish students to be?

Findings from the descriptive statistics for the Swedish raters showed signs of variability of ratings as well as differences in consistency. For example, the average scores for the Swedish raters varied between 5.6 and 8.0 on the ten-point scale. Further, rater profiles with differences in leniency and severity were identified. There were also oral performances that raters seemed to have more difficulty agreeing on than others. Considering the fact that the test used in the current study is an example of a so-called performance test, a certain degree of variability and inconsistency of rater behaviour was expected. As McNamara (1996) points out, performance assessment always involves interpretations by the raters and is thus subject to rater variability. Furthermore, there are several types of interactions involved in performance testing. First of all, the rater interprets the student's performance according to a rating scale and rating descriptors; secondly, there is the interaction between the two candidates in the test. Needless to say, these interactions make the rating process complex and there are many factors that can have an effect on the final outcome, i.e. the test scores. To summarise, however, the differences between the Swedish raters are not excessively large, indicating a reasonable degree of variability. In other words, rater effects seem to exist but are not striking.

As regards rater reliability, rank-order correlations, using Sperman's rho and Kendall's tau, were computed. The results showed that the median of correlations was .77 for Spearman's rho and .66 for Kendall's tau, pointing to reasonably satisfactory inter-rater reliability. Also, internal consistency was calculated using Cronbach's alpha. The result was .98 for the Swedish rater group (*n* = 17), indicating stable internal consistency.

There is a general claim in the literature that the OPI has high inter-rater reliability. One study often referred to is Adams (1978), whose findings on the FSI oral proficiency interview showed that inter-rater reliability between two raters was consistently .87, or higher. In other words, the inter-rater reliabilities in the present study seem somewhat lower. However, one major difference is that Adams based his study on a much larger sample (834 test performances). Further, the relationship between analytic factors and overall holistic scores was examined in Adams (1978). In the present study, only holistic scores were used (even though they are based on analytic descriptors for the CEFR raters and analytic, unscaled assessment factors for the Swedish raters, as well as holistic performance standards).

A third difference is that Adams (1978) and other previous studies with similar results investigate the OPI. However, it has been shown that inter-rater reliability is somewhat lower for group discussions and role plays (Shohamy et al., 1986), i.e. test formats with more than one test-taker. Therefore, it is also possible that the somewhat lower correlation coefficients in the present study are due to the fact that the paired speaking test format is more complex to rate than the OPI, thus generating more variability.

Finally, as pointed out in the research review on rater reliability, correlation coefficients do not take severity or leniency of raters into account. Therefore, it is important to investigate the descriptive statistics to get a broader view of the ratings. This was done in the present study, where the descriptive statistics were used as a complement to the correlation analyses.

To summarise, considering the fact that this study has a small sample size and is based on holistic scores, the overall results point to satisfactory inter-rater agreement. However, it should be noted that correlations are sensitive to the number of cases used, and so the inferences that can be made from a study with a small sample size, like the present, should be seen as tentative.

External CEFR raters

A secondary aim of the study was to make a small-scale empirical comparison between the Swedish performance standards for EFL and the common reference levels in the CEFR. Average ratings showed that the CEFR raters judged the performances of the Swedish test-takers to be between B1+ and C1 for all performances but two, which were clearly below B1+. These two performances were also rated as a Fail by some of the Swedish raters, which suggests that these two candidates' performances were considered borderline cases. The passing level of the test is intended to correspond to a low B2 (B2.1). Thus, the CEFR raters seem to be a little harsher around the cut-off point, i.e. B1+, than the Swedish raters.

In addition to examining the CEFR raters' scores in relation to the intended entrance level of the speaking test used in the present study, the rank ordering of performances was compared between the Swedish and CEFR raters. The results showed that the rank ordering was quite similar between the two groups. This is an interesting finding considering the fact that the raters come from different educational systems. What is more, the CEFR raters were not familiar with, or had any previous experience of, this specific speaking test. The Swedish raters, in comparison, rate this kind of test on a regular basis.

Rater orientations

Raters wrote summary comments regarding features of the performances that contributed to their judgement. These comments were segmented and coded using a coding scheme based on some of the illustrative scales for the different communicative competences (linguistic competence, pragmatic competence and sociolinguistic competence) and strategies described in the CEFR. Some additional categories were added to the coding scheme as well, based on features found in the rater comments. The research question relevant to this section is: What features of test-taker performance are salient to raters as they make their decisions?

Findings indicated that *accuracy* was the most salient feature, closely followed by *range*. Both *accuracy* and *range* are components of linguistic competence in the CEFR. In other words, test-takers' linguistic competence appeared to be highly salient to raters, with as many as 41% of the coded comments. Moreover, *coherence* and *fluency*, which are components of pragmatic competence in the CEFR, together accounted for 28% of the coded material. Candidates' pragmatic competence thus seems to be the second largest component that raters in this study heeded. *Interaction*, referred to as strategies in the CEFR, was the third largest category (15%), indicating that interactional skills were important in the rating decision.

Production strategies, also part of the strategies described in the CEFR, comprised 4% of the total number of coded comments. They thus seem to play a minor role in the rating decision, but are nevertheless a salient feature.

Surprisingly, *sociolinguistic appropriateness*, corresponding to the third component of communicative competence in the CEFR, turned out to be a small category with 1% of the coded comments, an issue in need of further exploration. *Sociolinguistic appropriateness* refers to students' ability to use language with the appropriate social meaning for the communicative situation at hand. However, since the test in the present study is a paired discussion between non-native speakers of the same age, sitting together in a test situation, opportunities for showing sociolinguistic awareness are somewhat limited. This may be the reason why there were very few clear examples of sociolinguistic appropriateness in the rater comments.

The last categories (*task realisation*, *other* and *intelligibility*) are features not explicitly mentioned in the rating criteria. It is therefore interesting to see that they represented about 12% of the rater comments. Partly, this has to do with the fact that many raters commented on how well the candidates summarised a short text, which is a task they need to fulfill in the first part of the test. Also, raters made some comments on overall performance, which is shown to be an important factor in both Hsieh (2011) and Annie Brown et al. (2005). The other non-criterion categories, *intelligibility* and *other*, together comprised a small part of the data (4%).

It is also interesting to make comparisons with previous research. The finding that *accuracy* was highly salient to raters can also be found in, for example, Brown (2007) and McNamara (1990). It is worth noting, however, that these studies do not investigate the paired speaking test format. In Brown (2007), the category *Sentence level syntax*, roughly corresponding to *grammatical* *accuracy* in the present study, was the most salient feature to raters (31%). A similar result can also be seen in McNamara (1990), in which a performance speaking test for health professionals was examined by investigating the relationship between an 'overall' speaking test score and analytic criteria. It was shown that grammar was the category that contributed most to the overall score.

Since *accuracy* in the present study does not only include *grammatical accuracy*, but also *phonological control* and *vocabulary control*, it is not possible to compare the results directly with the studies cited above, but they are clearly in line with the results of previous research. The reason why grammar is such a salient feature could be that it is quantifiable and systematically taught. In line with this reasoning, Wall, Clapham and Alderson (1994) point out that "grammar is less difficult to judge than the language skills" (p. 335). Iwashita, Brown, McNamara, and O'Hagan (2008) refer to several studies investigating the relationship between features of performance in determining overall speaking scores and conclude: "Taken as a whole, the studies cited above appear to show 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" (p. 27). In other words, the results of the present study, with 41% of the comments pertaining to candidates' *linguistic competence*, seem to confirm the general pattern of rater orientations observed in earlier research.

However, when looking at the two groups separately – Swedish raters and external CEFR raters – somewhat different rater orientations were noticeable. The CEFR raters did not focus on *accuracy* to the same extent as the Swedish raters. Instead the criteria specifically referred to in the CEFR rating scale (*accuracy*, *coherence*, *fluency*, *range*, and *interaction*) had a fairly even number of comments. In other words, there is no clear pattern of weighting the criteria. The only category that was significantly smaller than the others was *coherence*. Still, *coherence* is a large category with 11% of the coded comments. Consequently, the CEFR raters seem to adhere closely to the rating criteria and do not favour any of them significantly more than the others. In other research, it is often suggested that raters favour some performance features over others (see for example Brown, 2007). The results from the analysis of the CEFR raters' comments here do not seem to bring any convincing support to that finding.

The Swedish raters, in comparison, seemed to weight the criteria somewhat more, slightly favouring some over others. *Accuracy* was the most salient feature, with 24% of the coded comments. Further, *range* (18%), *coherence* (16%) and *interaction* (14%) seemed to play an equal role in rating decisions for the Swedish raters. An interesting difference between the rater groups is that *fluency* was not commented upon as much by the Swedish raters (10%), as it was by the external CEFR raters (19%), suggesting that this feature is less important to the Swedish raters. This could be explained by Swedish students' overall high proficiency in English. Possibly, Swedish raters take *fluency* more for granted since they know Swedish students are generally quite fluent in English, as compared to the CEFR raters for whom this factor is more important to comment on.

Finally, a result similar to that of the CEFR raters is that the categories not explicitly mentioned in the criteria had the lowest proportion of comments (about 11%). In other words, for both the CEFR raters and the Swedish raters, non-criterion features comprised only a small part of the comments. The findings in previous research confirm the result that raters include non-criterion, or self-generated features in their rating decision (May, 2006; Meiron, 1998; Orr, 2002). However, as stated above, they constituted a small proportion of the comments in the present study, whereas in previous studies they seem to have played a larger role.

For the Swedish raters, candidates' linguistic competence, i.e. *accuracy* and *range*, seemed to be highly salient (42%). This finding may seem somewhat surprising considering the fact that the holistic performance level descriptors used by the Swedish raters (i.e. the Swedish performance standards for course English 6, provided in Appendix 2) do not mention *accuracy*. *Accuracy* is only mentioned in the analytic assessment factors (Appendix 3), provided as a help and support for teachers in making their holistic judgement. In comparison, *range* is mentioned in the holistic performance level descriptors that the Swedish raters used, but rather vaguely: "students can express themselves in ways that are *varied*". In the assessment factors, the description of *range* is more explicit – range, variation, and complexity are to be taken into consideration. One reason for the CEFR raters' more balanced distribution of comments may be that these raters were experienced CEFR raters. In other words, they were used to rating with the help of scaled analytic descriptors, unlike the Swedish raters.

In summary, the results of the qualitative analysis of the written comments for both rater groups indicated that raters took a wide array of features into account in their holistic rating decision, although candidates' linguistic and pragmatic competences, as well as their interactional strategies seemed to be most salient. This finding coincides with the conclusion in Annie Brown et al.

(2005) that raters "take a range of performance features into account within each conceptual category and that holistic ratings are driven by all of the assessment categories rather than, as has been suggested in earlier studies, predominantly by grammar" (p. iv). The authors also state that the judges in their investigation of an English-for-Academic-Purposes Speaking Test "focused on the same general categories and tended to discuss the components of these categories in essentially similar ways" (p. 101), which is in line with the findings from the present study. The fact that there seems to be such strong agreement among raters as to the construct is positive with regard to validity.

Evaluative comments

In the analysis of rater orientations, the distribution of evaluative comments was also reported. Findings indicated that there were three categories, namely *accuracy*, *intelligibility*, and *task realisation*, for which a majority of the comments were Negative. For the other categories, the pattern looked rather similar with Positive as the largest category, Negative as the second largest and Mixed as the smallest. *Fluency* stood out somewhat with a fairly large proportion of Negative comments (32%), even though Positive comments were in a majority (52%) and there were also some Mixed comments (16%). For the other groups, Negative comments varied between 17% and 24%. The results for evaluative comments can be compared to Brown (2007). She only coded for Positive and Negative evaluative comments, and found that 55% of the comments on syntax were Negative, whereas 45% were Positive. In addition, she also found that all categories she used, except for strategies, had more Negative than Positive evaluative comments. Production, corresponding roughly to *fluency* and *pronunciation* in the present study, had as many as 81% Negative comments. However, a major difference is that Brown did not include a mixed category in her study, which could be one reason why our results diverge.

Analytic categories

In the results chapter, examples of comments for each main category and its subcategories were reported to further illustrate the research question about rater orientations. Evaluative comments in relation to the subcategories were also noted. In this section, some general remarks in relation to these data are made.

The main findings show that raters noted the same general categories, and their comments within the main and subcategories addressed the same kinds of features. In other words, raters seemed to understand and interpret the categories in a similar way.

As reported, linguistic features, *accuracy* and *range*, were the most highly salient to raters in the current study. In the *accuracy* category, comments pertained to vocabulary, phonology and grammar. Raters referred to, for example, frequency of errors, ability to produce well-functioning sentences, richness of vocabulary and language, nativeness of pronunciation, and adequacy and appropriateness of lexical choices. In many cases, *accuracy* was related to *intelligibility*; for example, lack of linguistic resources could lead to difficulty in understanding the candidate. There was a large majority of Negative comments on *vocabulary control*, whereas comments for *grammatical accuracy* and *phonological control* were more Mixed. In other words, the warning raised that raters might count all the errors a candidate makes, is partly justified, but on the other hand it is shown that raters also take notice of Positive features, for example good pronunciation, well-functioning syntax and complex grammar. It is important to emphasise that whereas the raters in the present study seem to find linguistic features, such as *accuracy*, highly important in the rating decision, this is not only because they find errors and slips in test-taker speech, but also because they are attentive to candidates who speak with good *accuracy*. As for the category *range*, comments referred to *variation, richness* and *sophistication* of the lexical and linguistic repertoire, including use of *idiomatic expressions*, as well as candidates' *ability to express viewpoints*. Positive evaluations were in the majority.

Comments pertaining to candidates' pragmatic competence, in the form of *coherence* and *fluency*, were also highly salient to raters (the second largest group). Comments on *coherence* were mainly Positive and referred both to *structure and organisation of speech*, as well as a candidate's ability to develop the topic, i.e. *the content of speech*. Furthermore, there was also a small category of comments referring to candidates' ability to vary formulations of what they want to say and adapt their language to the situation.

Fluency typically received general comments, which were mainly Positive, but there were also more specific comments on *pauses and hesitation*, which were mainly Negative or Mixed. The fact that comments on *pauses and hesitations* were predominantly Negative mirrors the results in Brown (2007). Where there was disfluency, for example *hesitation and pauses*, raters in many cases tried to infer the reasons for this behaviour. Brown (2007) noticed the same in her study.

