Oral proficiency in the workplace context:
Computer-simulated language testing and self-assessment

Tiina Salo
University of Tampere
School of Language, Translation and Literacy Studies
English Philology
Pro Gradu Thesis
September 2014

Tampereen yliopisto

Englantilainen filologia
Kieli-, käännös- ja kirjallisuustieteiden yksikkö

SALO, TIINA: Oral proficiency in the workplace context: Computer-simulated language testing and self-assessment

Pro gradu -tutkielma, 77 sivua + 2 liitettä
syyskuu 2014

Tämä pro gradu -tutkielma käsittelee suullisen kielitaidon testaamista ja arvioimista työelämäkontekstissa. Tutkimuksessa pyritään selvittämään, miten suullisen kielitaidon testaamista voisi toteuttaa käytännöllisellä tavalla työelämässä. Tutkimus on luonteeltaan pilottitutkimus ja sen tavoite on kaksiosainen. Ensinnäkin tutkimus tarkastelee kahta arviointimenetelmää (itsearviointia ja ulkoista arviointia) suullisen kielitaidon arvioimisessa. Toiseksi tutkimus arvioi LangPerform -konseptiin perustuvaa elokuvapohjaista kielisimulaatiotestiä suullisen kielitaidon testaamisessa ja arvioimisessa.

Tutkimusta varten suunniteltiin ja toteutettiin kielisimulaatiotesti, johon osallistui seitsemän englannin kielen aikuisopiskelijaa. Osallistujien tekemää itsearviointia verrattiin kielisimulaation testisuorituksiin sekä ulkoiseen arviointiin. Suoritukset arvioitiin Eurooppalaisen viitekehyksen puitteissa.

Tutkimuksen mukaan itsearviointia voidaan pitää melko luotettavana arviointimenetelmänä jopa siinä määrin, että sitä voitaisiin käyttää ulkoisen arvioinnin rinnalla kielitaidon arvioimisprosessissa, erityisesti juuri edistyneempien opiskelijoiden parissa. Tutkimuksessa kävi myös ilmi, että arvioinnissa käytetyillä arviointitaulukoilla ja niiden tulkinnalla on keskeinen rooli arvioitiprosessissa. Kielisimulaatiotesti osoittautui hyödylliseksi välineeksi suullisen kielitaidon testaamisessa helppokäyttöisyyden, joustavuuden ja monipuolisuuden ansiosta. Simulaatiokonsepti myös lisää suullisen kielitaidon testaamisen luotettavuutta ja tasapuolisuutta, sekä mahdollistaa autenttisen kielimateriaalin sisällyttämisen testiin. Simulaation haasteena ovat tekniset rajoitteet sekä aidon vuorovaikutuksen luominen keskustelutilanteissa.

Asiasanat: suullinen kielitaito, kielitaito työelämässä, itsearviointi, simulaatiotesti, funktionaalinen kompetenssi

Table of Contents

1 Introduction

Insufficient language skills can be harmful for business and therefore new and innovative ways for testing language proficiency in the workplaces are needed. Indeed, Neeley et al. (2012, 236) claim that inefficiencies in language use can cause "loss of information, added work, loss of learning opportunities, and disruption of the collaborative process". This view is confirmed by an extensive survey in Norway (Hellekjær 2007, 6) as well as a fairly recent European wide survey (PIMLICO 2011, 10) on the use and need of languages in the corporate environment. The surveys reveal that the lack of sufficient language skills can be the cause of financial losses through, among other reasons, lost contracts and incorrect deliveries.

The objective of this study is to present some perspectives on how language proficiency testing could be done in a practical and useful way in the occupational field. Firstly, this study examines oral proficiency in the workplace context and compares two evaluation methods: self-assessment and external rating. Secondly, the convenience of a novel computer-based language simulation concept for testing and assessing oral proficiency is considered. This study is a pilot study. According to Van Teijlingen and Hundley (2001, para. 1) a pilot study is a smaller scale study before a major study and it can be used for testing the appropriateness of a research method or instrument. Therefore in this study both the results as well as the methods and procedures, especially the creation and design of the language testing instrument, play an equally important role. Douglas (2012, 75) argues that self-assessment can have positive effects on learning motivation and self-awareness resulting in increased learner autonomy and learning capacity. Furthermore, the computer-simulated language test could offer a fast and easy way to test language proficiency. The simulation test is based on LangPerform concept (e.g. Haataja 2010, Haataja and Wewer, 2012; Haataja and Wicke, 2014). The concept uses computer technology to support and review communicative language competence (Haataja 2010, 187-189). It also enables authentic input and provides tools for evaluation and documentation of language performances. For the evaluation of oral proficiency in this study the aspect of functional competence was chosen because it represents well the language needs of oral communication in the workplace context. Seven adult students of business English participated in the simulation test and their performances were rated by nine external raters against the illustrative scales for functional competence of the Common European Framework of Reference (CEFR).

The research questions of the study concern the two evaluation methods, self-assessment and external rating; and the simulation test as a method for testing oral proficiency:

Q1 - How consistent is the evaluation among the external raters?

Q2 - To what extent does the test takers' self-assessment correlate with the external rating?

Q3 – How convenient is the language simulation concept for testing and assessing oral proficiency?

The following section introduces background information on central topics of this study. The topics are testing oral proficiency, self-assessment, the Common European Framework of Reference, computer-based language testing, and oral communication in the workplace context. In section 3 the research materials and methods of the study are discussed and the creation of the language simulation is described. Section 4 presents the results of the study and section 5 attempts to answer the research questions, discusses the results of the study and presents some ideas for further research. Finally, section 6 provides the concluding remarks of the study.

2 Background

In this section the central topics of this study are introduced. The topics include testing oral proficiency, self-assessment, the Common European Framework of Reference, computer-based language testing, and oral communication and language testing in the workplace.

2.1 Testing oral proficiency

Speech is a central part of language use for second language learners and therefore testing oral proficiency or speaking skill is essential. However, Lado (in Fulcher 2003, 18) argues that in the field of language testing the importance of oral proficiency testing has not always been recognised:

> The ability to speak a foreign language is without doubt the most highly prized language skill, and rightly so …Yet testing the ability to speak a foreign language is perhaps the least developed and the least practiced in the language testing field.

When testing language proficiency, language competence is traditionally divided into four different skills: reading, writing, listening and speaking (Cumming 2008, 4). In many cases different parts of a language test focus on each skill separately; although some tests, such as the TOEFL, attempt to integrate two or more skills together. Douglas (2010, 19) points out that in real life these skills are sometimes used separately, for example, while listening to the radio or reading for pleasure. However, in communicative language use these skills are often combined, such as while listening to a lecture and making notes, or speaking and listening simultaneously during a telephone conversation. Nonetheless, whether the skills are tested separately or integrated, the fundamental objective of a language test is not to test how well a language user performs in these skills or on a particular task. Instead, the interesting part is the "language *ability*, which is manifested through the *skills* of reading, writing, speaking and listening" (Douglas 2010, 19, original emphasis). Also Bachman (1991, 309) points out that the focus should be in what the performance of an individual indicates about the individual's ability, however, only in the limits of the testing context.

In order for the language ability to be evaluated there needs to be some observable and measurable features that can be scored. Fulcher (2003, 19) talks about operationalization of internal and external abilities and argues that in speaking the

internal abilities (such as knowledge and competence) and external abilities (such as interaction and communication) are hard to distinguish from each other. The process of human communication is really complex and it makes the operationalization of speech challenging. Therefore, the concept of speaking itself is considered next.

Fulcher (2003, 23) defines speaking as "the verbal use of language to communicate with others". Spoken language differs from the written language in style but also in sentence structure and the use of vocabulary (Chafe and Tannen 1987, 384, 388). In comparison to writing, speech also contains more repetition and repairs (Fulcher 2003, 23). This is because speech is dynamic and once something is uttered it cannot be corrected or erased in a way that written language can (Biber and Quirk 1999, 1066). Indeed, the correction of speech can only be done through hesitations, false starts, reformulations and other disfluencies. In addition, taking part in a conversation requires spontaneous language use where planning and executing the utterances happen in real time (Biber and Quirk 1999, 1048). Planning, formulation and articulation of speech needs to happen in a reasonably short amount of time and if the process is automatic, it happens without conscious attention (Fulcher 2003, 24).

Littlewood (1992, 41-42) introduces a four level model for language production. The highest level, 'message level', includes conceptual planning-processes and representation of ideas and meanings. 'Functional level' involves selecting word meanings and broad syntactic frames. In 'positional level' the exact word-forms and sentence-structures are defined before the final and the lowest level, where the actual articulation of the words takes place. When all the levels function well together the speech appears fluent to the listener. However, if the lower-level plans (word-forms and sentence-structures etc.) are not automated enough it can cause the speech to appear less fluent. This can be the case, for example, with a second or a foreign language learner who has an idea about what to say but lacks the right words or grammar to actually articulate it. Other factors that have an effect on the speed and fluency of the speech are how complex the message is, how familiar the speaker is with the topic, how accurate the speech is expected to be and what are the consequences for making a mistake (Fulcher 2003, 24).

In speaking tests the learners typically encounter tasks through which their speaking abilities are displayed. Ellis (2005, 713, 721) describes a task as something that presents a communicative problem that has a relation to real-world activities. A task needs to be solved by using the language. Tasks motivate communication, and

using as well as learning language skills happens simultaneously. Bygate's (1999, 186) definition is much like the one presented by Ellis but he describes tasks as classroom activities in which "learners use language communicatively". Bygate also brings up the twofold challenge that a task sets for language: firstly, language is used "to achieve an outcome" and secondly, the overall purpose of the task is to learn the language. Fulcher (2003, 50, 47) links a task's meaning to language testing and describes tasks as "the means by which we can elicit a sample of language that can be scored". Testing students' language ability in all possible speech contexts would be absurd and impossible. Therefore tasks provide sample performances from which the "likely success or failure of the learner's future performance" can be estimated (Fulcher 2003, 47, 51). Fulcher goes on discussing the importance of context for language use and notes that tasks help the students to show what they can do with the language and provide a framework for using language in a test.

In a method called task-based language learning (TBL) tasks are used primarily for learning language. However, the principles of TBL can also been applied to language testing (Wigglesworth 2008) and in research TBL method has been used in evaluating oral language proficiency (e.g. Ellis 2009). In TBL the given instructions do not specify or restrict which aspects of the language should be learnt but direct the speaker to learn language through communicating in the language (Ellis 2005, 713, 721). Primary focus is on meaning but occasionally it is also acceptable to direct the attention to a specific form of language in order to practice it. TBL as well as other meaning-centred approaches have been criticised for not being able to teach students specific grammatical structures or sociolinguistic competence. However, although meaning is emphasised over form in TBL, focusing on language forms are not totally excluded. Furthermore, in TBL learner engagement is emphasised and if the task succeeds in providing reasonable challenge for the learners, they can be cognitively involving and also contribute to motivation in learning (Ibid.).

Typical tasks for testing oral proficiency include, for instance, argumentation (Bygate 1999); narration on films (Wood 2006) or on cartoons (Foster and Skehan 2013); and narrated picture stories (Rossiter 2009; Mochizuki and Ortega 2008). Tavakoli and Foster (2011, 37) studied how narrative task design influences the oral performance of second language learners in terms of accuracy, fluency, complexity and lexical diversity. In total 100 foreign language learners participated in the study, 60 in Teheran and 40 in London. The participants were shown cartoon frames and asked to

tell about them aloud. The results of the study revealed that task design has an effect on the second language performance. For example, complex storylines in the cartoons resulted in syntactic complexity in the language performance. In addition, the study showed that the participants who lived in London and used the target language in everyday tasks outside classroom had more diverse vocabulary. However, they did not necessarily produce more grammatically correct language.

In a more recent study Leaper and Riazi (2014, 177) argue that the focus of speaking tests is nowadays more in interactive communication instead of judging the participant's speaking ability on the basis of linguistic features in the speech. Therefore, speaking tests where the participants interact with each other are more common, such as group discussions, rather than conducting interviews with a trained interlocutor. Leaper and Riazi (2014, 200) conclude that group oral tests, where several students have discussions making use of task prompts and questions, are a convenient way of testing oral proficiency. Moreover, the testing focuses on interaction with peers and can have positive influence on teaching and learning the target language.

2.2 Self-assessment

In the evaluation of language proficiency teaching, learning and assessment can be seen as interconnected. Furthermore, self-assessment can play an important role as an evaluation method (Stoynoff 2012, 530). Sometimes two different terms are used: 'self-evaluation' for judgements that are used for grading; and 'self-assessment' which refers to informal judgements about attainment (Ross 2009, 3). However, in this study no distinction is made between the two terms and the term 'self-assessment' is mainly used. Underhill (1987, 22) attaches the term self-assessment to the process of interaction and defines self-assessment as an automatic, constantly present activity which enables communication. The speakers assess the process of communication by listening to themselves and observing how other people react and reply to what has been said. This type of self-assessment is unconscious and it is only noticeable when the communication breaks down. This study is more concerned with the definition provided by Klenowski (1995, cited in Ross 2009, 3), who defines self-assessment as the evaluation or judgment of 'the worth' of one's performance and the identification of one's strengths and weaknesses with a view to improving one's learning outcomes".

This type of self-assessment is done consciously and its objective is to promote learning.

Self -assessment plays a central role in learner-centred pedagogies (Little 2005, 322), the objective of which is to develop learner autonomy by encouraging students to be involved in the process of making decisions about the learning. Such decisions include setting learning targets, selecting activities and evaluating learning outcomes. However, the learners should be taught self-assessment skills so that they would be able to evaluate their learning outcomes. Indeed, learning self-assessment skills can help the learners to view assessment as a "shared responsibility" (Ibid.). Making self-assessment a part of normal evaluation procedures can also have positive effects on learning processes. Douglas (2010, 75) recognises the contribution that self-assessment and learner autonomy can have on the individual's learning capacity. Self-assessment supports reflection on language learning objectives and provides learners with "enhanced awareness of learning goals and criteria for judging the quality of their own learning" (Ibid., 75). Bullock (2011, 114) also promotes learner involvement in all learning processes, evaluation of learning outcomes included. Indeed, her study on teachers' attitudes on learner autonomy and students' self-assessment confirms some of the benefits of self-assessment discussed above. According to the study, most attitudes about self-assessment were favourable (Ibid., 119):

- When supported, learners benefit from assessing their own work
- Self-assessment raises learners' awareness of their strengths and weaknesses
- Self-assessment stimulates motivation and involvement in the learning process

Mikkonen et al. (2013, 75) argue that learning is based on reflection and define reflection as a process of critical observation of one's feelings, attitudes, thoughts and actions. Furthermore, Zavistanavičienė et al. (2006) studied university students' self-assessment in writing essays. The study described self-assessment as "a practical tool" in the classroom and reported that self-assessment "promotes students' autonomy and independent learning skills, makes students more active in judging their own progress and encourages them to see the value of what they have learned" (Ibid., 86). Similar results about the effects of self-assessment on self-motivation were retrieved by Weisi and Karimi (2013) who studied self-assessment among Iranian English learners and found out that self-assessment created positive attitudes towards English learning.