Raters mentioned lack of linguistic resources (such as searching for a word) and *cognitive planning* as possible reasons for disfluency. However, it is not clear in all cases what the pauses and hesitations were attributed to. Brown (2007) makes the case that "lack of evidence cannot always be assumed to indicate non-mastery" (p. 122). In other words, *hesitations and pauses* that arise from cognitive planning are prevalent in native speakers' speech as well, and could thus be viewed as positive, or at least neutral, in a second/foreign language context as well. The problematic issue here is that when raters make inferences about reasons for disfluency, they seem to believe that pauses and hesitations are generally signs of shortcomings (Ginther, Dimova, & Yang, 2010).

Comments referring to interactional strategies were common and mainly Positive. There were three main subcategories of interaction, namely cooperating strategies, turn-taking strategies, and dominant or passive behaviour of the candidate/interlocutor. These categories can be compared to Ducasse and Brown (2009) where three main features of raters' perception of paired interaction were found, namely *non-verbal interpersonal communication*, *interactive listening*, and *interactional management*. Body language cannot be taken into account in the present study, because interactions were not video-filmed. However, *interactive listening*, which involves both showing involvement and supportive listening, and *interactional management*, which is about management of the topics and turns, can be found in the subcategories of *interaction* in the present study. Interactional management is actually part of both *interaction* and *coherence* (topic development) in the present study, since this category also refers to "candidates' ability to develop the conversation by extending the topic" (Ducasse & Brown, 2009, p. 436)

Further, Galaczi (2010) reviews research on the paired speaking test format and concludes that collaborative interactional skills include: *topic development skills* (expanding one's own and others' topics), *turn taking skills*, *active listening skills*, *equality and mutuality in interaction*, and *non-verbal support*. It is clear that these features are salient to raters in the present study as well, with the exception of non-verbal support, which cannot be analysed.

As regards equality and mutuality, raters were concerned with *asymmetric* interactions (Galaczi, 2008), i.e. when a candidate took over and dominated the discussion at the expense of the other candidate. Practically all raters commented on one pair, where the girl was considered to be dominating, but the raters had slightly different views on how this should be interpreted. In general, comments were Negative, indicating that this behaviour could have a negative impact on the grades. However, when looking at the results of the ranking, this candidate was ranked first among the Swedish and second among the CEFR raters. There was another candidate who was also perceived as dominant by some of the raters, and she was ranked second among the Swedish and first among the CEFR raters. In other words, a proficient but somewhat dominant speaker does not seem to have been penalised by raters for this kind of behaviour. On the other hand, the opposite, i.e. a candidate who does not speak much, was also a concern for raters. Many commented that they would have wanted to hear more from this candidate to be able to rate him/her fairly. However, there were also raters who believed that the more passive candidate was helped by the more talkative candidate to receive a higher score. Thus, raters did not completely agree on this issue, it appears.

Earlier studies on *interlocutor effects* show quite contradictory results, but they do seem to indicate that scores are not affected to a very great extent by different proficiency levels or personality traits (Berry, 1993; Davis, 2009; Nakatsuhara, 2009). One exception is Galaczi (2008), who studied paired candidate discourse and found three different patterns of interaction: *collaborative interaction*, *parallel interaction* and *asymmetric interaction*. Pairs with *parallel interaction*, i.e. two speakers who initiate and develop topics but do not build on each other's ideas, and *asymmetric interaction*, i.e. one dominant and one passive speaker, received the lowest scores for the criterion "Interactive Communication". In Galaczi's data of 30 paired candidate performances, only 10% were oriented towards an asymmetric pattern of interaction. However, even though they were few, Galaczi indicates that asymmetric dyads were the most problematic from a rating perspective. In summary, the present study confirms that pairing of candidates is an important issue for the paired speaking test. However, it does not seem to be the case that low equality in a conversation renders lower grades on the part of the dominant interlocutor.

Production strategies refer to candidates' ability to spot, backtrack and correct their own errors, *monitoring and repair*, as well as candidates' ability to paraphrase or use circumlocution when not finding the right vocabulary or grammatical structure, *compensating*. Comments were mainly Positive, indicating that raters reward test-takers who can monitor and cover gaps in their language.

Finally, the last three categories, *other*, *intelligibility* and *task realisation*, refer to features not explicitly mentioned in the rating criteria. Together they had a relatively low proportion of comments, about 12% of the total number of coded comments. *Task realisation* was the largest category in this group and comments in this subcategory were predominantly Negative. This was because raters made many comments on how well candidates summarised the short text they were given to read in advance. Raters' references to *intelligibility* were mainly Negative, indicating that this feature was taken into account in the rating decision when there was a problem. This result is in line with Brown (2007), in whose study 38 out of 40 comments on *intelligibility* were Negative. In many cases, *intelligibility* was attributed to other performance features, such as *accuracy*, *coherence* and *fluency*. Comments in the *other* category were about degree of confidence, degree of relaxation or use of "safe" language. In other words, these sorts of comments were more behaviour-based.

In addition to the categories just mentioned, another coding layer was added to the study in the form of *inter-* and *intra-candidate comparisons*. Raters frequently made comparisons between candidates in a pair with regard to differences or similarities between them. They also commented on the *interaction* in the pair, as well as on whether the two candidates seemed to be well matched in terms of proficiency level. In addition, raters also commented on candidates' development, or lack thereof, during the test. Somewhat different rater orientations were discovered in the two rater groups. CEFR raters seemed to make proportionally more comments on inter- and intra-candidate comparisons, compared to the Swedish raters. The more prominent emphasis on inter-candidate comparisons for CEFR raters may be a result of the fact that this speaking test model was new to them, whereas Swedish raters are used to rating paired orals. A hypothesis is that when raters rate paired orals without receiving specific training on what to focus on, more inter-candidate comparisons are made. Swedish raters have a more specific focus on individual assessment, since they both mark national tests and award final grades to candidates. A similarity, however, was that both Swedish and CEFR raters made many intra-candidate comparisons.

Making inter-candidate comparisons in a paired speaking test seems inevitable, because of the co-constructed nature of this kind of speaking task. The question is, however, how these comparisons between the candidates in the pair affect the individual grade. Moreover, it is a question that should be addressed in test specifications, so that raters are advised on how to handle this issue. Meiron (1998) focused on this question in her study and found that when candidates had different proficiency levels, the tendency for raters was to focus on linguistic features that were shared by the candidates, instead of salient features of the specific individual performances. In Pollitt and Murray (1996), findings indicated that when students had different proficiency levels, raters focused on the features of the lower-level candidate. In the present study, it was not possible to examine the relationship between the proficiency levels of the two candidates in the pair, and features that raters paid attention to. However, it is an important area in need of further investigation.

Finally, raters in the present study made inferences about candidate behaviour and reflected on their rating decision (5% of the total number of coded comments, excluding evaluative response), and this group of comments was named *rater reflections*. The largest proportion of comments within this category pertained to justifications of rating decisions and more general comments on candidate behaviour, whereas comments on matching of candidates and how this affected the overall performance or grade made up a minor subcategory. Especially CEFR raters compared candidate performance to the CEFR descriptors for different levels when reflecting on their rating decision. This might indicate that the Swedish raters, who used holistic performance level descriptors, were not able to refer to the descriptors for the different grading levels to the same extent as the CEFR raters.

The finding that raters make inferences is also made in Pollit and Murray (1996) and Brown (2007), who found that raters' comments consisted of inferences about, for example, candidates' personality, maturity, world knowledge, and exam-consciousness. As can be seen, the focus of the inferences in this study, which examines a paired speaking task, was on general behaviours, but also on the matching of candidates, confirming once again that this is an important aspect for raters in paired orals.

In summary, the analysis of analytic categories indicates that raters made similar comments on candidate performances within a wide range of categories, both pertaining to linguistic and non-linguistic features. It was shown that there seem to be many features that contribute to raters' perceptions of oral proficiency in a paired speaking test, making the rating process a complex and challenging task.

Relationship between comments and scores

An attempt was made to analyse the potential relationship between comments and scores. First of all, the distribution of coded features per candidate was reviewed. Findings showed that the distribution of categories across candidates generally seemed to be fairly similar regardless of proficiency level. It was also shown that individual candidates seemed to have at least one or two features each that were more salient to raters than others, and these features seemed to be quite individual. When looking at candidates who were ranked at either end of the scale (high or low), *interaction* seemed to be more salient for the two highest-ranking candidates, whereas *intelligibility* seemed more salient for the two lowest-ranking candidates. However, there was another candidate with low grades and the raters had not commented on *intelligibility* in her case. In other words, the results should be seen as tentative and of limited applicability.

A clearer pattern emerged, however, when evaluative response per candidate was looked into. Results showed that the three lowest-scoring candidates had a majority of Negative comments, whereas the other candidates had a majority of Positive comments. In addition, the two highest-scoring candidates had a proportionally higher distribution of Positive comments than the rest, and very few Negative comments. Previous research in this area shows somewhat different results. Pollit and Murray (1996) found that grammatical accuracy was more salient to raters at the lower levels and sociolinguistic and stylistic competence at higher levels. This is not confirmed in the current study, where *accuracy* had a fairly even distribution for all candidates, irrespective of proficiency level. However, as mentioned above, the coding of the present study allowed insight into what kind of evaluative response was made, and it was shown that there were both Positive and Mixed evaluations of *accuracy*, even though the Negative comments were in the majority. Furthermore, Brown (2007) found that comprehensibility, corresponding to *intelligibility* in the present study, and production, corresponding to *fluency* and *phonological control,* were more salient at the lower proficiency levels. The relationship between intelligibility and proficiency level is tentatively confirmed in the present analysis, where *intelligibility* was found to be proportionally more salient for the two lowest-ranking candidates, however not for the third lowest-ranking student.

Further analyses were made to investigate whether the same performance elicits comments of the same kind or not from different raters. To this end, three types of oral performances were chosen: (1) two performances with a large range of grades, indicating that raters are in disagreement about the grade for this particular candidate, (2) one performance with little range of grades, indicating that raters agree to a large extent, and (3) one performance where Swedish and CEFR raters seemed to disagree slightly on the ranking. Results showed that in some cases, raters noticed fairly similar features across performances, but there were differences in how they evaluated them (i.e. positively or negatively). In other cases, raters actually seemed to base their decision on partly different features, thus indicating that they may focus on different features of the same performance when making their judgement. These results can be compared to Brown (2007), who found that it was generally the case that raters focused on different features of a performance, and this could be a reason why they award different scores to the same performance. Furthermore, Orr (2002) draws the conclusion from his study that raters, in addition to heeding many non-criterion aspects of the performance, also heeded different features of the rating criteria. One example was that raters who awarded the same score still perceived the performance in different ways. Finally, a study by Douglas and Selinker (1992) investigating the relationship between test-taker discourse and scores, found that raters who use the same scoring rubric might award similar scores to candidates who produce qualitatively different performances.

Chapter Seven: Conclusion

In this chapter, some concluding remarks are made, based on the main findings of the present study. In particular, comments are given regarding positive and negative aspects of validity. Finally, didactic implications are outlined and some suggestions for future research are made.

Concluding remarks

In the present study, rater-related variability was examined in relation to the Swedish raters' judgements of candidates' speaking proficiency. Findings showed signs of variability of ratings as well as differences in consistency, which was expected given the complex nature of performance testing. In addition, inter-rater reliability was computed and was found to be reasonable, even though the correlation coefficients were not as high as in some previous research on speaking tests, thus showing room for improvement.

Considering these results, indicating rater-related variability, double marking for the paired speaking test in the Swedish national tests of English is highly recommended. This is also in line with recommendations made in previous research in relation to rating of performance tests. Using a procedure where at least two raters are involved in the rating decision would contribute both to reliability and validity. Of course, further research would be needed to confirm this. It is also possible to employ methods such as multifaceted Rasch analysis (Eckes, 2005, 2009) as a means to achieve a better understanding of the variability of rater severity in research studies.

It was also found that ranking of the performances between the CEFR and Swedish raters was similar, pointing to agreement between the two rater groups. Further, the CEFR group rated most of the performances (10/12) at level B1+ and above, which is in line with the intention of the test. This is an interesting finding, since, up until now, very little empirical validation has been done in relation to Swedish national tests of English and the CEFR levels (Erickson, 2010b). It also strengthens the validity of the assessment.

As regards salient features, raters seemed to pay attention to the same categories and displayed similar ways of commenting on the performances. They even used similar terminology. In other words, raters seemed to understand and interpret the categories in a similar way, and it could therefore be argued that they have a broad level of agreement regarding the construct that the test intends to measure. This also is positive for the validity of the assessment.

Further, raters did not seem to favour any features significantly more than others. Even though it was found that comments pertaining to linguistic competence seemed to be the most salient, the other features, which focus on pragmatic competences as well as interactional and production strategies, were highly salient too. This broad view of candidates' communicative competence, displayed in the rater comments, enhances validity.

However, there was a slight difference in rater orientations between the CEFR and Swedish raters. The findings showed that the distribution of coded comments was slightly more evenly balanced for the CEFR raters than for the Swedish raters, who seemed to find *accuracy* and *range* particularly salient. The conclusion thus seems to be that the Swedish raters weight the criteria somewhat more as compared to the CEFR raters. Nevertheless, as pointed out already, there were no significant differences regarding rank ordering or consistency of rating between the groups, which seems to indicate that this did not have any effect on the scores given.

One of the positive aspects noticed was that raters focused mainly on the criteria described in the descriptors, and assessment factors for the Swedish raters. There seemed to be only a small proportion of non-criterion features in the rater comments. Moreover, these non-criterion features were in no way irrelevant to the test, the main proportion dealing with candidates' ability to summarise a short text, which was part of the test requirements.

The co-constructed nature of the paired speaking task seems to bring some challenges for raters. A large proportion of comments pertained to the *interaction* in the pairs, and there were also frequent *inter-candidate comparisons*, as well as *rater reflections* on the matching of candidates. In some cases, raters had different opinions on whether a dominant partner helped or took over in relation to the other candidate in the pair. There were also discussions on how this could affect the rating decision. Consequently, it seems that the matching of candidates is an essential aspect of this test format, which has consequences both for raters and for the individual test-taker. Thus, rater and interlocutor effects pertaining to paired interaction are important to account for in the speaking test format. May (2009) even proposes shared scores for interactional skills in paired speaking tests. This may provide a possible solution to this issue, but of course it would further complicate the already complex rating procedure.