Self-assessment can have positive effects on language learning but it can also serve the purposes of language testing. In fact, Underhill (1987, 22) recognises the benefits of self-assessment and argues that as opposed to oral test assessment, where the judgement is based on only short speech samples, the judgement in self-assessment is based on a much wider perspective on the speakers' impressions of their own communication skills. Moreover, Leblay (2013) studied self-assessment in connection to oral proficiency testing. The study concerned adult French learners' self-assessment in oral skills and how well self-assessment correlated with external rating. Elements of peer review were also included in the assessment process. The speaking test consisted of two tasks and the performances were recorded. The learners were engaged in self-assessment in two ways. Firstly, the students define their language proficiency level according to proficiency levels and respond on paper to can-do statements that were based on DIALANG, an online-based language testing system. Secondly, the students listened to their own performances and compared them with the performances of their fellow students. The results of the self-assessment were then compared with external rating. Similarly to the present study, Leblay also used CEFR as a framework for both self-assessment and external rating. The results showed that the correlation between self-assessment and external rating was good. The best correlation between self-assessment and external rating was with the evaluation of language proficiency through DIALANG can-do statements (Leblay 2013, 230). The can-do statements were similar to the ones used in the present study for self-assessment. Likewise Ross (1998, 16) received similar results on self-assessment. The study compared self-assessment and external evaluation in functional competence and found correlation between them.

Further research on the reliability of self-assessment shows that self-assessment results reflect the real competence of the test takers, especially among adult students. For instance, Malabonga et al. (2005) studied oral proficiency testing with a computer-mediated test among second language learners of Arabic, Spanish and Chinese. Self-assessment was used to choose an appropriate starting level for the test. The study results suggest that self-assessment was a reliable way to do this for most of the students. Additionally, Brantmeier et al. (2012, 153) received similar results about the reliability of self-assessment. They investigated self-assessment among second language learners of Spanish in all of the four language skills and found out that advanced students had particularly good abilities to rate their performances. Apart from

that, the study emphasised that good rating criteria should be provided for the students in order for the self-assessment to be successful (see also Underhill 1987, 23).

The method of self-assessment has also received criticism. For example, Niemelä's (2012) Master's Thesis on self-assessment among university students revealed that some students found self-assessment uninspiring and burdening. Similarly Ross (2009, 8) reported on students' negative attitudes towards self-assessment. According to the study some students considered self-assessment boring or unfair because they felt that the teacher was trying to make them do the teacher's job. Little (2005, 322) emphasises that it is important that the teacher gradually teaches self-assessment skills to the students**.** Additionally, according to Underhill (1987, 23) certain factors can hinder the reliability of self-assessment. For example, lack of experience in comparing the language performance against external standards can cause problems in self-assessment (Ibid.). Furthermore, deliberate under-rating or over-rating of the performance can be problematic when using self-assessment. Also Douglas (2010, 75) reminds that the use of self-assessment should be considered carefully, especially if the learners profit from a higher rating. Such circumstance can occur, for example, if the self-assessment affects the final course evaluation (Ross 2009, 3). Other factors that may have an effect on self-assessment are age, gender and level of proficiency (Underhill 1987, 23).

2.3 Common European Framework of Reference (CEFR)

The Common European Framework of Reference (CEFR) has been named among the "most relevant and controversial documents in the field [of language learning and testing] in the twenty-first century" (Figueras 2012, 477). CEFR provides levels of proficiency which allow measuring the learners' learning progress in communicative language use and describe learning objectives in a "comprehensive way" (Council of Europe 2001, 1). CEFR separates different components in language competence and presents illustrative scales to describe skills and competence connected to those components. Six levels in language proficiency are distinguished and each level has a description on what skills are expected from the language user in order to reach that level. The levels are marked with six labels that form three groups: Basic User (A1 and A2), Independent User (B1 and B2), and Proficient User (C1 and C2) (Council of

Europe 2001, 23). The level descriptions are clear, brief and independent, and they are empirically developed and validated (Figueras 2012, 480).

Little (2005, 334) emphasises the usefulness of CEFR in planning curriculums and programs, selecting learning materials and developing assessment procedures. CEFR encourages self-directed learning which includes raising the learners' awareness of their present knowledge; setting objectives for learning; and self-assessment (Council of Europe 2001, 6). As an example, the European language portfolio (ELP), which is based on CEFR, is a useful tool for using self-assessment in language learning (e.g. Little 2005). Furthermore, CEFR represents a view of lifelong language learning among people of different ages (Council of Europe 2001, 5) as well as in different domains of life. Indeed, CEFR can be used in a flexible way for educational, personal and occupational purposes (Council of Europe 2001, 46). Many language tests in the occupational field base their assessment on CEFR, such as the Business English Certificates (BEC) which is aimed at people who are preparing for a career in business. O'Loughlin (2008, 77) estimates that in the future an increasing number of workplace assessments will be linked to CERF. However, despite the significance that CEFR has in different educational fields around Europe, CEFR has also received criticism. It has been criticised for "insufficient definitions and incoherencies" in the descriptions of language proficiency. In addition, the usefulness of the level descriptions for the second language acquisition has been challenged (Figueras 2012, 483; for more critical views on CEFR see Alderson 2007).

In this study CERF is used for assessing spoken language. Three categories were chosen to represent different aspects of oral proficiency:

- Functional competence: fluency and propositional precision
- Speaking skills: spoken production and spoken interaction
- Comprehensive language ability: Global scale

Functional competence represents language use requirements in the occupational field where communication has to be clear and understandable (see section 2.5). According to Amos (2012, 457) fluency of speech and precise expression are essential features of clear communication of the message. Functional competence is "concerned with the use of spoken discourse and written texts in communication for particular functional purposes" (Council of Europe 2001, 125). Two illustrative scales in CEFR describe the functional competence of a learner (Council of Europe 2001, 128):

a) *fluency*, the ability to articulate, to keep going, and to cope when one lands in a dead end

b) *propositional precision*, the ability to formulate thoughts and propositions so as to make one's meaning clear

In fluency the continuous flow of speech is emphasised (Council of Europe 2001, 128). The descriptor of the fluency scale includes flow of language, spontaneity, ease of expressions, tempo of speech, pauses and hesitations. De Jong and Perfetti (2011, 533) define fluency as the ability to "express thoughts easily, with more attention to meaning than form, in any given situation" with "smooth" communication and "relatively fast and automatic" production processes. Littlewood (1992, 66) gives a similar definition for fluency and talks about 'fully automated' and 'semi-automated' language which causes the speech to be more or less fluent depending on the complexity of the situation. Semi-automated forms can be used with familiar topics only, whereas fully automated forms can be used in any context. The more language forms are automated, the more fluent the speech appears.

In propositional precision the clarity of speech is emphasised (Council of Europe 2001, 128). The descriptor of the propositional precision scale includes shades of meaning, accuracy, comprehensibility as well as precision and details of given information. The illustrative scales for fluency and propositional precision can be found in Appendix 1.

Speaking skills are represented by two subcategories: spoken production and spoken interaction. CEFR separates four language activities involving spoken (and/or written) language, of which the two in the middle are relevant for us: reception, production, interaction and mediation (Council of Europe 2001, 14). Productive activities, such as oral presentations or reports, are important in academic and professional fields. In interactive activities, on the other hand, individuals participate in (written or) oral exchange where production and reception of messages take turns and may even happen simultaneously. The category of spoken production concentrates on producing "an oral text which is received by an audience of one or more listeners", such as a speech or a presentation (Council of Europe 2001, 58). The category of spoken interaction focuses on the learner's ability to interact spontaneously and take part in a conversation (Ibid., 73).

The global scale is used for defining the comprehensive language ability or the overall language proficiency of a language user (Council of Europe 2001, 24). The global scale takes a holistic view into language proficiency. It is recommended to be used as an "orientation point" (Ibid.). In this study CERF is used to guide the design of the methods and procedures of the study. In addition, CEFR provides a framework for the two evaluation methods used in this study, self-assessment and external rating.

2.4 Computer-based language testing

Computer technology can open up new and creative ways for learning and testing language skills. Computer-based testing (CBT) refers to using computers to deliver, score and report scores of assessment (Ockey 2009, 836). CBT has received much attention in the field of second language education. Since the 1980's, when the personal computers became affordable also for ordinary people, the use of computer technology for learning and testing language ability has increased rapidly (Wilson and Fox 1982, 145; see also Davies 1984). Computer-based language testing (CBLT) refers to using CBT for testing language proficiency. Many language tests that are used worldwide make use of computer technology. As an example two language tests should be mentioned, IELTS (International English Language Testing System) and TOEFL (Test of English as a Foreign Language), which are designed for evaluating academic English skills.

Brown (1997, 45) analysed the use of computers in language testing some 20 years ago stating that the easiest language tests that can be adapted to computer-based testing medium include tasks that test receptive skills or grammar, such as multiple-choice, true-false, and matching items, fill-in or cloze. Often such tasks, however, could easily be done on paper as well. Likewise James (1996, 18), writing in the same period as Brown, describes practical speaking activities on computer and argues that activities that would also work without a computer are not worth doing with a computer. Brown (1997, 45) continues by suggesting that the most challenging tasks on computer test productive skills such as compositions, oral presentation and role-plays. Much has changed since Brown and James discussed their ideas. However, the challenge to use computers for testing oral skills in an authentic way remains the same. Despite the challenges, the work for developing computer technology to suit the needs of oral proficiency testing should be continued. Indeed, interactive tasks and role-plays seem

to be effective ways to measure oral proficiency. In fact, role-play has been claimed to be a "reliable instrument for assessing candidates' performance on the given task" and the competence that is displayed in the role-plays correlate well with the competence that is employed in ordinary conversations (Okada 2010, 1648). Underhill (1987, 51) defines role-play as an activity where the participant "is asked to take on a particular role and to imagine himself in that role in a particular situation" and communicate in a way that is "appropriate to the role and the situation".

Lee (2000, under the headline "Why use CALL?") lists some benefits that working with computers can offer for language learning but these can also be applied to language testing: experiential learning, motivation, authentic materials, greater interaction, individualisation, independence from a single source of information and global understanding. Additionally, Chapelle (2003, 28) recognises the possibilities of CBT in "authentic input that is relative to the context" and the positive influence that it can have on the validity of the test. Research on computerised language testing also reveals further benefits for using computer in language testing. For instance, Jamieson et al. (2013) examined a fully automated computerised speaking test for admission and placement of students in a university. Besides delivering the test and collecting the responses, a fully automated test uses technology also in scoring the responses. The test turned out to be useful in the following ways (Ibid., 290):

- Can be taken at many locations, improves test accessibility and availability
- Can be delivered to large number of test takers
- Economizes the test administration efforts
- Standardized instruction and prompts improve the test's reliability and fairness

In the study it was also argued that because the results are scored by a computer program, fully automated computerised language tests can lower the costs of doing language proficiency testing. However, the expenses can increase if the results are scored by human raters (Brown 1997, 46). Nevertheless, CBT can be used in a cost-effective way to test language skills. For example, Álvarez and Laborda (2011, E136) examined a university entrance test in Spain and found out that CBLT saves time, it can be delivered and rated in a lower price, and because of this, tasks that are usually costly can be included in the test, such as the speaking section. In addition, learners generally have positive attitudes towards using computers and the technology allows for individuals to work on their own pace. Moreover, El Hmoudova (2013, 411) states that

CBLT allows, for example, increase of delivery, effective administration and scoring, new advanced and flexible item types as well as improved test security, consistency and reliability. On the other hand, using computer in language proficiency testing can also have negative effects such as computer anxiety (Brown 1997, 48). Also differences in familiarity of using computer can influence the performance negatively. In addition, making use of computers in language proficiency testing can be challenging if the computer equipment are not available or they are out of order.

The speaking test used in this study was carried out as a computer-based language simulation test. The test is based on the LangPerform concept developed by Kim Haataja, the director of RULE (the Research and Development Unit for Languages in Education at the University of Tampere). The LangPerform concept is designed to support, review and evaluate communicative language competence with the help of video-based computer simulations and web-based assessment instruments (Haataja 2010, 187-189). Furthermore, the context of the language performance is made as authentic as possible. Bachman (1991, 307) defines authenticity as the "extent to which test tasks replicate 'real-life' language use tasks" and argues that authenticity plays an important part in the generalizability of test results (Ibid., 301). In addition, Fulcher (2003, 54-55) points out that authenticity is a matter of perception and can mean different things to different people. Furthermore, creating a sensation of authenticity is far more complex than simply copying situations, topics or discourses from the real world. Indeed, there is some debate about the extent to which tasks can represent authentic real world activities (Wigglesworth 2008, 117-118).

In the LangPerform concept the language testing situation is supposed to be pleasant and relaxing for the test takers (Haataja 2010, 187-189). The simulation as a testing instrument enables testing, evaluation and documentation of communicative language use in situations that are natural, variable and reality-like. In addition, the LangPerform concept enables creating a learning profile for each participant. The test performances are stored in a central server and the online evaluation instrument allows objective, valid and reliable external rating as well as self-assessment on each individual performance, independent of time or place. The performances are reviewed e.g. against the CEFR scales. Each simulation can be designed according to the needs and interests of the target group.

The LangPerform concept is new and innovative, and already several research and development projects on the concept have been conducted in the recent years and

some of them are still ongoing. Funders for these projects include Federal Foreign Office of Germany, the Goethe-Institut, the Nordplus programme and the Finnish National Board of Education. Many of the projects are closely related to Content and Language Integrated Learning (CLIL), an educational approach where subject content is taught through a language other than the first language of the learners. Thus the working language is both a target of learning and an instrument for learning a subject. Projects that use the LangPerform concept and its instruments for learning and teaching various target languages include e.g. ProfiDaF, CLILiG-SCAN and INNOCLILiG, as well as PROFICOM and INTOCOM. The LangPerform concept and the simulation instruments have also played a key role in EVALANG, a national in-service training programme for language teachers in upper secondary education in Finland (for details, visit http://rule.uta.fi/en/about/).

So far three Master's theses have been written on the LangPerfom concept in the recent years, one in German and two in French. Tuuna-Kyllönen (2011) discusses the experiences of German learners in upper secondary school and compares the computer-simulated speaking test to the speaking test of National Certificate of Language Proficiency which is used as an optional language proficiency test by Finnish National Board of Education. The strengths of the simulation seem to be the possibility for individual and impartial performances and a technology which enables encounters with native speakers. However, the lack of authentic interaction, technical problems and getting used to the testing procedure were seen as weaknesses. Furthermore, Ilkankoski (2012) discusses the use of the simulation test in measuring language proficiency in different language skills among French learners, while Hasan (2011) concentrates on learning French pronunciation with the help of the simulation.

Just recently a doctoral dissertation based on the use of the LangPerform concept has been published at the University of Turku (Wewer 2014). The dissertation considered language assessment at primary level in Finnish basic education. In addition, two doctoral dissertations are in the making at the University of Tampere on interpreter education (Viljanmaa) and on French proficiency in upper secondary education (Kemppainen).