On the other hand, there are many positive aspects of the paired speaking test format. If one of the constructs we want to measure is interaction, a paired conversation is definitely appropriate to use as a test format. As can be seen in the results, raters made frequent comments on the interactional skills of the candidates and in a majority of the cases, positive examples were noted where candidates cooperated and helped move the conversation forward.

Finally, the relation between raters' justifications of the scores, in the form of written comments in this study, and the scores themselves, proved complex to analyse. The tentative comparison made between comments and scores seems to point to two problematic areas for the reliability and validity of scores. Firstly, raters may heed the same features of a performance but evaluate them differently. Secondly, raters may heed partly different features of a performance, thus basing their decisions on different perceptions to some extent. However, the relationship between comments and scores needs to be explored further, with a more comprehensive analysis, to enable any firm conclusions.

In sum, this study has shown that the rating of communicative performances is a complex task. As stated at the onset, performance assessment always involves subjective judgements and thus rater variability needs to be taken into account. However, due to its inherent complexity, it is not realistic to expect to find ratings that are consistent across all performances and across all raters in performance testing. Nevertheless, taken together, the findings of this study indicate that to a very large extent, raters are in agreement about the construct. Furthermore, rater variability and rater effects do exist, but are reasonable considering the nature of performance-based testing. One step to improve reliability may be to use double marking.

Didactic implications

The findings of this study also have didactic implications for the interpretation of oral test scores, especially with regard to the speaking test in the Swedish national test of English at upper secondary school level. First of all, cooperating with colleagues in the marking process has been suggested as a means to improve inter-rater reliability in the speaking test. This has also been emphasised at the national level, by the national school authorities as well as the university departments responsible for test development (Erickson, 2009; The

Swedish Schools Inspectorate, 2012). A questionnaire is routinely conducted with the teachers who mark the national test of English and in the questionnaire from spring term 2013, when the present study was conducted, teachers answered a question about the extent to which they "co-rated" (i.e. rated together with one colleague or more) the test[12].

For the essay, co-rating was quite common: only 11% of the teachers rated the essays solely on their own. However, for the speaking test fewer teachers used co-rating: 57% answered that they were the only raters. This shows that the speaking test is co-rated to a lesser degree than the essay. This may be due to practical reasons, since the conversations need to be recorded in order for another teacher to listen to and rate the performances, or two teachers need to be present at the same time in the test situation (which might be difficult to organise). In the guidelines for teachers, it is strongly recommended that the speaking test be recorded, since it facilitates co-rating and thereby enhances fairness, and also provides the opportunity to go back and listen to the conversation one more time. In the 2013 questionnaire referred to above, 50% of the respondents answered that they recorded the test, which is quite positive but still shows considerable room for improvement. Hopefully, the findings of this investigation, where co-rating is highlighted as important, will indicate to everyone involved, including head teachers, that cooperation between colleagues in the rating process is equally important for the speaking test and the essay.

Furthermore, the raters in the present study were very positive towards the opportunity to discuss their rating decisions with other teachers in the short group discussion we had at the end of the rating seminar. Organising gatherings where teachers can listen to the same performances and discuss and compare their ratings, as well as features they pay attention to and rating criteria they employ, would provide useful in-service training, and could also lead to more reliable test results in the long run. For example, when teachers award different scores to the same performance, rater orientations could be compared to find differences and similarities. In addition, the different components of communicative competence could also be discussed in relation to test performance. As was seen in the results of the present study, the main focus

[12] Results from the regular questionnaires distributed to teachers who mark the national tests of English in Sweden are published on the National Assessment Project webpage: www.nafs.gu.se. The results from the questionnaire from spring term 2013 were retrieved from: www.nafs.gu.se/prov_engelska/engelska_gymn/resultat/.

seems to be on linguistic features for the Swedish raters; possibly a broader communicative view may be desirable. It can also be mentioned that the test development group at the University of Gothenburg has recently suggested to the National Agency for Education that materials for this type of activity could be developed and offered to schools for in-service purposes. This has been done before for French, German and Spanish in 2006, and reactions were very positive.

The co-constructed performance in the paired speaking test format also has didactic implications. Consequently, the organisation of the oral part of the national test at schools should be considered crucial for validity. Not only should teachers be provided with enough time to mark the test together with colleagues as far as possible (and thus also record the test); the matching of candidates with regard to both their proficiency level and their personality also needs to be taken into account when organising the test. Especially asymmetric pairs, with one dominant and one more passive candidate, are problematic from a rating perspective. It is of course a complex undertaking to organise pairs with matching proficiency levels and personality, and it may not always be possible, but this aspect should at least be considered. This is also emphasised in the teacher guidelines for the test.

Further, examiner intervention should be addressed. It says in the test guidelines, that the examiner (i.e. the teacher in this case) should "keep in the background" and let the students show that they can initiate discourse, interact and advance the conversation on their own. In addition, the teacher should encourage students to give each other equally much space in the conversation. Raters in the present study commented both on excessive and insufficient examiner intervention. This seems like a somewhat problematic area, where perhaps more specific instructions and examples should be given.

Finally, this study provides some pedagogical insights for the classroom. Strategic competence is of primary importance in language education in general, not least in oral interaction (Malmberg, 2000). Learners might benefit from explicit teaching of interaction and production strategies, such as showing active listening skills, initiating and ending turns, using paraphrasing and circumlocution. As is stated in the CEFR, these strategies serve as a bridge between "the learner's resources (competences) and what he/she can do with them (communicative activities)" (Council of Europe, 2001, p. 25). The importance of collaborative interaction with a high degree of mutuality and equality in the pair should also be highlighted to students taking the test.

Future research

There are some possible options for future research, emanating from the present investigation. Firstly, it would be interesting to further explore the relationship between rater comments and scores in a more systematic way. This seems like a complex but important area to investigate from the point of view of validity and reliability. Consequently, further development of the present study may be to compare the relationship between rater comments, scores and test-taker discourse. For example, rater comments on test-taker performance could be examined in relation to (1) the scores that raters awarded, and (2) test-taker discourse. Also, it would be interesting to see to what extent raters focus on features shared by candidates, instead of salient features of the individual performances, as suggested in previous research.

Further, as mentioned in the concluding remarks, double marking of paired orals would be interesting to investigate. Paired (or group) discussions where raters compare their reasons for awarding a score to a particular candidate, could be the focus of analysis, as a complement to justifications of individual ratings.

Another possibly confounding factor, not addressed in this study, is the fact that teachers are given the choice in the test specifications to either organise pairs, or groups of three students for the speaking test. The regularly distributed questionnaire of teachers' opinions following all national tests, showed that 65% of teachers in spring 2013, when the test in the present study was used, answered that they organised their students in pairs, 21% in groups, and 14% used both types. Hence, a majority used the paired speaking test format, but as many as 35% used either groups (of three) or a combination of both pairs and groups. As the present study only included paired conversations, the effect of group size was not considered. In future research, it would therefore be interesting to compare ratings where candidates were first placed in pairs, and then groups of three, to see if this has an effect on individual scores.

Finally, in future studies it would be of considerable interest to further explore the results for the total groups of raters and/or students. In this, studies of subgroups, not least based on gender, seem highly relevant.

Swedish summary

Inledning

Föreliggande studie avser bedömning av muntlig språkfärdighet i engelska i ett så kallat autentiskt prov eller performance-prov. Denna typ av prov innehåller uppgifter som är utformade för att så långt som möjligt likna verkliga situationer, där eleverna får utföra olika slags uppgifter eller aktiviteter för att visa upp sin förmåga (McNamara, 1996). Detta kan jämföras med mer traditionell bedömning där ofta enskilda frågor i ämnet besvaras. Ett exempel på ett autentiskt prov är det parsamtal som genomförs i den muntliga delen av de nationella proven i engelska. Bedömningen av ett sådant prov, nämligen kursprovet för engelska 6 i gymnasieskolan, är också fokus i denna studie. Mer specifikt undersöks bedömarvariabilitet och bedömarprocess.

En svårighet med autentisk bedömning, som alltså mäter komplexa kunskaper, är att det finns en risk för *bedömarvariabilitet*, eftersom subjektiva uppfattningar påverkar bedömningen. Termen *bedömareffekter* beskriver variation i bedömningen, som kan hänföras till bedömare snarare än elevens prestation. Eftersom bedömareffekter utgör ett hot mot validitet och reliabilitet (Messick, 1989) är det viktigt att försöka att begränsa deras inverkan.

En av de vanligaste bedömareffekterna är att en bedömare konsekvent bedömer strängare eller mildare jämfört med andra bedömare (Bachman et al., 1995). Det finns dock flera andra faktorer som kan påverka bedömningen i autentiska prov. Till exempel kan bedömare tolka och använda bedömningskriterierna på olika sätt, och därigenom ge olika betyg till samma elevprestation, eller ge samma betyg men av helt olika skäl (McNamara, 1996; Orr, 2002). Det har även visats i tidigare forskning att bedömare lägger märke till olika aspekter av elevprestationer beroende på vilken språklig nivå eleven befinner sig på (Adams, 1980; Annie Brown, 2007; Pollitt & Murray, 1996). En svårighet med bedömning av parsamtal är dessutom att interaktionen skapas tillsammans av deltagarna, vilket komplicerar den individuella bedömningen. Forskning visar till exempel att matchningen av elever är viktig, då variabler hos samtalspartnern, som t.ex. språknivå och personlighetstyp, kan påverka interaktionen på skilda sätt, såväl positivt som negativt (Davis, 2009; Galaczi, 2008).

De forskningsresultat som finns kring hur bakgrundsvariabler hos eleverna påverkar bedömningen är dock inte entydiga. Vissa studier (Davis, 2009; Iwashita, 1996) visar att mängden talat språk kan påverkas av att eleverna har olika språknivåer, men att detta i sin tur inte påverkar betygen. Galaczis (2008) undersökning pekar dock på att det finns en tydlig koppling mellan elevernas gemensamt konstruerade samtal och deras betyg.

Nationella prov i engelska

De nationella proven i Sverige konstrueras på uppdrag av Skolverket av olika universitet i landet. Göteborgs universitet, Institutionen för pedagogik och specialpedagogik, är ansvarig för att ta fram de nationella proven i främmande språk, samt olika typer av bedömningsstöd i engelska, franska, spanska och tyska. Detta görs i en kollaborativ process tillsammans med lärare, forskare och elever (Erickson & Åberg-Bengtsson, 2012). Proven i engelska innehåller tre delprov: receptiva färdigheter testas i hör- och läsförståelseuppgifter, skriftlig produktion och interaktion i en uppsats och muntlig produktion och interaktion i ett parsamtal. Det muntliga provet genomförs i par (eller grupper om tre) och behandlar ett tema (t.ex. stress). I första delen av det prov för engelska 6 som ingår i studien testas muntlig produktion, då eleverna får sammanfatta en kort text de har läst och sedan diskutera denna med sin partner; i den andra delen är fokus på interaktion, och elevernas diskuterar och argumenterar utifrån givna frågor eller påståenden.

Kommunikativ språkbedömning

Under 1970- och 80-talen började kommunikativa teorier om språkinlärning påverka hur språkprov utformades. Tidigare hade proven testat delar av språklig förmåga separat utan tydlig kontext (Oller, 1973). De nya kommunikativa språkproven hade istället fokus på att bedöma språk i en tydlig kontext och med så autentiska uppgifter som möjligt. De olika språkfärdigheterna (tala, läsa, skriva och lyssna) används dessutom ofta i kombination med varandra i kommunikativa språktest. Den mest inflytelserika teorin bakom den kommunikativa språksynen härrör sig från Dell Hymes (1972), som introducerade begreppet kommunikativ kompetens, i vilket bruket av språk i olika sociala sammanhang tillmättes central betydelse.

Gemensam europeisk referensram för språk

Gemensam europisk referensram för språk (GERS) publicerades av Europarådet år 2001 och är baserad på mer än tjugo års forskning (Council of Europe, 2001). Dess huvudsyfte är att ge en gemensam grund för lärande, undervisning och bedömning av andraspråk och främmande språk och på så sätt också underlätta internationell samverkan. GERS bygger på en kommunikativ och handlingsorienterad syn på språkinlärning och språkanvändning, vilket innebär att språkinlärare ska kunna använda språket för olika syften och i olika sammanhang. I GERS finns skalor för olika kommunikativa språkaktiviteter och strategier, samt för de olika delarna av den kommunikativa språkkompetensen: lingvistisk, pragmatisk och sociolingvistisk kompetens. Vidare är GERS-skalorna indelade i olika nivåer, så kallade gemensamma referensnivåer, nämligen A1, A2, B1, B2, C1 och C2, där A står för nybörjarnivå, B för självständig och C för en avancerad nivå. Kurserna i svenska för invandrare, engelska och moderna språk i det svenska skolsystemet är explicit knutna till GERS. Ett godkänt resultat i kursen engelska 6, till exempel, som ingår i denna undersökning, ska motsvara lägstanivån för B2 i GERS. Det har gjorts en del textuella jämförelser mellan nivåerna i GERS och de svenska kursplanerna i främmande språk, men hittills endast få empiriska undersökningar. Därför är ett sekundärt syfte med denna studie att tentativt jämföra de svenska kunskapskraven i kursen engelska 6 med GERS referensnivåer.

Syfte

Denna studie undersöker bedömning av muntlig färdighet i det nationella provet i kursen engelska 6 på gymnasienivå. Det första syftet är att studera variabilitet i bedömningarna. Det andra syftet är att undersöka bedömarnas beslutsprocesser genom att identifiera och jämföra *bedömarprofiler*, det vill säga aspekter i elevprestationerna som bedömarna tar hänsyn till när de sätter betyget. Slutligen är ett sekundärt syfte att göra en tentativ, empirisk jämförelse av de svenska, nationella kunskapskraven och referensnivåerna i GERS. Forskningsfrågorna är som följer:

1. Vad kan uppmärksammas vad gäller variabilitet i bedömningarna?
2. Vilka aspekter av elevernas prestationer är framträdande för bedömare när de fattar sina beslut om betyg?

3. Vilken är den möjliga relationen mellan betyg och bedömarnas motivering av dessa betyg?
4. Vilka nivåer i GERS anser de externa bedömarna att de svenska eleverna ligger på?