2.5 Oral communication in the workplace

Language use and communication in the workplace context has been widely studied in the recent years both internationally (e.g. Ehrenreich 2010; Nickerson 2005; Holmes 2000) as well as in Finland (e.g. Räisänen 2013; Louhiala-Salminen and Kankaanranta 2011). English is the most widely spoken foreign language in Europe, including the workplace context (European Commission 2012). Although multilingualism is increasingly important in international encounters, the central role of English language is unquestionable. PIMLICO (2011, 16), an empirical study of foreign language use in European companies, reveals the dominance of English language in global trade. The study reports that people operating in international trade are assumed to have good English skills. Furthermore, English skills are often considered as generic skills in a similar way as computing skills, which further emphasises the importance of good English skills. Like tendencies can be seen in the Finnish workplaces where English skills are often regarded as self-evident. Lehtonen and Karjalainen (2008, 495) report on situations where some applicants are not even considered for a job because of poor English skills. A large scale national survey (Leppänen et al. 2009, 41) on the attitudes of Finns towards English language shows that more than 40% of Finns use English in their workplaces. English is not only used in external communication with customers and partners but in some cases English is becoming the language of internal communication within a company as well (e.g. Louhiala-Salminen et al. 2005).

The present study focuses on testing oral proficiency in a workplace context. Moslehifar and Ibrahim (2012, 530) argues that in order to be successful in the workplace, good oral communication skills are needed, even more than good written communication skills (Kassim and Ali 2010, 168). Both written and spoken skills in the workplace context have been studied in the recent years, such as the use of emails (e.g. Warren 2013; Evans 2012; Louhiala-Salminen et al. 2005). Additionally, research on oral communication in the workplace covers, for example, negotiations and sales interactions (Charles 1998; Firth 1995); business meetings (Räisänen 2012); and everyday spoken business language (Handford and Matous 2011). Oral communication can refer to many different kinds of situations that involve speaking. These situations include, for example, formal presentations or participating in teams and meetings as well as telephone conversations and informal work related discussions (Crosling and Ward 2002, 41; Moslehifar and Ibrahim 2012, 530). Crosling and Ward (2002, 47) claim that good oral communication skills are important in many different areas of working

life such as recruitment, job success and promotion. Employees benefit from strong oral communication skills but they will be "disadvantaged in the workplace if they lack these skills" (Ibid., 56).

According to Crosling and Ward (2002, 53) the most often used forms of oral communication are “informal work-related discussions, listening and following instructions, and informal conversations." Interestingly, this kind of informal communication and small talk have also been mentioned as the most challenging language use situations. Indeed, Charles (2007, 272) argues that using field specific terminology or language in formal communication is less problematic than the “discursive conventions of informal communication”, or finding the right expressions in “ordinary small talk” and being able to “suddenly and effectively express opinions or convey nuances”.

Successful communication consists of many different elements. In this study three elements that contribute to successful communication in the workplace context are discussed: politeness, context-dependency and clarity. The discussed elements are interdependent and cannot fully be separated from each other.

Politeness, or “making it sound nice” (Kankaanranta and Louhiala-Salminen 2010, 207), is important in formal encounters but it is also very important in informal communication which often takes place, for example, around the coffee machine or in between meetings. It is essential for networking and creating bonding relationships between employees, which in turn contributes to knowledge sharing and the accumulation of social capital within the company (Charles 2007, 272). Kankaanranta (2010, 207) studied business communication among international business professionals and found out that politeness can be displayed in overall interpersonal orientation where the “non-business” part of communication is emphasised, such as in small talk situations. Indeed, small talk plays a significant role in being polite and it can be used to create rapport between clients or “building relations and trust between staff in companies” (Pullin 2010, 455).

The significance of context awareness in communication is undeniable. Language is never used in a vacuum or only for its own sake (Douglas 2010, 20) and therefore it is very important to have a context for language use. Bachman (1991, 82-83) argues that context defines the appropriate use of language and it includes “both the discourse, of which individual utterances and sentences are part, and the sociolinguistic situation which governs, to a large extent, the nature of that discourse, in both form and

function". In other words, the context in which a person is speaking guides the speaker's choices on what vocabulary to use, how to address other speakers and what formality of language should be used etc. Apart from that, Louhiala-Salminen and Kankaanranta (2011, 247) emphasise flexibility in language use and the ability to formulate the message "appropriately in a given context of interaction". In brief, successful communication is always context-dependent (Crosling and Ward 2002, 43).

Clarity of speech ensures that what has been said can be understood by the listeners. Indeed, both Kankaanranta (2007, 255) and Sweeney and Hua (2010, 481) emphasise the importance of communicating messages clearly and so that they can be understood by other speakers. Being clear and getting the message through is the primary focus of communication in the workplace. Kankaanranta and Louhiala-Salminen (2010, 205) suggest that the purpose of language in workplace communication is to "get the job done". In a similar way, Sweeney and Hua (2010, 481) emphasise that working and doing business is "inherently goal oriented" and those goals are reflected in the behaviour of employees, language use included. Furthermore, according to Crosling and Ward (2002, 42) communication and social interaction are "the means for achieving occupational activity, enabling employees to learn and acquire new skills which facilitate the development of solutions to problems".

When using language in the workplace context, the functional aspects of language use are emphasised. Indeed, most important is what can be achieved through the language use and how well the use of language is serving the purpose for which it is being used. Therefore, when discussing language use in the workplace context the primary focus is not in language itself. Linguistic correctness or grammaticality play a secondary role. Indeed, the grammatical or idiomatic correctness are not considered the most important aspects of the language when communicating in the workplace (Kankaanranta and Louhiala-Salminen 2010, 207). Moreover, the use of business English can vary "enormously" in quality and it can be seen as deviating from the ideal of native speaker English (Louhiala-Salminen and Kankaanranta 2011, 248). Firth (2009, 152) holds a similar view about workplace interaction among non-native speakers of English and describes it as

> variously 'marked' and at times linguistically and discursively extraordinary, [but] also real, authentic, effective, expedient and, it appears, endogenously treated as contextually appropriate and ordinary.

Indeed, much of the workplace communication in English takes place between speakers whose first language is other than English. It has been estimated that more than 80% of the worldwide daily English business communication takes place in ELF, English as a *lingua franca* (Kankaanranta & Louhiala-salminen 2007, 56). In ELF, English is seen as a "shared resource" and it is used for communication between non-native speakers of English (Louhiala-Salminen and Kankaanranta 2011, 245). A special term, BELF, has been coined to refer to the use of English as a *lingua franca* in business. BELF is "the language that business professionals from different cultural and linguistic backgrounds use to conduct their daily work activities" (Louhiala-Salminen and Kankaanranta 2011, 248). Effective communicative skills and competence in BELF are very important in business and they can be considered a part of the overall business know-how (Räisänen 2013, 19-20).The uses of ELF and BELF are interesting topics but they are not the primary focus of this study. In this study ELF only plays a minor role as non-native speakers of English are included in the language simulation (see section 3.2.2).

2.5.1 Language testing in the workplace

Assessment of second language skills in the workplace is part of Language for Specific Purposes (LSP), an established branch of applied linguistics (O'Loughlin 2008, 69). *English for Specific Purposes* is a peer-reviewed journal that publishes articles related to LSP. LSP is divided into two parts, languages for academic purposes (e.g. Davies 2001) and languages of occupational purposes (e.g. Huhta 2010). Language tests for occupational purposes are designed to assess "whether an individual has the language skills to assume … relevant professional or vocational duties" (O'Loughlin 2008, 69). According to Tratnik (2008, 7) language proficiency assessment in the workplace should be economical in terms of administration, time and money, and it should also be authentic, accurate, reliable and have beneficial effects. Lockwood (2012, 111), on the other hand, emphasises a need for workplace specific language training which can be "tailored" according to the work that is done in the particular organisation. Language tests that concentrate in language use in the field of business include, for example, the Business English Performance Test (BEPT), the Oxford International Business Certificate (OIBEC) and the Business Language Testing Service (BULATS) by Cambridge ESOL which is used in 30 countries by businesses for recruitment, training

and staff development (O'Loughlin 2008, 77). Sometimes large-scale tests such as IELTS or TEOFL that are designed for testing English for academic purposes are used in the occupational field, although this is considered unethical by many researchers (O'Loughlin 2008, 78).