Material och metod

Data och deltagare

Data i studien består av betyg och bedömarnas skriftliga kommentarer som motiverar betygen. Den första gruppen bedömare är gymnasielärare i engelska i Sverige (*n* = 17), som individuellt bedömde sex inspelade parsamtal i relation till nationella kunskapskrav. Dessutom bedömde två grupper av europeiska bedömare (*n* = 14) samma elevsamtal i relation till referensnivåerna i GERS, detta med syfte att göra en tentativ, empirisk jämförelse av de svenska, nationella kunskapskraven och referensnivåerna i GERS.

De svenska bedömarna kommer från två olika städer, och från olika skolor. De europeiska bedömarna är vana vid GERS-baserad bedömning och kommer från Finland och Spanien. I dessa länder används skalor baserade på referensnivåerna i GERS i större utsträckning än i svenska sammanhang.

Bedömarna använde två olika skalor; de svenska en tiogradig skala baserad på det svenska betygssystemet, de europeiska en niogradig baserad på referensnivåer i GERS. På grund av denna olikhet i betygsskalor är syftet inte att jämföra deras bedömningar. Vad gäller de svenska bedömarna är fokus på att undersöka variabilitet, vad gäller de europeiska bedömarna på att studera vilka nivåer i GERS som de svenska elevernas prestationer anses motsvara. Det som dock går att jämföra är de två bedömargruppernas ranking av eleverna, eftersom denna inte bygger på betygsskalorna. Dessutom går det att jämföra de svenska och europeiska bedömarnas bedömarprofiler, eftersom båda grupperna skrev kommentarer där de motiverade betygen.

Analys av data

Data samlades in under en dag, vid olika tillfällen för de olika bedömargrupperna. Eftersom data består av en kvantitativ del med betyg och en kvalitativ del med bedömarnas kommentarer till betygen, delades analysen in i två delar. I den kvantitativa delen gjordes deskriptiva analyser bland annat av medelvärden, spridning, korrelationer och reliabilitet. Den kvalitativa delen undersöktes enligt metoder för *verbal protocol analysis* (VPA). Bedömarnas kommentarer delades in i segment, som utgjorde en enhet eller idé, och kodades med hjälp av ett kodningsschema. Kodningskategorierna bygger på bedömningskriterierna som bedömarna använde, samt skalor för kommunikativ kompetens och kommunikativa strategier beskrivna i GERS (Council of Europe, 2001).

Resultat

Den statistiska analysen av de svenska bedömarnas betygssättning visar på rimlig samstämmighet, även om viss variabilitet förekommer både vad gäller betyg och korrelationer mellan bedömarna. Den deskriptiva statistiken visar att det finns tydliga bedömarprofiler med skillnader i stränghet. Till exempel varierar medelbetygen för bedömarna mellan 5,6 och 8,0 på den tiogradiga skalan. Det framkom också att vissa elevprestationer var mer svårbedömda än andra, och därmed hade större variabilitet. Vidare låg medianen av de parvisa korrelationer mellan bedömarna på .77 med Spearman's rho och .66 med Kendall's tau, vilket kan ses som relativt god samstämmighet men med utrymme för förbättring. Cronbach's alpha, som mäter den interna konsistensen i gruppen, var dessutom mycket hög, .98.

Resultaten visar 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 för alla elevprestationer utom två. De två elevprestationer som bedömdes ligga under provets minimumnivå av de europeiska bedömarna hade även bedömts som underkända av några av de svenska bedömarna. Rankingen av elevprestationer jämfördes mellan den svenska och europeiska gruppen och resultaten visar på stora likheter.

Innehållsanalysen av de skriftliga kommentarerna, med hjälp av NVivo 10, pekar på att bedömarna tar hänsyn till en mängd olika aspekter i sin holistiska bedömning, men att elevernas lingvistiska och pragmatiska kompetenser, samt deras interaktionsstrategier, verkar vara mest framträdande. Bedömarna höll sig väl till bedömningskriterierna, och kommenterade andra aspekter i relativt liten utsträckning. 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, såsom hur elevernas prestation påverkades av den andra partnern. De gjorde också jämförelser mellan eleverna i paret, till exempel i förhållande till likheter och skillnader, språklig nivå och interaktionen mellan eleverna. Vidare uppmärksammade bedömarna i stort sett samma aspekter av elevprestationerna och använde liknande sätt att uttrycka sig på, vilket tyder på en god samstämmighet angående den kompetens som avsågs. En tentativ jämförelse mellan bedömarnas kommentarer och betyg visar också att fördelningen av kommentarer för de kodade kategorierna var liknande oavsett elevernas språkliga nivå, men att bedömarna i vissa fall värderade aspekterna olika.

Diskussion och slutsatser

Resultaten visar på bedömareffekter och bedömarvariabilitet, vilket var förväntat med tanke på att det muntliga parsamtalet i nationella provet i engelska är ett exempel på så kallat performance-prov, eller autentiskt prov. Autentisk bedömning är komplex, eftersom ett flertal aspekter tas hänsyn till i bedömningen. I detta muntliga prov med parsamtal ska bedömaren till exempel tolka bedömningskriterierna och applicera dem på elevprestationen, samt ta hänsyn till interaktionen mellan eleverna i sin bedömning. En åtgärd för att öka validitet och reliabilitet är därför att använda sambedömning då två bedömare diskuterar sina betygsgrunder för att sedan kunna fatta ett beslut om betyg.

Vad gäller de aspekter som är framträdande för bedömare när de fattar beslut om betyg, visar analysen av bedömarnas kommentarer att de tar hänsyn till olika delar av den kommunikativa språkkompetensen, med att de lingvistiska och pragmatiska aspekterna, samt elevernas interaktionsstrategier, verkar vara mest framträdande. Däremot kommenterade bedömarna inte elevernas sociolingvistiska kompetens i någon större utsträckning, vilket kan ha att göra med att provet inte ger förutsättningar för detta.

Positivt för validitet är att bedömarna verkade vara överens om vilka aspekter som ska bedömas. De kommenterade på ett liknande sätt och använde till och med liknande terminologi. Resultaten tyder även på att bedömarna inte viktar kriterierna i någon större utsträckning. Det är också positivt för validiteten att bedömarna fokuserade på och använde bedömningskriterierna i en mycket stor utsträckning. De kommentarer som inte hänvisar direkt till kriterierna utgjorde ca 10% av hela materialet. Dessa kommentarer var även högst relevanta även om de inte direkt beskrivs i kriterierna.

Analyserna av relationen mellan kommentarer och betyg var komplexa. De tentativa resultaten visar att fördelning av de aspekter bedömarna uppmärksammade var liknande oavsett elevens språkliga nivå, dock med några undantag. Två områden som kan vara problematiska för reliabilitet och validitet framkom, och som därför kräver djupare undersökning. Dels kan bedömare uppmärksamma samma aspekter av en elevprestation men värdera dem olika, dels kan de uppmärksamma delvis olika aspekter i samma elevprestationer, och alltså basera sitt beslut på olika grunder.

Bedömarna gjorde många kommentarer angående interaktionen i paret, och de gjorde även jämförelser mellan de två eleverna, till exempel angående språklig nivå. Bedömarna var dock inte alltid överens om hur interaktionen mellan eleverna påverkade betyget, till exempel då en av eleverna var mer dominant än den andra. Slutsatsen är att sättet på vilket eleverna paras ihop är en viktig fråga. Det fanns även många positiva exempel på hur interaktionen och samarbetet mellan eleverna fungerar för att utveckla och föra samtalet vidare. Detta tyder på att provet fungerar väl för att mäta muntlig interaktion.

Didaktiska implikationer

Studien har betydelse för hur muntliga provresultat i främmande språk kan tolkas och förstås. Resultaten visar på bedömareffekter och bedömarvariabilitet, varför sambedömning starkt rekommenderas. Detta har också påpekats i tidigare forskning och rekommenderas även av Skolverket. I den enkät som genomförs med lärare i anslutning till de nationella proven visar resultat från våren 2013, då denna studie genomfördes, att uppsatsen sambedöms i stor utsträckning men inte den muntliga delen av provet. Förhoppningsvis kan denna studie, som visar på vikten av sambedömning för att öka reliabiliteten, bidra till en större medvetenhet om att det är lika viktigt att sambedöma den muntliga delen av nationella provet som den skriftliga. Dessutom rekommenderas kompetensutveckling då lärare får bedöma elevsamtal och diskutera sin bedömning och sina bedömningsgrunder. Hur eleverna paras ihop är också en viktig aspekt av provet. Både språklig nivå och personlighet bör tas hänsyn till.

Slutligen visar studien även att lärare kan förbereda eleverna för det muntliga provet genom att i undervisningen ta upp vikten av strategisk kompetens, både vad gäller interaktionen och för att lösa språkliga problem.

References

Adams, M. L. (1978). Measuring foreign language speaking proficiency: a study of agreement between raters. In J. L. D. Clark (Ed.), *Direct testing of speaking proficiency: theory and application* (pp. 129-149). New Jersey: Educational Testing Service.

Adams, M. L. (1980). Five coocurring factors in speaking proficiency. In J. R. Firth (Ed.), *Measuring spoken language proficiency* (pp. 1-6). Washington D.C.: Georgetown University Press.

Aijmer, K. (2004). Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies. Special Issue., 3*(1), 173-190.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford [u.a.]: Oxford Univ. Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257. doi: 10.1177/026553229501200206

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.

Bagarić, V., & Mihaljević Djigunović, J. (2007). Defining communicative competence. *Metodika, 8*(14), 94-103.

Bejar, I. I. (1985). *A Preliminary Study of Raters for the Test of Spoken English*. Princeton, N.J.: Educational Testing Service.

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. (2004). *A study of the interaction between individual personality differences and oral performance test facets*. (Unpublished PhD thesis), King's College, University of London.

Bringer, J. D., Johnston, L. H., & Brackenridge, C. H. (2004). Maximizing Transparency in a Doctoral Thesis1: The Complexities of Writing About the Use of QSR*NVIVO Within a Grounded Theory Study. *Qualitative Research, 4*(2), 247-265. doi: 10.1177/1468794104044434

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

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15. doi: 10.1177/026553229501200101

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, A., & Davies, A. (1999). *Dictionary of language testing*. Cambridge [u.a.]: Cambridge Univ. Press.

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

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

Börjesson, L. (2012). Om strategier i engelska och moderna språk [On strategies in English and modern languages]. Retrieved March 30, 2013, from http://www.skolverket.se/polopoly_fs/1.177027!/Menu/article/attachment/Strategier i engelska och moderna spr%C3%A5k.pdf

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards, & Schmidt, R. W. (Ed.), *Language and Communication* (pp. 2-27). London: Longman.

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. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33. doi: 10.1177/026553229501200102

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

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: M.I.T. Press.

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

Council of Europe. (2009). *Relating examinations to the Common European Framework: a manual*. Council of Europe Retrieved from http://www.coe.int/t/dg4/linguistic/manuel1_en.asp.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51. doi: 10.1177/026553229000700104

Cumming, A. (2008). Assessing Oral and Literate Abilities. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of Language and Education* (2 ed., Vol. 7: Language Testing and Assessment, pp. 3-18). New York: Springer

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

Douglas, D., & Selinker, L. (1992). Analysing oral proficiency test performance in general and specific purpose contexts. *System, 20*, 317-328. doi: 10.1016/0346-251X(92)90043-3

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

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies.* Oxford: Oxford University Press

Eckes, T. (2005). Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly, 2*(3), 197-221. doi: 10.1207/s15434311laq0203_2

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H).* Strasbourg, France: Council of Europe/Language Policy Division.

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

Erickson, G. (1991). Muntliga prov i engelska i årskurs 8. [Oral tests of English in grade 8.]. *LMS Lingua, 3*, 97-102.

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

Erickson, G. (2010a). A New Look at Teaching and Testing: English as Subject and Vehicle. In T. Kao & Y. Lin (Eds.), *Good Practice in Language Testing and Assessment – A Matter of Responsibility and Respect* (pp. 237-258). Taipei, Taiwan: Bookman Books Ltd.

Erickson, G. (2010b, 29- 30 October). *Putting the CEFR to Good Use – A Collaborative Challenge.* Paper presented at the IATEFL Testing, Evaluation and Assessment Special Interest Group (TEA SIG) and EALTA Conference, Barcelona, Spain.

Erickson, G. (2012). National assessment of foreign languages in Sweden. 2014, from http://www.nafs.gu.se/digitalAssets/1389/1389767_nat_assessment_of_foreign_lang_in_swe.pdf

Erickson, G., & Åberg-Bengtsson, L. (2012). A collaborative approach to national test development. In D. Tsagari & I. Czepes (Eds.), *Collaboration in Language Testing and Assessment* (pp. 93-108). Frankfurt: Peter Lang Verlag.

Ffrench, A. (2003). *The change process at the paper level.* Cambridge: Cambridge University Press.

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

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

Fulcher, G. (2008). *Language Testing and Assessment.* New York: Springer.

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. (2010). *Paired Speaking Tests: An Approach Grounded in Theory and Practice.* Paper presented at the Recent Approaches to Teaching and Assessing Speaking, IATEFL TEA SIG Famagusta Conference Proceedings, Canterbury, UK.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*(3), 379-399. doi: 10.1177/0265532210364407

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

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

Habermas, J. (1970). Toward a theory of communicative competence. *Inquiry, 13*, 360-375.

Hadden, B. L. (1991). Teacher and Nonteacher Perceptions of Second-Language Communication. *Language Learning, 41*(1), 1-20. doi: 10.1111/j.1467-1770.1991.tb00674.x

Haertel, F. (1992). Performance measurement. In M. C. Alkin (Ed.), *Encyclopedia of Educational Research* (6th ed., pp. 984-989). New York: MacMillan.

Harley, B., Allen, Patrick, Cummins, Jim, Swain, Merrill. (1990). *The Development of Second Language Proficiency*. Cambridge: Cambridge University Press.

Hasselgren, A. (1998). *Smallwords and valid testing.* Dept. of English, University of Bergen, Bergen.

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing, 13*(1), 53-61. doi: 10.1177/026553229601300104

Hsieh, C.-N. (2011). *Rater Effects in ITA Testing: ESL Teachers' versus American Undergraduates' Judgments of Accentedness, Comprehensibility, and Oral Proficiency.* (Ph.D.), Michigan State University, Ann Arbor.

Hymes, D. H. (1971). Competence and performance in linguistic theory. In R. Huxley & E. Ingram (Eds.), *Language acquisition: Models and methods* (pp. 3-28). London: Academic Press.

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

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51-66.

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

Iwashita, N., & Grove, E. (2003). A comparison of analytic and holistic scales in the context of a specific- purpose speaking test. *Prospect, 18*(3), 25-31.

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

Jakobovits, L. A. (1970). *Foreign language learning: a psycholinguistic analysis of the issues.* Rowley, MA: Newbury House.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485-505. doi: 10.1177/0265532209340186

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.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217. doi: 10.1177/0265532208101010

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

Kramsch, C. (2006). From Communicative Competence to Symbolic Competence. *The Modern Language Journal, 90*(2), 249-252. doi: 10.1111/j.1540-4781.2006.00395_3.x

Lado, R. (1961). Language testing: the construction and use of foreign language tests: A teacher's book. New York, N.Y.: McGraw-Hill.

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. (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.

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

Magnan, S. (1988). Grammar and the ACTFL oral proficiency interview: Discussion and data. *The Modern Language Journal, 72*(3), 266-276.

Malmberg, P. (2000). *I huvudet på en elev. Projektet STRIMS. Strategier vid inlärning av moderna språk. [In the head of a student. The STRIMS project. Strategies when learning modern languages]*. Stockholm: Bonnier utbildning.

May, L. (2000). Assessment of oral proficiency in EAP programs: A case for pair interaction. *Language and Communication Review, 9*, 13-19.

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* (Vol. 24). 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

McNamara, T. F. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-76. doi: 10.1177/026553229000700105

McNamara, T. F. (1995). Modelling Performance: Opening Pandora's Box. *Applied Linguistics, 16*(2), 159-179. doi: 10.1093/applin/16.2.159

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

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

Meiron, B. E. (1998). *Rating oral proficiency tests: A triangulated study of rater thought processes*. Paper presented at the Language Testing Research Colloquium, Monterey, CA.

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

Messick, S. A. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256. doi: 10.1177/026553229601300302

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.

Nattinger, J., & DeCarrico, J. (1992). *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

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

Oller, J. W., Jr. (1973). Discrete-Point Tests Versus Tests of Integrative Skills. In J. W. Oller Jr. & J. C. Richardson (Eds.), *Focus on the Learner: Pragmatic Perspectives for the Language Teacher* (pp. 184-199). Rowley, Massachusetts: Newbury House Publishers.

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

Papajohn, D. (2002). Concept Mapping for Rater Training. *TESOL Quarterly, 36*(2), 219-233. doi: 10.2307/3588333

Perret, G. (1990). The language testing interview: A reappraisal *Individualising the assessment of language abilities* (pp. 225-228). Philadelphia, PA: Multilingual Matters.

Pollitt, A., & Murray, N. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3: Performance Testing, Cognition and Assessment* (pp. 74-91). Cambridge: UCLES/Cambridge University Press.

Purpura, J. E. (2008). Assessing Communicative Language Ability: Models and their Components. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education* (2 ed., Vol. 7: Language Testing and Assessment, pp. 53-68). New York: Springer.

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In B. A. Elder C, Grove E, Hill K, Iwashita N, Lumley T, McNamara T, O'Loughlin K (Ed.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 82-96). Cambridge: Cambridge University Press.

Saussure, F. d. (1959). *Course in general linguistics.* New York: Philosophical Library.

Savignon, S. J. (1972). *Communicative Competence: An Experiment in Foreign-Language Teaching.* Philadelphia: Center for Curriculum Development.

Shohamy, E. (1983a). Interrater and intrarater reliability of the oral interview and concurrent validity with cloze procedures in Hebrew. In J. W. Oller (Ed.), *Issues in Language Testing Research* (pp. 229-236). Rowley, MA: Newbury House.

Shohamy, E. (1983b). The stability of oral proficiency assessment in the oral interview procedure. *Language Learning, 33*(4), 527-540. doi: 10.1111/j.1467-1770.1983.tb00947.x

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

Sundh, S. (2003). Swedish School-leaving Students' Oral Proficiency in English: Grading of Production and Analysis of Performance *Acta Universitatis Upsaliensia 123*. Uppsala: Almqvist & Wiksell

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275-302. doi: 10.1177/026553220101800302

Taylor, L. (2000). Investigating the paired speaking test format. *Cambridge ESOL Research Notes, 2*, 14-15.

The National Assessment Project. (2014). Engelska 6 - Exempel på uppgiftstyper [English 6 - Examples of tasks]. Retrieved 26 November, 2014, from http://www.nafs.gu.se/prov_engelska/exempel_provuppgifter/engelska_5_exempeluppg/

The Swedish National Agency for Education. (2011). Ämne - Engelska [Subject - English]. Retrieved 26 November, 2014, from http://www.skolverket.se/laroplaner-amnen-och-kurser/gymnasieutbildning/gymnasieskola/eng?tos=gy&subjectCode=ENG&lang=sv&courseCode=ENGENG05 - anchor_ENGENG05

The Swedish Schools Inspectorate. (2012). Lika för alla? Omrättning av nationella prov i grundskolan och gymnasieskolan under tre år [Remarking of national tests for comprehensive school and upper secondary education]. Retrieved from: http://www.skolinspektionen.se/sv/Beslut-och-rapporter/Publikationer/

Vygotsky, L. S. (1986). *Thought and language*. Cambridge, MA: MIT Press.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223. doi: 10.1177/026553229401100206

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178. doi: http://dx.doi.org/10.1016/S1075-2935(00)00010-6

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

Wigglesworth, G. (2008). Task and Performance Based Assessment. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of Language and Education* (2 ed., Vol. 7: Language Testing and Assessment, pp. 111-122). New York: Springer.

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.

Young, R. F. (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, Canada.

List of appendices

Appendix 1: Rater background variables for Swedish raters

Table: Rater background variables for Swedish raters (n = 17)

Rater	Gender	Teaching experience/years
Rater 1	F	<10
Rater 2	M	>10
Rater 3	M	<10
Rater 4	F	>10
Rater 5	F	>10
Rater 6	F	10
Rater 7	F	>10
Rater 8	F	>10
Rater 9	F	<10
Rater 10	F	<10
Rater 11	F	<10
Rater 12	M	>10
Rater 13	F	>10
Rater 14	M	<10
Rater 15	F	>10
Rater 16	F	<10
Rater 17	F	>10

Appendix 2: Performance standards for course English 6 in Swedish upper secondary school

Grade E

In oral and written communications of various genres, students can express themselves in relatively varied ways, relatively clearly and relatively coherently. Students can express themselves with some fluency and to some extent adapted to purpose, recipient and situation. Students work on and make improvements to their own communications.

In oral and written interaction in various, and more formal contexts, students can express themselves clearly and with some fluency and some adaptation to purpose, recipient and situation. In addition, students can choose and use essentially functional strategies which to some extent solve problems and improve their interaction.

Grade C

In oral and written communications of various genres, students can express themselves in a way that is relatively varied, clear, coherent and relatively structured. Students can also express themselves with fluency and some adaptation to purpose, recipient and situation. Students work on and make well grounded improvements to their own communications.

In oral and written interaction in various, and more formal contexts, students can express themselves clearly with fluency, and with some adaptation to purpose, recipient and situation. In addition, students can choose and use functional strategies to solve problems and improve their interaction.

Grade A

In oral and written communications of various genres, students can express themselves in ways that are varied, clear, coherent and structured. Students can also express themselves with fluency and some adaptation to purpose, recipient and situation. Students work on and make well grounded and balanced improvements to their own communications.

In oral and written interaction in various, and more formal contexts, students express themselves clearly, relative freely and with fluency, and also with adaptation to purpose, recipient and situation. In addition, students can choose and use wellfunctioning strategies to solve problems and improve their interaction, and take it forward in a constructive way.

Source: The Swedish National Agency for Education (2011)

Appendix 3: Assessment factors provided in Teacher Guidelines for the national test for course English 6 in Swedish upper secondary school

Innehåll (Content)

- tydlighet *(clarity)*
- fyllighet och variation *(complexity and variation)*
    - Olika exempel och perspektiv *(different examples and perspectives)*
- sammanhang och struktur *(coherence and cohesion, structure)*
- anpassning till syfte, mottagare, situation och genre *(adaption to purpose, recipient, situation and genre)*

Språk och uttrycksförmåga *(Language and ability to express oneself)*

- kommunikativa strategier *(communicative strategies)*
    - för att utveckla och föra samtal vidare *(to develop and advance the conversation)*
    - för att lösa språkliga problem genom t.ex. omformuleringar, förklaringar och förtydliganden *(to solve linguistic problems by e.g. rephrasing, explaining and clarifying)*
- flyt och ledighet *(fluency and ease)*
- omfång, variation, komplexitet, tydlighet och säkerhet *(range, variation, complexity, clarity and accuracy)*
    - vokabulär, fraseologi och idiomatic *(vocabulary, phraseology and idiomatic expressions)*
    - uttal och intonation *(pronunciation and intonation)*
    - grammatiska strukturer *(grammatical structures)*
- anpassning till syfte, mottagare, situation och genre *(adaption to purpose, recipient, situation and genre)*

Source: The National Assessment Project (2014)

Appendix 4: Scales from the Manual for Relating Language Examinations to the CEFR (Council of Europe, 2009) used by the CEFR raters