When designing a test for specific purposes a number of factors should be taken into consideration (Hamp-Lyons and Lumley 2001, 130). Firstly, why is the test designed in the first place, what is the need that it fulfills? Secondly, what content will be included in the test? This question is important because the content of the test provides both the target and vehicle for communication. Thirdly, it should be considered whether the test serves only for a particular purpose or can the same test be used for other purposes as well. And finally, when designing a test the point of view of the test taker should also be taken into consideration.

In workplace language assessment, the readiness of the test takers is often tested by designing tasks that simulate real world situations related to particular employment situations (O'Loughlin 2008, 69). The test takers are required to achieve particular communicative functions instead of just displaying their linguistic knowledge. This is in line with the idea that in the workplace setting language is used as an instrument and not only for its own sake (see section 2.5). Typical methods for testing language ability in the workplace are, for example, face-to-face interviews (Edwards 2000; Lumley 1998) or interviews on the phone (Lockwood 2012). Additionally, Räisänen (2013) used audio and video recordings in real communication situations. Recordings have the benefit that the data can be easily stored. On the other hand, gathering the data can be time-consuming and laborious. Furthermore, language testing and assessment can also be linked to language training such as a course where the language abilities are tested before and after attending the course (Roberts 2005, 128).

The use of simulated situations and role-plays is typical for language testing in the workplace (e.g. O'Loughlin 2008; Edwards 2000; Jacoby and McNamara 1999; Lumley 1998). For example, already in the beginning of the 1990's role-plays were included in assessing the English language competence of medical and health professionals in Australia in the Occupational English test (OET) (O'Loughlin 2008, 71). In the speaking part of OET the test takers participated in role-plays with a trained interlocutor. In the OET the clarity of communication was emphasised more than grammaticality. The assessment of the test was restricted to the aspects of language performance, and non-linguistic factors, such as background knowledge, were not

evaluated. Instead, factors that were assessed included overall communicative effectiveness, intelligibility, fluency, comprehension and appropriateness of language as well as grammar and expression. As a conclusion, O'Loughlin (2008, 72) emphasises that when testing language use in occupational field, the most important criteria for assessing the performances should not be linguistic ability, but instead the communicative competence should be tested.

3 Methods and materials

The methods and materials used in this study are presented in this section. First, the participants of the study are introduced and secondly, the procedure and the data collection of the study are described. This includes introducing the self-assessment questionnaire and the simulated language test. In addition, the rating of the samples as well as the transcription of the data is described. Because this is a pilot study, the procedures of the study play an important role and therefore the process of creating the data collection method is described in detail.

3.1 Participants

Two groups of participants took part in the test: test takers and raters. Seven test takers filled in a self-assessment questionnaire and completed the simulated speaking test. Recorded samples from the speaking test performances were evaluated by nine external raters.

3.1.1 Test takers

The test takers were mature students participating in a business English course at the Summer University of Tampere in fall 2012. The recommended overall language proficiency level for the course participation was B1-B2 of the CEFR levels. Altogether 23 students were enrolled on the course. Participating in the speaking test was optional for the students. It was agreed with the course teacher that the speaking test would be included in the course curriculum and could therefore be carried out during the regular course hours. The test takers were asked to fill in an online self-assessment questionnaire about their English language skills. After that they participated in the simulated speaking test. The performances of seven test takers were included in this study on the basis that both the self-assessment questionnaire and the speaking test were successfully completed. Two of the test takers were men and five were women. Most of them were working in middle management positions in different organisation in Finland.

3.1.2 Raters

The language samples were rated by nine raters. Eight of the raters were English language student teachers at Tampereen yliopiston normaalikoulu, a teacher training school of the University of Tampere. Student teachers are university students who teach in a school for a period of time as a qualification for the teacher training. Five of the student teachers (from now on referred to as *student raters*) were women and three were men. The student raters were not expected to have any previous experience in evaluating spoken language performances. For this reason it was decided that also a rating of a language teacher (from now on referred to as *expert rater*) with previous experience in evaluating spoken language performances should be included in the study as a point of comparison. The expert rater chosen for the task was a male English language teacher at Tampereen yliopiston normaalikoulu and he had previous experience in teaching and assessing oral English courses at upper secondary school level.

3.2 Procedures and data collection

The procedures and data collection for the study involved three different phases. In the first phase the instruments for collecting data were created. This included designing and creating the self-assessment questionnaire and the simulated speaking test. In the second phase the test takers filled in the questionnaire and completed the speaking test. In the third and final phase the language samples from the speaking test performances were rated. The final phase also included transcription of data.

3.2.1 Self-assessment questionnaire

The test takers filled in a self-assessment questionnaire concerning their English language skills. The questionnaire was online-based and was titled “How is your Business English?”. The test takers completed the questionnaire before participating in the speaking test in fall 2012. The questionnaire was seven pages long and it took about 20 minutes to finish it. The test takers answered the questions in the questionnaire according to their general understanding of their own language skills and on the basis of their overall previous experiences as language users. Many of the questions and their formulations were inspired by CEFR and most of the assessments were done against the CEFR scales. The purpose of the questionnaire was to serve as a language profile

for the simulation test (Haataja 2010, 189) and also to get data for the present study about the self-assessment of the test takers. The questions and instructions in the questionnaire were presented in English.

The questions in the questionnaire concerned personal information, such as name, gender, occupation and mother tongue; language skills in foreign languages other than English in writing, reading, listening and speaking; and the use of English in different domains: occupation, education, public places and personal life (see Council of Europe 2001, 14). Skills in English language were then inquired in more detail. Firstly, the test takers were asked to choose a reference level against the global scale (Council of Europe 2001, 24) to represent their overall language skills in English. Secondly, a Likert-type rating scale from 1 to 10 was presented for reading, writing and listening skills, in which 1 represented "very basic skills" and 10 "near native skills". For evaluating speaking skills, the test takers were encouraged to choose a level description of the CEFR scales for spoken production and spoken interaction (Council of Europe 2001, 58, 74) that best described their skills.

In the last section of the questionnaire the test takers were asked to evaluate their skills in functional competence which included fluency and propositional precision (Council of Europe 2001, 129). The test takers were presented with 17 can-do statements concerning their skills in functional competence. The statements were slight modifications of the illustrative scales for functional competence in CERF and were worded as 'I can (do X) ...', such as "I can express myself with relative ease and keep going understandably and effectively without help." The test takers could then choose one of the two alternatives as a response to the statement ("Yes, I can" or "Not yet") according to their estimation on whether or not they thought they were able to do what was described in the statement. Similar statements are also used in an online-based language testing system DIALANG (e.g. Figueras 2012, 480; Alderson and Huhta 2005, 303). Similarly to the level descriptions in CEFR, the statements in the questionnaire were worded positively with the emphasis on what the learners already know instead of focusing on what they do not now (Council of Europe 2001, 205; Figueras 2012, 480).

The parts of the questionnaire that concerned English speaking skills (spoken production and spoken interaction) and functional competence (fluency and propositional precision) were included in this study. Furthermore, the self-assessment on the overall English skills against the global scale was included as a point of

comparison. However, the results of the self-assessment against the global scale were not taken into consideration in the analysis and comparison of the different ratings. The self-assessments that concerned reading, writing or listening were not included in this study. The scales and can-do statements as well as level descriptions can be found in Appendix 2.

3.2.2 Computer-simulated language test

The computer-simulated language test used in this study was designed and created for the purposes of the study. The test is based on the LangPerform concept developed by Kim Haataja, the director of RULE (the Research and Development Unit for Languages in Education at the University of Tampere). The making of the simulation was realised in cooperation with Mr. Haataja and RULE. Because many of the projects relating to the LangPerform concept were designed for the needs of pupils, students and teachers in educational institutions (see section 2.4), there seemed to be a need to expand the research into other domains of language use as well. For this reason language use in a workplace context was chosen. Designing the tasks for the simulation and writing a manuscript was done together with Johanna Litmanen, a research assistant at RULE. This included discussing ideas about characters, writing and rewriting dialogues, planning scenes and locations as well as refining the text. In addition, the manuscript included planning of visualisation, sound effects and timing. All actors in the simulation film were non-professional and participated on a voluntary basis. The three main actors were native speakers of English with British and Australian backgrounds and the four supporting actors were non-native speakers of English with Indian, Italian and Finnish backgrounds. In addition, the five extras had Finnish backgrounds. Native speakers were chosen for the leading parts to ensure authentic language material. Indeed, according to Sweeney and Hua (2010, 48) native speakers are often involved in EFL communication and can cause "misunderstanding in intercultural interactions". Non-native speakers were included because, as discussed in section 2.5, the majority of worldwide English business communication takes place between non-native speakers of English (Kankaanranta and Louhiala-salminen 2007, 56) and therefore also contribute to authentic language input. Most of the scenes for the simulation were filmed during one evening in fall 2012. Jussi Hiidenuhma filmed the scenes while Ms. Litmanen directed. She was also responsible for editing picture and sound for the

simulation. Markus Ackermann, a software development expert at RULE, integrated the film to the simulation software and added instruction texts and other visual aids on the screen, such as a time bar and a continue-button. The Langperform Lab, an information network based language laboratory, where the simulation can be completed and reviewed, required an internet connection and the operating system Windows 7.

In the simulation the test takers encounter different language use situations and are required to solve them by speaking English. The test taker is invited as a guest speaker to an imaginary conference called "Focus on Finland" where different aspects of Finnish culture and nature are discussed. The frame story provides contextual information about the proper use of language and gives a reason for completing the tasks. Indeed, Fulcher (2003, 51) highlights the importance of a framework in language test that guides the test taker's language use. Without a framework, the test taker would be required to complete random assignments and tasks without a context, which can create a very unnatural situation from the point of view of the test taker. The frame story and the tasks in the simulation were designed to represent different working life situations, such as talking on the phone and giving a short speech. Inspiration for the different types of tasks were sought in various business English course books and other literature as well as in informal interviews and conversations with teachers of business English. In total, the simulation included five tasks and four self-reflective sections as follows:

Task 1: Introduction
Task 2: Presentation + Self-reflection
Task 3: Speech + Self-reflection
Task 4: Conversation + Self-reflection
Task 5: On the Phone + Self-reflection

In the first task the test taker arrives to the location where the conference is held and is required to inquire a way to the conference from a receptionist. The second task takes place after arriving in the conference room and the test taker is asked to shortly introduce himself/herself to other people attending the conference. In the third task the test taker is asked to give a short speech on the topic "Four seasons and thousands of lakes". The test taker is instructed to include in the speech characteristics of the Finnish nature and the four seasons, and to describe a bar chart that shows the average temperatures of each season in Finland. For the third task the test taker is given some

time to prepare for the task. Ellis (2009, 474) calls this strategic pre-task planning which involves "planning what content to express and what language to use but without opportunity to rehearse the complete task". The context of the third task is semi-formal and represents the spoken production skills (Council of Europe 2001, 58). After the main task the test taker is presented with a question from the audience about polar bears in Finland and the test taker is expected to answer the question spontaneously. In task four the test taker engages in small talk, an informal conversation with a colleague on a coffee break. Small talk can be "more or less" work related and it has an important function in the workplace because through it interpersonal relationships are created and maintained between co-workers and clients (Holmes 2000, 36). Furthermore, good relationships indirectly serve the organisation's goals. In order to complete the fourth task the test taker needs to use language spontaneously and is not allowed any time to be prepared for what to say. In such within-task planning "the planning … occurs on-line while learners are actually performing a task" (Ellis 2009, 474). The fourth task represents spoken interaction skills (Council of Europe 2001, 73). In the fifth and final task the test taker makes a phone call and is asked to leave a voicemail and set a meeting with a colleague.

After tasks 2-5 the test taker is encouraged to reflect on the previous task performance. The test taker meets an imaginary character Mr. C who is the "English language conscience" in the test taker's imagination. His function is to help the test taker to reflect on the language learning process during the speaking test. Mr. C asks questions such as "What went well?" and "What could be improved?" and the test taker can freely answer the questions within the given time limit.

All the tasks in the simulation include instructions and hints about how to react and what to say on each task. The instructions are written and appear in the top part of the screen, such as: "Answer the question" or "Ask: enjoy travelling?". For each task and self-reflection section the test taker is given a certain amount of time to complete the task. The time limits vary from 6 seconds to 2 minutes and 30 seconds according to the character of the task. This type of pressured planning (Ellis 2009, 474) was chosen because it enabled controlling the overall duration of the simulation test. Another reason for providing limited time for completing the tasks was to imitate real-world situations where interaction often needs to be spontaneous and automatic, and there is no time for long planning. In some of the tasks the test takers were given a choice to move forward in the test by pressing a 'continue' button that appeared on the screen after a certain

time. This way the test taker was given a minimum and a maximum time that could be used for completing the tasks and the test taker could adjust their performance according to those limits. The total duration of the simulation test was about 25 minutes.

The figure 1 illustrates the written instructions and the time bar which indicates the amount of time that can be used for the task.

Figure 1. Still frame of the simulation test.

3.2.3 Rating of samples

Two language samples from each test taker's performance on the simulated speaking test were chosen to be rated. The two samples were chosen to represent two aspects of speaking skills presented in CEFR, spoken production and spoken interaction. Sample 1 represents the spoken production skills and consists of the task three "Speech" (see previous section 3.2.2). Sample 2 represents spoken interaction skills and includes four excerpts from tasks one, two and four, which all require spontaneous language use. Samples 1 and 2 are both two minutes long. The samples were converted to MP3 format for the rating.

The two samples were rated by eight student raters and one expert rater. The student rating took place in fall 2013 and the expert rating in spring 2014. In order to

improve the reliability of the rating all the raters received the language samples in different order. The raters were given the following material:

- Introduction and instruction sheet
- Rating scales for spoken fluency and/or propositional precision
- Samples 1 and 2 of three, four or eight test takers in MP3 format
- Rating form for marking down the CEFR level
- Comment sheet

One student rater and the expert rater submitted the ratings via email. Other raters did the rating individually on appointed times and each rating lasted for about 30 minutes. Student raters evaluated both samples in either spoken fluency or propositional precision. The expert rater evaluated the performances both in fluency and propositional precision. The student raters evaluated all seven test takers' samples: four student raters evaluated the samples of three test takers and four student raters evaluated the samples of the remaining four test takers. This procedure was chosen because none of the student raters had previous experience in evaluating oral performances and the workload was eased.

The rated samples were 2 minutes in duration. Are such short samples sufficient for making any assumptions on learners' speaking skills? As the rater cannot be present in all the communication situations of a language user, the judgment of spoken language abilities is often based on only short samples of speech. In the 1940's Knower (1944, 492) evaluated spoken proficiency on the basis of "a great variety of data" which included observations of speech performances and "a review of records of interests and achievements". In the 1980's Underhill (1987, 22) commented on the unsatisfactory evidence that short speech samples provide for speaking competence. By 'short' he meant samples of about 10 minutes in duration. The samples used in this study were significantly shorter. However, a quick review on research on spoken language testing reveals that using language samples of similar length is actually quite common. Leblay (2013), Frost et al. (2012) and Rossiter (2009) used samples of 1-2 minutes in duration. On the other hand, in several studies (Tavakoli and Foster 2011; Birjandi and Ahangari 2008; Mochizuki and Ortega 2008; Wood 2006; Bygate 1999) the duration of the samples is undefined.