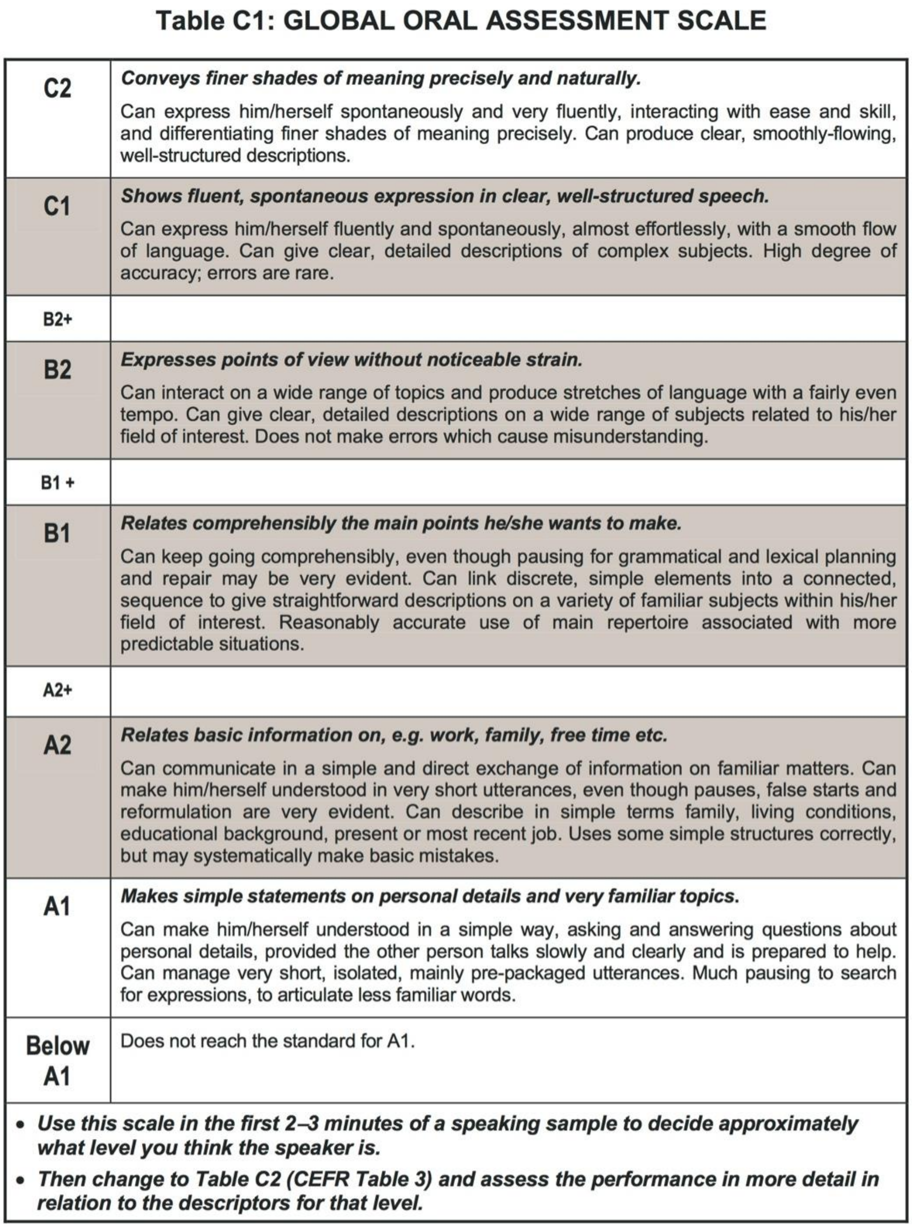


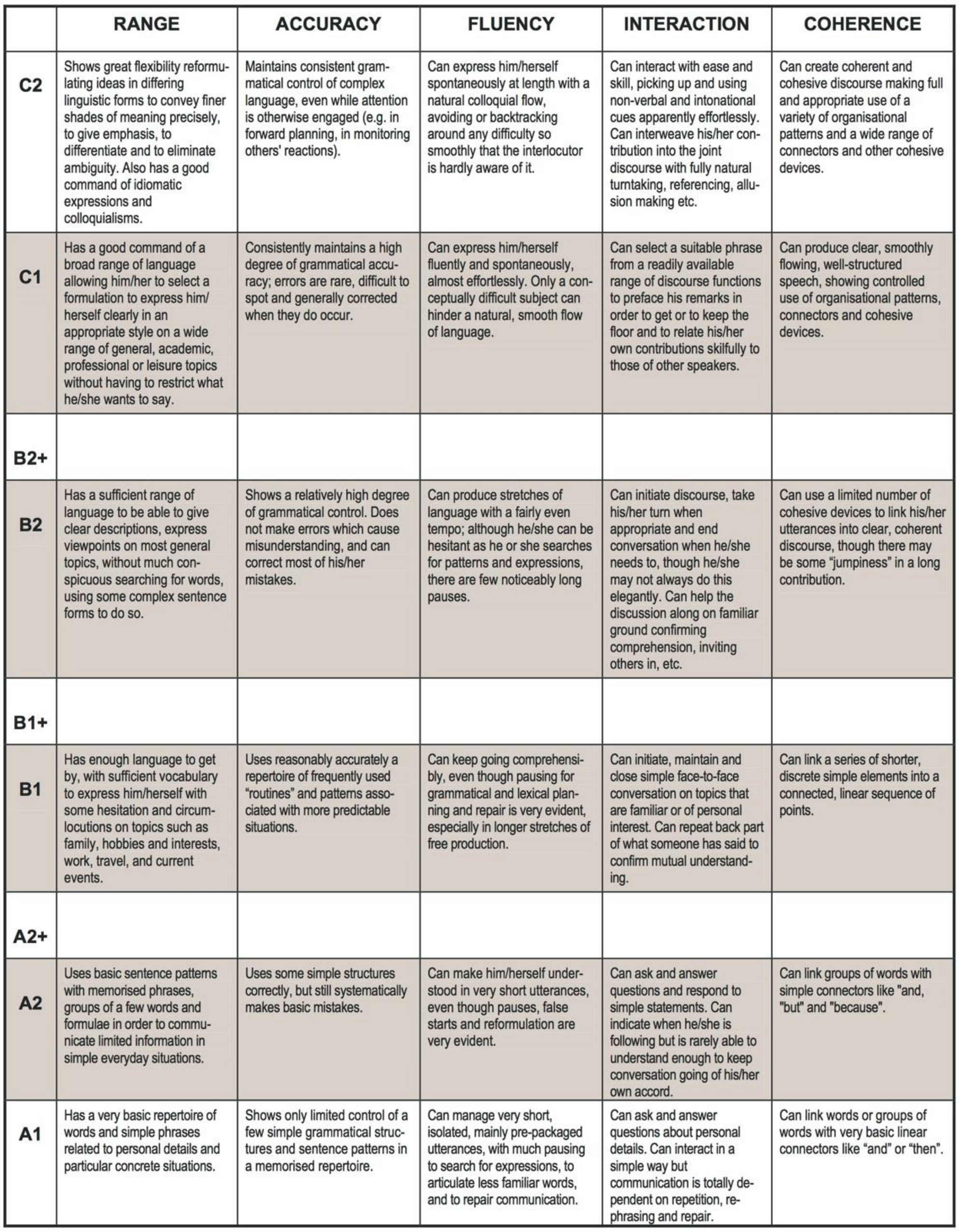
Table C1: GLOBAL ORAL ASSESSMENT SCALE

C2	***Conveys finer shades of meaning precisely and naturally.*** Can express him/herself spontaneously and very fluently, interacting with ease and skill, and differentiating finer shades of meaning precisely. Can produce clear, smoothly-flowing, well-structured descriptions.
C1	***Shows fluent, spontaneous expression in clear, well-structured speech.*** Can express him/herself fluently and spontaneously, almost effortlessly, with a smooth flow of language. Can give clear, detailed descriptions of complex subjects. High degree of accuracy; errors are rare.
B2+	
B2	***Expresses points of view without noticeable strain.*** Can interact on a wide range of topics and produce stretches of language with a fairly even tempo. Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest. Does not make errors which cause misunderstanding.
B1 +	
B1	***Relates comprehensibly the main points he/she wants to make.*** Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair may be very evident. Can link discrete, simple elements into a connected, sequence to give straightforward descriptions on a variety of familiar subjects within his/her field of interest. Reasonably accurate use of main repertoire associated with more predictable situations.
A2+	
A2	***Relates basic information on, e.g. work, family, free time etc.*** Can communicate in a simple and direct exchange of information on familiar matters. Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can describe in simple terms family, living conditions, educational background, present or most recent job. Uses some simple structures correctly, but may systematically make basic mistakes.
A1	***Makes simple statements on personal details and very familiar topics.*** Can make him/herself understood in a simple way, asking and answering questions about personal details, provided the other person talks slowly and clearly and is prepared to help. Can manage very short, isolated, mainly pre-packaged utterances. Much pausing to search for expressions, to articulate less familiar words.
Below A1	Does not reach the standard for A1.
• ***Use this scale in the first 2–3 minutes of a speaking sample to decide approximately what level you think the speaker is.*** • ***Then change to Table C2 (CEFR Table 3) and assess the performance in more detail in relation to the descriptors for that level.***	

Source: Council of Europe, 2009, p. 184

Appendix 4 (continued):

Table C2: ORAL ASSESSMENT CRITERIA GRID (CEFR Table 3)



	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/ herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.
B2+					
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.
B1+					
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
A2+					
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can ask and answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connectors like "and, "but" and "because".
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like "and" or "then".

Source: Council of Europe, 2009, p. 185

Appendix 4 (continued):

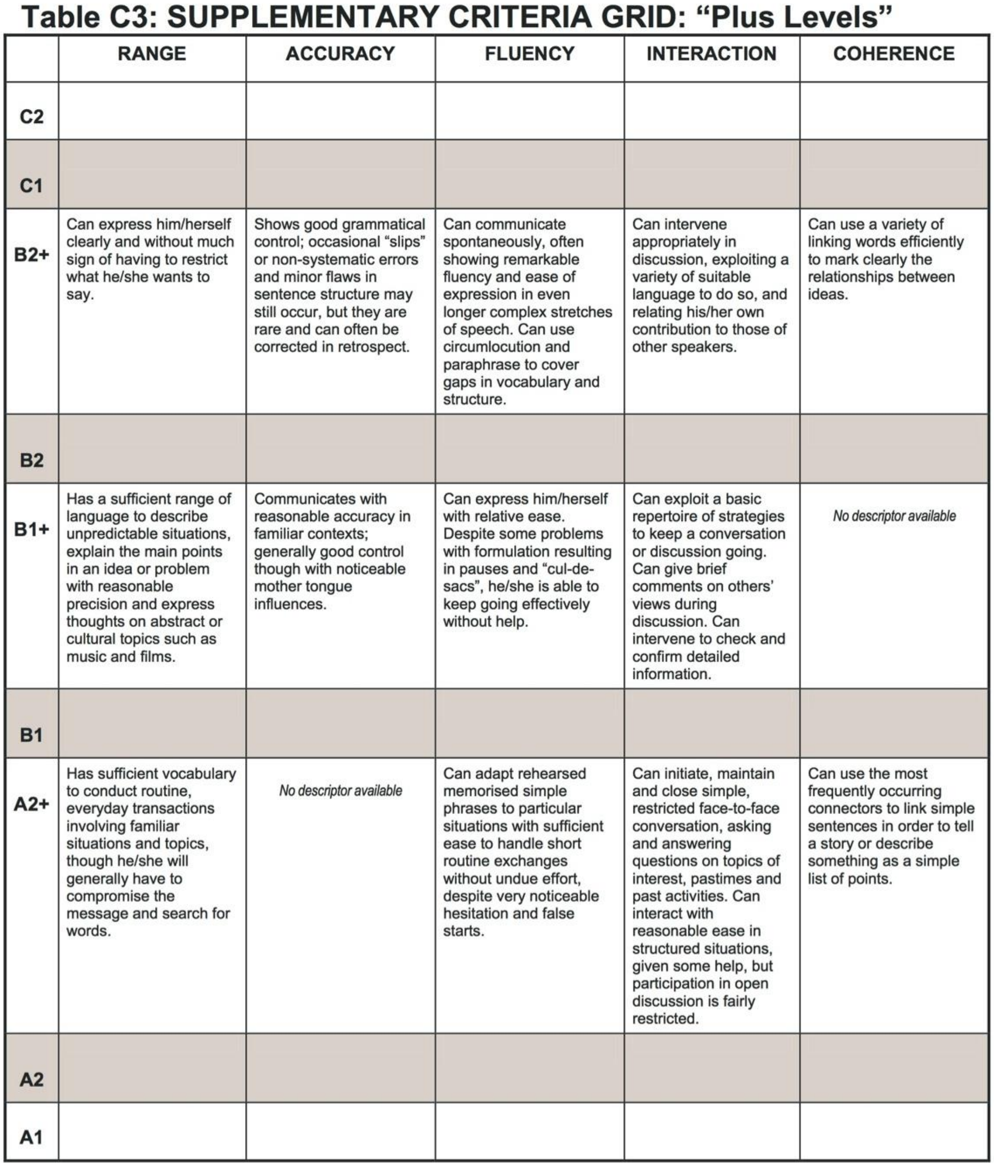


Table C3: SUPPLEMENTARY CRITERIA GRID: "Plus Levels"

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
C2					
C1					
B2+	Can express him/herself clearly and without much sign of having to restrict what he/she wants to say.	Shows good grammatical control; occasional "slips" or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech. Can use circumlocution and paraphrase to cover gaps in vocabulary and structure.	Can intervene appropriately in discussion, exploiting a variety of suitable language to do so, and relating his/her own contribution to those of other speakers.	Can use a variety of linking words efficiently to mark clearly the relationships between ideas.
B2					
B1+	Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.	Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influences.	Can express him/herself with relative ease. Despite some problems with formulation resulting in pauses and "cul-de-sacs", he/she is able to keep going effectively without help.	Can exploit a basic repertoire of strategies to keep a conversation or discussion going. Can give brief comments on others' views during discussion. Can intervene to check and confirm detailed information.	*No descriptor available*
B1					
A2+	Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics, though he/she will generally have to compromise the message and search for words.	*No descriptor available*	Can adapt rehearsed memorised simple phrases to particular situations with sufficient ease to handle short routine exchanges without undue effort, despite very noticeable hesitation and false starts.	Can initiate, maintain and close simple, restricted face-to-face conversation, asking and answering questions on topics of interest, pastimes and past activities. Can interact with reasonable ease in structured situations, given some help, but participation in open discussion is fairly restricted.	Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points.
A2					
A1					

Source: Council of Europe, 2009, p. 186

Appendix 5: Verbatim quotations in Swedish

Extract 13. Hon använder också en del uttryck/ord fel (learn to handle with money). (negative) /Sw

Extract 33 Eleven stannar ofta upp och detta i kombination med uttalet gör att flytet uteblir. (negative) /Sw

Extract 50 Eleven har dock en tendens att ta över samtalet och släpper inte in sin partner i samtalet. Hon ger inte sin partner tid att tänka när han t.ex. inte hittar orden vilket stressar honom och gör att hon tar över ännu mer – detta är något som drar ner betyget något då det blir mer monolog än dialog ibland. (negative) /Sw

Extract 58: Hade det inte varit för henne så hade han haft svårt för att klara av den här uppgiften, men hon ställde bra frågor till honom som fick honom att tänka till. /Sw

Extract 63. Hon har väldigt bråttom, och upprepar mycket vilket får henne att kännas som osäker. (negative) /Sw

Extract 69 och han använder strategier när han inte hittar orden, han förklarar t.ex. vad han menar. (positive) /Sw

Extract 73: Eleven har ett relativt gott ordförråd och några riktigt bra formuleringar (comfort zone, interpret, appreciate the littel things, life experience, hard to settle down). (positive) /Sw

Extract 82: Han diskuterar dock, och kommer med sin åsikt, kring det hans partner pratat om. (positive) /Sw

Extract 84: Använder ordet ”crap” vilket inte hör hemma i sammanhanget – han ber dock om ursäkt för detta, så han är medveten om det. (mixed) /Sw

Extract 89: Hon berättar mycket kort om sitt kort, därför blir också diskussionen kortfattad. (negative) /Sw

Extract 97 Han följer instruktionerna för uppgiften och det känns som att han har en tydlig bild över vad han vill säga (även om det blev tyst i början). (positive) /Sw

Extract 99 Jag tror att hon bidrar till att hennes partner får ett högre betyg än vad han har presterat tidigare för hon anpassar sitt språk och ställer bra frågor. /Sw

Extract 101 Det är också svårt för honom att komma in i samtalet ibland eftersom han blir avbruten flera gånger. Kanske hade han kunnat visa mer med en annan samtalspartner men det vet vi inte. /Sw

Extract 103 Hans språk är inte det bästa, och inte heller hans uttal. Men han förtjänar ett högre betyg med tanke på innehållet. /Sw

Extract 104 Varför hon får E och inte han är för att hon har bättre idéer och följer instruktionerna på ett bättre sätt än vad han gör. /Sw

Extract 109 Till en början trodde jag att hon lyssnade aktivt och var intresserad av vad han sa, men märkte efter ett tag att hon upprepade allt han sa, och att hon inte hade så många egna tankar kring det som diskuterades. Hon avbryter till viss del samtalet med sina ”yes”, ”I think so too” och ”yeah”. /Sw

Extract 122 Hon ställer bra och relevanta frågor till sin partner. Det för samtalet vidare och det bidrar till intressanta diskussioner. Det är bra interaktion i deras par, och hon bidrar till stor det till det. (positive) /Sw

Extract 127 Flytet störs ibland av att hon inte hittar orden, detta blir dock bättre i del två och när hon tänker fritt. (mixed) /Sw

Appendix 6: Written instructions to CEFR raters

You have received a CD with six recorded conversations (one female student and one male student in each conversation). You are going to listen to one conversation at a time with stops and repetition where needed. While you are listening you are asked to take notes on the piece of paper that is provided. I would like you to take notes freely in order to capture your thoughts as you are assessing. For that reason you do not have to write complete sentences, but rather just jot down your thoughts. Please note as many aspects as possible that you pay attention to while listening and forming your judgment.

After listening to each conversation I would like you to fill in an assessment form for each student. The assessment form is available on the memory stick. In the assessment form you are asked to fill in your score and also explain it by writing a summary comment about the performance. In other words, I would like you to explain what qualities and aspects of the oral performance you attended to in making your decision. Finally, please save the document on your memory stick. Table C1, C2 and C3 from the CEFR Manual are provided.

You will use your notes when we have the group discussion to help you remember. Both the notes and the assessment forms you fill in are part of the research material.

Step by step summary

1) Listen to each conversation and take notes by hand
2) Fill in assessment form (on memory stick) for each student with rating and summary comment
3) Save the document on the memory stick

Appendix 7: Coding scheme

CRITERION FEATURES

COMMUNICATIVE LANGUAGE COMPETENCES (AS DESCRIBED IN THE CEFR)

- **LINGUISTIC (Range and Accuracy)**
- **PRAGMATIC (Fluency and Coherence)**
- **SOCIOLINGUISTIC**

RANGE

RA: GLR	general linguistic range (range mentioned in general)
RA: VOC	vocabulary range
RA: EXP	ability to express viewpoints

ACCURACY

AC: GRA	grammatical accuracy
AC:VC	vocabulary control
AC:PC	phonological control

FLUENCY

FL:FLU	fluency – mentioned in general
FL: SPE	speed of delivery – fast/slow
FL:HES	hesitation and pauses

COHERENCE

CO:CC	coherence and cohesion
CO: TOD	topic development, complexity of ideas
CO: FC	flexibility to circumstances

SOCIOLINGUISTIC APPROPRIATENESS

SL:SA	sociolinguistic appropriateness

COMMUNICATION STRATEGIES (Interaction and production strategies)

INTERACTION

IN:TT	turntaking
IN:COOP	cooperating
IN: DOM	dominates the discussion – usually mentioned negatively
IN: MAN	manages/controls interaction – usually mentioned positively
IN:HEL	helps partner out
IN: PAS	has a passive role in the conversation

PRODUCTION STRATEGIES

PS:CS	compensating
PS:MR	monitoring and repair

Appendix 7: Coding scheme (continued)

FEATURES NOT EXPLICITLY STATED IN THE CRITERIA

INTELLIGIBILITY

IB	intelligibility to rater

TASK REALISATION

TR: LEN	length of response - extended or very brief discourse by candidate
TR: COT	completing and understanding the task
TR: OV	comments on the overall performance
TR: ST	summary of text (how well the candidate summarises the text)

OTHER coded comment that does not fit any of the above categories

RATER REFLECTION

RR:REF	rater reflection in general
RR:DEC	rater reflection about rating decision
RR:MAT	matching of candidates – how candidates perform in relation to each other

EVALUATIVE RESPONSE OF RATER

Pos	Positive
Neg	Negative
Mix	Mixed

FOCUS OF RESPONSE

✓ Inter-candidate comparison, finding similarities (ICCS)
✓ Inter-candidate contrast, finding differences (ICCD)
✓ Inter-candidate comparison, aspect to do with interaction strategies (ICCI)
✓ Intra-candidate comparison of an aspect of candidate's performance over time (ICCT)
✓ Comparison with other pairs (COMP)
✓ Refers to /uses a candidate's exact words (RCW)

Appendix 8: Scaled descriptors for sociolinguistic appropriateness and flexibility

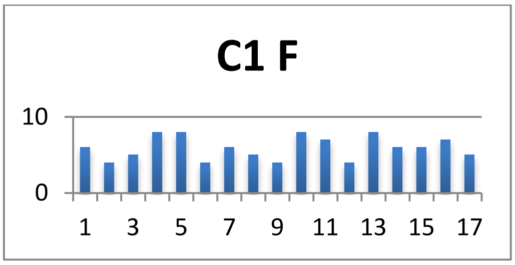
SOCIOLINGUISTIC APPROPRIATENESS
C2
Has a good command of idiomatic expressions and colloquialisms with awareness of connotative levels of meaning. Appreciates fully the sociolinguistic and sociocultural implications of language used by native speakers and can react accordingly. Can mediate effectively between speakers of the target language and that of his/her community of origin taking account of sociocultural and sociolinguistic differences.
C1
Can recognise a wide range of idiomatic expressions and colloquialisms, appreciating register shifts; may, however, need to confirm occasional details, especially if the accent is unfamiliar. Can follow films employing a considerable degree of slang and idiomatic usage. Can use language flexibly and effectively for social purposes, including emotional, allusive and joking usage.
B2+
Can express him or herself confidently, clearly and politely in a formal or informal register, appropriate to the situation and person(s) concerned.
B2
Can with some effort keep up with and contribute to group discussions even when speech is fast and colloquial. Can sustain relationships with native speakers without unintentionally amusing or irritating them or requiring them to behave other than they would with a native speaker. Can express him or herself appropriately in situations and avoid crass errors of formulation. Can perform and respond to a wide range of language functions, using their most common exponents in a neutral register.
B1
Is aware of the salient politeness conventions and acts appropriately. Is aware of, and looks out for signs of, the most significant differences between the customs, usages, attitudes, values and beliefs prevalent in the community concerned and those of his or her own.
A2+
Can perform and respond to basic language functions, such as information exchange and requests and express opinions and attitudes in a simple way. Can socialise

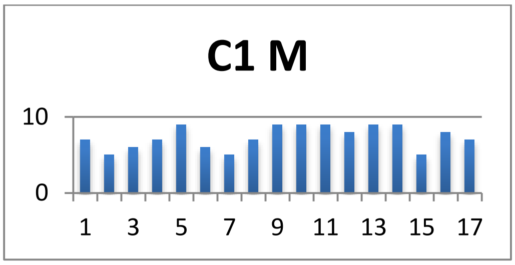
simply but effectively using the simplest common expressions and following basic routines.
A2
Can handle very short social exchanges, using everyday polite forms of greeting and address. Can make and respond to invitations, suggestions, apologies, etc.
A1
Can establish basic social contact by using the simplest everyday polite forms of: greetings and farewells; introductions; saying please, thank you, sorry, etc.

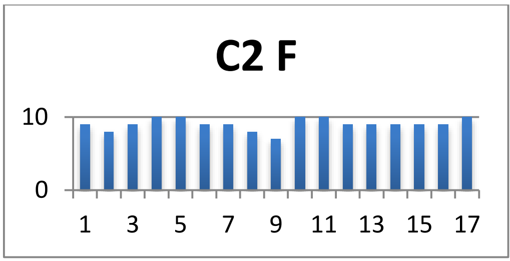
FLEXIBILITY
C2
Shows great flexibility reformulating ideas in differing linguistic forms to give emphasis, to differentiate according to the situation, interlocutor, etc. and to eliminate ambiguity.
C1 As B2+
B2+
Can adjust what he/she says and the means of expressing it to the situation and the recipient and adopt a level of formality appropriate to the circumstances.
B2
Can adjust to the changes of direction, style and emphasis normally found in conversation. Can vary formulation of what he/she wants to say.
B1+
Can adapt his/her expression to deal with less routine, even difficult, situations.
B1
Can exploit a wide range of simple language flexibly to express much of what he/she wants.
A2+
Can adapt well rehearsed memorised simple phrases to particular circumstances through limited lexical substitution.
A2
Can expand learned phrases through simple recombinations of their elements.
A1
No descriptor available

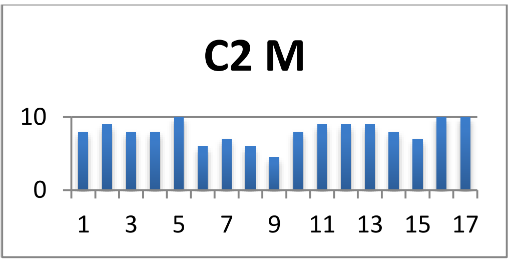
Source: Council of Europe, 2001, p. 122 and 124

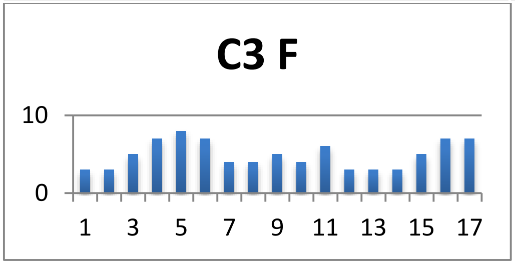
Appendix 9: Distribution of scores per candidate for Swedish raters

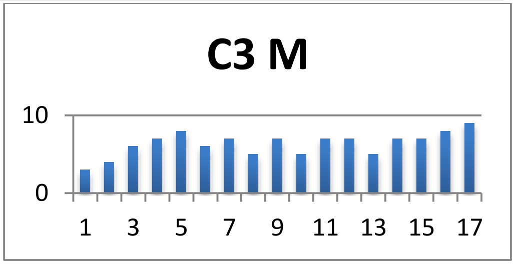


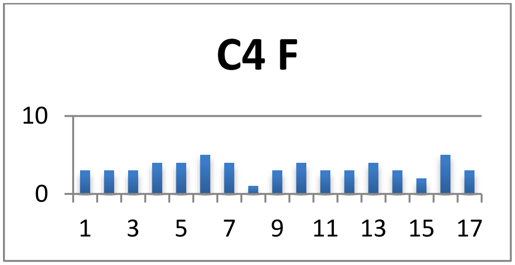


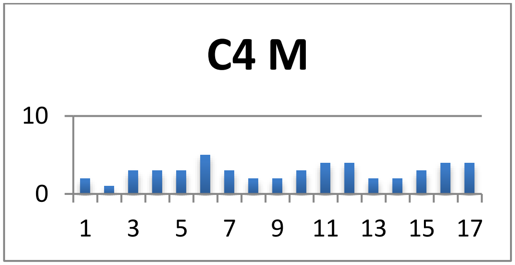


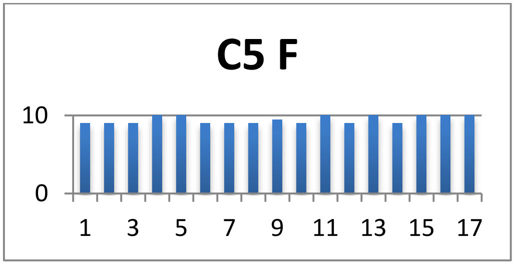


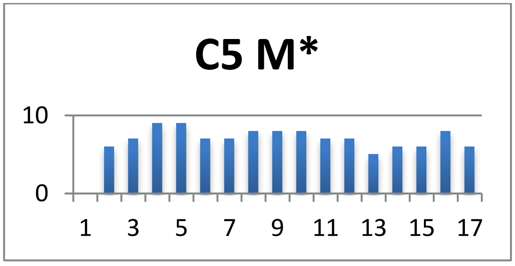




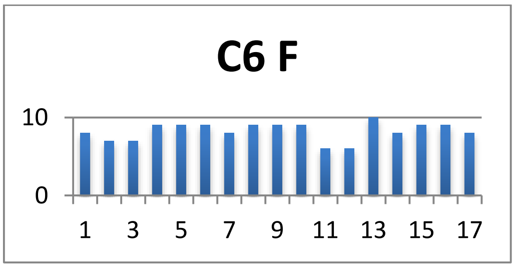


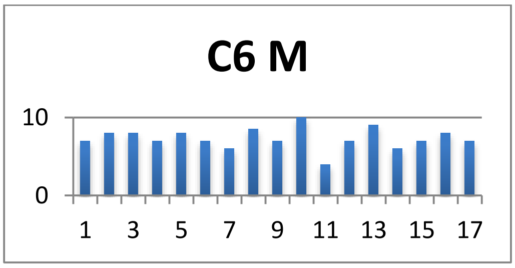






**Rater 1 did not award a score for C5M.*





Appendix 10: Correlations between Swedish raters using Kendall's Tau and Spearmans's rho

Kendall's tau_b	R1	Correlation Coefficient	1,000	,840**	,828**	,834**	,873**	,583*	,780**	,726**	,580*	,735**	,626*	,673**	,845**	,776**	,727**	,816**	,673**
		Sig. (2-tailed)		,001	,001	,001	,001	,022	,002	,003	,019	,004	,012	,007	,001	,002	,004	,001	,007
		N	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11
	R2	Correlation Coefficient	,840**	1,000	,885**	,656**	,795**	,509*	,705**	,704**	,468*	,700**	,496*	,700**	,750**	,650**	,689**	,875**	,639**
		Sig. (2-tailed)	,001		,000	,006	,001	,033	,003	,002	,042	,003	,035	,003	,001	,006	,003	,000	,006
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R3	Correlation Coefficient	,828**	,885**	1,000	,684**	,791**	,638**	,750**	,732**	,525*	,712**	,571*	,729**	,661**	,661**	,783**	,838**	,717**
		Sig. (2-tailed)	,001	,000		,004	,001	,008	,001	,002	,024	,003	,016	,002	,005	,006	,001	,000	,002
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R4	Correlation Coefficient	,834**	,656**	,684**	1,000	,793**	,602*	,855**	,634**	,505*	,591*	,604*	,504*	,591*	,626**	,701**	,684**	,530*
		Sig. (2-tailed)	,001	,006	,004		,001	,014	,000	,007	,033	,014	,012	,037	,014	,010	,003	,005	,026
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R5	Correlation Coefficient	,873**	,795**	,791**	,793**	1,000	,546*	,738**	,583*	,588*	,590*	,762**	,750**	,661**	,786**	,615*	,883**	,703**
		Sig. (2-tailed)	,001	,001	,001	,001		,028	,002	,015	,014	,016	,002	,002	,007	,001	,011	,000	,004
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R6	Correlation Coefficient	,583*	,509*	,638**	,602*	,546*	1,000	,587*	,673**	,611*	,509*	,296	,351	,491*	,421	,621**	,566*	,518*
		Sig. (2-tailed)	,022	,033	,008	,014	,028		,015	,004	,010	,037	,222	,150	,043	,084	,010	,020	,031
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R7	Correlation Coefficient	,780**	,705**	,750**	,855**	,738**	,587*	1,000	,634**	,557*	,576*	,555*	,559*	,627**	,729**	,850**	,787**	,650**
		Sig. (2-tailed)	,002	,003	,001	,000	,002	,015		,006	,017	,016	,019	,019	,008	,002	,000	,001	,006
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R8	Correlation Coefficient	,726**	,704**	,732**	,634**	,583*	,673**	,634**	1,000	,688**	,728**	,443	,480*	,761**	,579*	,732**	,668**	,504*
		Sig. (2-tailed)	,003	,002	,002	,007	,015	,004	,006		,003	,002	,058	,041	,001	,013	,002	,005	,029
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R9	Correlation Coefficient	,580*	,468*	,525*	,505*	,588*	,611*	,557*	,688**	1,000	,534*	,413	,417	,617**	,634**	,508*	,606*	,426
		Sig. (2-tailed)	,019	,042	,024	,033	,014	,010	,017	,003		,023	,078	,076	,009	,007	,029	,010	,067
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R10	Correlation Coefficient	,735**	,700**	,712**	,591*	,590*	,509*	,576*	,728**	,534*	1,000	,444	,534*	,759**	,672**	,559*	,574*	,441
		Sig. (2-tailed)	,004	,003	,003	,014	,016	,037	,016	,002	,023		,064	,026	,002	,005	,019	,017	,064
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R11	Correlation Coefficient	,626*	,496*	,571*	,604*	,762**	,296	,555*	,443	,413	,444	1,000	,735**	,496*	,735**	,504*	,604*	,639**
		Sig. (2-tailed)	,012	,035	,016	,012	,002	,222	,019	,058	,078	,064		,002	,038	,002	,033	,012	,007
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R12	Correlation Coefficient	,673**	,700**	,729**	,504*	,750**	,351	,559*	,480*	,417	,534*	,735**	1,000	,552*	,707**	,559*	,713**	,695**
		Sig. (2-tailed)	,007	,003	,002	,037	,002	,150	,019	,041	,076	,026	,002		,021	,003	,019	,003	,003
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R13	Correlation Coefficient	,845**	,750**	,661**	,591*	,661**	,491*	,627**	,761**	,617**	,759**	,496*	,552*	1,000	,741**	,695**	,748**	,542*
		Sig. (2-tailed)	,001	,001	,005	,014	,007	,043	,008	,001	,009	,002	,038	,021		,002	,003	,002	,022
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R14	Correlation Coefficient	,776**	,650**	,661**	,626**	,786**	,421	,729**	,579*	,634**	,672**	,735**	,707**	,741**	1,000	,661**	,765**	,661**
		Sig. (2-tailed)	,002	,006	,006	,010	,001	,084	,002	,013	,007	,005	,002	,003	,002		,006	,002	,005
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R15	Correlation Coefficient	,727**	,689**	,783**	,701**	,615*	,621**	,850**	,732**	,508*	,559*	,504*	,559*	,695**	,661**	1,000	,769**	,717**
		Sig. (2-tailed)	,004	,003	,001	,003	,011	,010	,000	,002	,029	,019	,033	,019	,003	,006		,001	,002
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R16	Correlation Coefficient	,816**	,875**	,838**	,684**	,883**	,566*	,787**	,668**	,606*	,574*	,604*	,713**	,748**	,765**	,769**	1,000	,804**
		Sig. (2-tailed)	,001	,000	,000	,005	,000	,020	,001	,005	,010	,017	,012	,003	,002	,002	,001		,001
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R17	Correlation Coefficient	,673**	,639**	,717**	,530*	,703**	,518*	,650**	,504*	,426	,441	,639**	,695**	,542*	,661**	,717**	,804**	1,000
		Sig. (2-tailed)	,007	,006	,002	,026	,004	,031	,006	,029	,067	,064	,007	,003	,022	,005	,002	,001	
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12

Appendix 10: Correlations between Swedish raters using Kendall's Tau and Spearmans's rho (continued)

Spearman's rho	R1	Correlation Coefficient	1,000	,918**	,914**	,900**	,931**	,712*	,862**	,859**	,682*	,833**	,724*	,800**	,915**	,866**	,836**	,901**	,776**
		Sig. (2-tailed)		,000	,000	,000	,000	,014	,001	,001	,021	,001	,012	,003	,000	,001	,001	,000	,005
		N	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11
	R2	Correlation Coefficient	,918**	1,000	,954**	,751**	,866**	,629*	,806**	,821**	,593*	,792**	,634*	,848**	,862**	,758**	,829**	,935**	,779**
		Sig. (2-tailed)	,000		,000	,005	,000	,028	,002	,001	,042	,002	,027	,000	,000	,004	,001	,000	,003
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R3	Correlation Coefficient	,914**	,954**	1,000	,796**	,862**	,767**	,859**	,845**	,649*	,820**	,674*	,863**	,794**	,763**	,880**	,911**	,832**
		Sig. (2-tailed)	,000	,000		,002	,000	,004	,000	,001	,022	,001	,016	,000	,002	,004	,000	,000	,001
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R4	Correlation Coefficient	,900**	,751**	,796**	1,000	,858**	,703*	,911**	,763**	,628*	,647*	,731**	,623*	,706*	,701*	,815**	,788**	,657*
		Sig. (2-tailed)	,000	,005	,002		,000	,011	,000	,004	,029	,023	,007	,030	,010	,011	,001	,002	,020
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R5	Correlation Coefficient	,931**	,866**	,862**	,858**	1,000	,654*	,825**	,712**	,656*	,666*	,861**	,853**	,768**	,870**	,739**	,928**	,821**
		Sig. (2-tailed)	,000	,000	,000	,000		,021	,001	,009	,021	,018	,000	,000	,004	,000	,006	,000	,001
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R6	Correlation Coefficient	,712*	,629*	,767**	,703*	,654*	1,000	,712**	,799**	,747**	,636*	,387	,474	,607*	,545	,743**	,689*	,664*
		Sig. (2-tailed)	,014	,028	,004	,011	,021		,009	,002	,005	,026	,214	,120	,036	,067	,006	,013	,018
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R7	Correlation Coefficient	,862**	,806**	,859**	,911**	,825**	,712**	1,000	,781**	,657*	,667*	,697*	,726**	,761**	,775**	,934**	,879**	,797**
		Sig. (2-tailed)	,001	,002	,000	,000	,001	,009		,003	,020	,018	,012	,007	,004	,003	,000	,000	,002
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R8	Correlation Coefficient	,859**	,821**	,845**	,763**	,712**	,799**	,781**	1,000	,846**	,857**	,465	,639*	,874**	,696*	,829**	,788**	,599*
		Sig. (2-tailed)	,001	,001	,001	,004	,009	,002	,003		,001	,000	,127	,025	,000	,012	,001	,002	,040
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R9	Correlation Coefficient	,682*	,593*	,649*	,628*	,656*	,747**	,657*	,846**	1,000	,671*	,530	,564	,728**	,742**	,636*	,679*	,551
		Sig. (2-tailed)	,021	,042	,022	,029	,021	,005	,020	,001		,017	,077	,056	,007	,006	,026	,015	,063
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R10	Correlation Coefficient	,833**	,792**	,820**	,647*	,666*	,636*	,667*	,857**	,671*	1,000	,496	,674*	,855**	,754**	,714**	,687*	,556
		Sig. (2-tailed)	,001	,002	,001	,023	,018	,026	,018	,000	,017		,101	,016	,000	,005	,009	,014	,061
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R11	Correlation Coefficient	,724*	,634*	,674*	,731**	,861**	,387	,697*	,465	,530	,496	1,000	,834**	,558	,847**	,596*	,728**	,745**
		Sig. (2-tailed)	,012	,027	,016	,007	,000	,214	,012	,127	,077	,101		,001	,059	,001	,041	,007	,005
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R12	Correlation Coefficient	,800**	,848**	,863**	,623*	,853**	,474	,726**	,639*	,564	,674*	,834**	1,000	,691*	,851**	,692*	,852**	,801**
		Sig. (2-tailed)	,003	,000	,000	,030	,000	,120	,007	,025	,056	,016	,001		,013	,000	,013	,000	,002
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R13	Correlation Coefficient	,915**	,862**	,794**	,706*	,768**	,607*	,761**	,874**	,728**	,855**	,558	,691*	1,000	,847**	,814**	,856**	,681*
		Sig. (2-tailed)	,000	,000	,002	,010	,004	,036	,004	,000	,007	,000	,059	,013		,001	,001	,000	,015
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R14	Correlation Coefficient	,866**	,758**	,763**	,701*	,870**	,545	,775**	,696*	,742**	,754**	,847**	,851**	,847**	1,000	,724**	,851**	,794**
		Sig. (2-tailed)	,001	,004	,004	,011	,000	,067	,003	,012	,006	,005	,001	,000	,001		,008	,000	,002
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R15	Correlation Coefficient	,836**	,829**	,880**	,815**	,739**	,743**	,934**	,829**	,636*	,714**	,596*	,692*	,814**	,724**	1,000	,873**	,847**
		Sig. (2-tailed)	,001	,001	,000	,001	,006	,006	,000	,001	,026	,009	,041	,013	,001	,008		,000	,001
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R16	Correlation Coefficient	,901**	,935**	,911**	,788**	,928**	,689*	,879**	,788**	,679*	,687*	,728**	,852**	,856**	,851**	,873**	1,000	,904**
		Sig. (2-tailed)	,000	,000	,000	,002	,000	,013	,000	,002	,015	,014	,007	,000	,000	,000	,000		,000
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
	R17	Correlation Coefficient	,776**	,779**	,832**	,657*	,821**	,664*	,797**	,599*	,551	,556	,745**	,801**	,681*	,794**	,847**	,904**	1,000
		Sig. (2-tailed)	,005	,003	,001	,020	,001	,018	,002	,040	,063	,061	,005	,002	,015	,002	,001	,000	
		N	11	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Appendix 11: Relationship between comments and scores

Table 1. Comments on C3M

Score	**Comments**	**Score**	**Comments**
E- ; E		**C+;B**	
Range	"Basic vocabulary" "He tries to interact but **as the vocabulary isn't really as wide as necessary** there isn't much of a discussion and the topics are just briefly dealt with."		"He expresses himself in a varied way with a good and relatively broad vocabulary but there are some unidiomatic expressions" (Tr.) "Clear speaker with limited vocabulary in the beginning"
Pronunciation/ Fluency	"Intonation and pronunciation ok but needs practice and is influenced by Swedish."		"Throughout the conversation, he expresses himself with good fluency and good pronunciation" (Tr.)
Production strategies	"Gets stuck on a word and it takes some time for him to work around it."		"However, on occasion he gets stuck, and he has some difficulty paraphrasing and moving on" (Tr.) "Good strategies to discuss around a topic when faced with a tricky word or phrase (performance anxiety)." "Corrects himself often, showing an awareness of the mistakes he is making."
Interaction	"Interaction between the two is okay, they comment on each other but they could help each other more." **"He tries to interact** but as the vocabulary isn't really as wide as necessary **there isn't much of a discussion** and the topics are just briefly dealt with.		"He expresses his views and relates to his partner's contributions" (Tr.)

Appendix 11: (continued)

Table 2. Comments on C6M

Score	Comments	Score	Comments
E		**A**	
Accuracy	“grammar and phrasal errors” “makes a few language mistakes again in part two”		
Fluency/Coherence	“From time to time it is fluent but in other occasions he feels a bit unclear” “He has a lot of good examples that weigh up to his grade despite the errors he makes and despite the unclear parts”		“Structured and complex” “very “relaxed” and calm”
Interaction	“Interacts well with his partner”		“The speakers help each other well here, they give and take, ask for clarifications, examples”
Vocabulary			“Broad vocabulary, varied”

Appendix 11: (continued)

Table 3. Comments on C5F

Score	Comments	Score	Comments
A		**A**	
Accuracy	"Few grammar errors (verbs: people likes, I have chosed…)" "Good pronunciation and intonation"		"there are very few mistakes when it comes to expressions and grammar. The language and sentence structures are varied and quite advanced." "Correct pronunciation"
Fluency			"She speaks with very good fluency"
Coherence	"develops her line of thinking very well."		"She uses different examples (both from the card and /…/ which contributes to quite a few perspectives. The content is coherent and structured"
Interaction	"Really good interaction: nice nuanced discussion. Asks partner to develop or clarify. Invites partner." "Brings conversation forward (hogs the conversation a little bit, maybe)"		"She starts by commenting on the other speaker's comments. This is a smooth conversation but he tends to be a bit quiet and she gradually starts to take over the conversation." "she adapts to the male speaker by adding questions and comments throughout the session."
Vocabulary	"Varied and extensive vocabulary. Loads of nice idiomatic expressions and complex language structures."		"The language is varied and contains several idiomatic phrases/expressions"

Appendix 11: (continued)

Table 4. Comments on C6F

Score	Comments	Score	Comments
2 CEFR raters		**2 Sw raters**	
Accuracy	“However, she can at times show good command of structures, which makes her performance a bit irregular”		“she displays some very good language and is mostly correct” “Only a little unidiomatic on occasion, e.g. it’s benefit for beneficial, keep up it for keep it up.” “Very good pronunciation” “very good pronunciation” (Tr.)
Fluency	“She manages to put her message across all along, though she’s clearly finding it hard to show consistent fluency” “There are usually no major problems in getting the message across even though there are pauses and hesitation. However, the speech is not very coherent or fluent” “The speaker sometimes has difficulty in finding the correct way of expressing herself – from time to time there are longer pauses and hesitation.”		“lacks fluency at times” “needs to work on fluency” “/…/ even if the student occasionally gets stuck and can’t keep going” (Tr.) “Good fluency” (Tr.)
Coherence	“There are usually no major problems in getting the message across even though there are		“Also, some good discourse management, e.g. on the one hand.”

	pauses and hesitation. However, the speech is not very coherent or fluent." "and she clearly lacks /…/ connecting devices needed to express herself with ease."		"The content is well-developed and she gives plenty of examples to support her views. She summarises the discussion on one occasion allowing the discussion to continue in a constructive way. Good structure and coherence." (Tr.) "Uses connectors like "On the other hand" links different parts of the discussion." (Tr.)
Interaction			"She summarises the discussion on one occasion allowing the discussion to continue in a constructive way" (Tr.)
Range	"The speaker uses fairly simple vocabulary" "Her English is rather broken and she clearly lacks the vocabulary /…/ needed to express herself with ease."		"Overall good language" (Tr.) "Many idiomatic expressions and relatively formal language at the beginning at least." (Tr.)

Appendix 11: (continued)

Verbatim quotations in Swedish

These quotations are translated into English in Tables 1 and 4 in Appendix 11

Table 1

”Han formulerar sig varierat med gott och relativt brett ordförråd men några oidiomatiska uttryck förekommer”

“Han uttrycker sig genomgående med bra flyt och gott uttal”

“vid något tillfälle fastnar han dock och har vissa svårigheter att omformulera och ta sig vidare.”

“Han uttrycker åsikter och anknyter till partners inlägg

Table 4

“mycket bra uttal”

“även om eleven stundtals hakar upp sig och inte kommer vidare”

“Gott flyt”

“Innehållet är fylligt och hon ger gott om exempel för att stödja sina åsikter. Sammanfattar vid något tillfälle diskussionen så här långt vilket gör att diskussionen kan fortsätta på ett konstruktivt sätt. Bra struktur och tydliggjort sammanhang.”

“Sammanbindningsfraser som: ”On the other hand…” länkar ihop olika delar av diskussionen.”

“Sammanfattar vid något tillfälle diskussionen så här långt vilket gör att diskussionen kan fortsätta på ett konstruktivt sätt”

“Bra språk överlag”

“Många idiomatiska uttryck och relativt formellt språk i alla fall i början.”