The language samples were rated against the illustrative scales for functional competence, which include spoken fluency and propositional precision (Council of

Europe 2001, 129). These rating scales were chosen because they represent a functional aspect of language use in CEFR and were seen to represent the context of this study, oral communication in the workplace. The raters were asked to choose one of the reference level labels A1, A2, B1, B2, C1 or C2 to represent the sample they were rating. Both samples 1 and 2 were rated separately. In the illustrative scale for spoken fluency the rater's attention is directed to hesitations and pauses, flow and tempo of the speech, the ease of production, naturalness, spontaneity, false starts and reformulations. The illustrative scale for propositional precision relates to shades of meaning, modification and qualification, modality, informational details and precision. Green (2012, 31) analyses the illustrative scales for functional competence and finds that in order to receive a C level marking the learner needs to show "greater sensitivity to context, greater spontaneity and less hesitancy than at the lower levels". The illustrative scales for spoken fluency and propositional precision are presented in Appendix 1.

After marking down the reference level label in the rating form the raters were asked to shortly comment on why they chose the label for the sample. When the raters had finished rating the given samples they filled in a comment sheet which contained three questions about the rating process:

1. Which aspects did you focus on in the rating?
2. What was difficult / easy in doing the rating?
3. How would you comment on the length of the language samples in rating spoken fluency / propositional precision?

3.2.4 Transcription of data

The data of the study consists of the self-assessment questionnaire, transcriptions of the language test performances and the comment sheets of the raters. The language test performances in samples 1 and 2 were transcribed according to the conventions of Conversation Analysis (CA) based on Jenks (2001). All the words in the transcript are written in lower-case letters. This was done because capitalising letters was not seen to give any additional information to the content of the transcript. Instead, it could have distracted the flow of speech. Transcripts are representations of talk that do not as such represent sentences but utterances (Jenks 2001, 96), and therefore the words that begin a new utterance are not capitalised. Table 2 presents the markings of the features of talk that were included in the transcript.

Table 2. Markings of the features of talk in the transcript.

Feature	Marking	Description
Silent pause	(.)	≤ 0.1 sec
Silent pause	(0.3)	= 0.3 sec, etc.
Filled pause	hmm, emm, etc.	
Audible inhalation, short / long	.h / .hh	≤ 0.5 sec / ≥ 0.5 sec
Audible exhalation, short / long	h / hh	≤ 0.5 sec / ≥ 0.5 sec
Laughter	ha, hha, haha, etc.	
Elongation	:	
Abrupt stop	-	
Other	(s); (c)	= smack; caugh
Inarticulate word or sound	*[italics], italics*	

Here is an excerpt from the transcript:

> eeh especially in the winter time we have a lot of snow (0.3) and actually also a lot of: .hh *all*- (.) *al*- very cold (.) in lapland (.) we have this lovely polar lights (.) which (.) are very exceptional .hh in finland (0.5) (s) .h then in the autumn (0.3) eeh (.) the colorfu- (0.2) colourful leaves: (.) are (.) are (0.6) beautiful .h (0.9) a:nd (.) and summertime (.) we ha- (.) we have w- warm and sunny days (.) and of course the: .h midnight summer when (.) the nightless .h (0.7) then is the nightless time h (0.5) (s) .h and in general (0.2) we have a beautiful nature in finland we are (0.2) we are called .h hund- (.) thousand (0.4) lakes country (2.7) eeh (.) the nature if- (.) is fresh (2.2) eeh we have a lot of: eeh (0.6) woods (4.4) a:nd (1.5) .h a:nd if we the (.) if we are discussing the (.) temperatures in finland they varies a lot depending what season .hh is in question (TT4, sample 1)

A digital sound editing software WavePad was used to measure the lengths of the pauses. The transcript was made as a closed transcript to selectively highlight the features of talk that were relevant for the analysis of the data. The features that were relevant to take into consideration in the transcription were those that indicated fluency, such as pauses, hesitations and other disfluencies of speech. Fluency of speech can be characterised by "perceptions of ease, eloquence, and smoothness" (Housen and Kuiken 2009, 463). On the other hand, speech that lacks fluency can be described as "slow and uneven", "hesitant", "jerky" and "disconnected" (Fulcher 2003, 30). There are particular observable speech behaviours, or subdimensions, that are associated with fluency (Fulcher 2003, 27):

- Hesitations consisting of pauses, which can be unfilled (silence) or filled (with noises like ‘erm’)
- Repeating syllables or words
- Changing words
- Correcting the use of cohesive devices, particularly pronouns

According to Biber and Quirk (1999, 1048) pauses, hesitations and repetitions are “quite normal” characteristics of speech as long as they do not interfere with understanding. These can occur in situations where the speaker needs to keep talking even though the mental planning happens slower and the “planning needs to catch up” (Biber and Quirk 1999, 1048). It is often difficult to interpret whether a disfluency is intentional or natural and when it should be seen as interfering with the fluency of speech. In the following the features that are included in the transcript are introduced: silent pauses, filled pauses, repetition, elongations and abrupt stops.

There are two types of pauses, unfilled or silent pauses and filled pauses. An unfilled pause is “a period of silence where the speaker appears to plan what to say next” (Biber and Quirk 1999, 1053). Pradas Macías (2006, 28) studied the role of silent pauses in fluency and argues that from a communicative point of view, the audience perceives a silence as “interrupting the flow of talk.” Filled pauses, on the other hand, are occupied by vowel sounds and are also called hesitators (Biber and Quirk 1999, 1053). Hesitators can be used, for instance, for signalling a wish to continue speaking (Ibid., 1092). Repetition or repeats can be used as strategies for trying to gain time by beginning and rebeginning the same piece of speech. A word is repeated until the speaker is able to continue. Repetitions can be unplanned, involuntary or deliberate, such as when a speaker wants to intensifying what is said or to get the attention of the listener (Ibid., 1056). Furthermore, elongations and abrupt stops are also considered disfluencies of speech (Jenks 2001, 59). They can be signs of “inner workings of the brain” (Ibid.). Apart from that, they can also denote interactional conventions and practices, maintaining speakership or “yielding conversational floor” (Ibid.). Elongation refers to lengthening of a word and extension of a sound, whereas abrupt stops are cut-off sounds that can occur anywhere during the talk (Ibid., 60).

4 Results and analysis

In this section the results of the study are presented and analysed. First, the results of the self-assessment and external ratings are presented. The external rating includes the ratings of the student raters and the expert rater. Secondly, the results of the self-assessment and the external ratings are compared to each other. Thirdly, some noteworthy observations from the ratings are presented with excerpts from the transcript of the actual language performances. The evaluation outcomes are presented in two forms according to the needs of the analysis: firstly, in detailed division of the reference level labels A1-A2-B1-B2-C1; and secondly, in three broad levels of the reference level labels: Basic User (including A1-A2), Independent User (including B1-B2) and Proficient User (including C1) (see Council of Europe 2001, 23). The reference level C2 is not included in the analysis of this study and therefore the highest level of rating can be C1 (Proficient User).

4.1 Self-assessment

The self-assessment was made in three categories: speaking skills, functional competence and comprehensive language proficiency. Speaking skills were divided into two sub-categories: spoken production and spoken interaction. Spoken production refers to speaking in front of an audience, such as giving a speech or a presentation (Council of Europe 2001, 58), whereas spoken interaction is spontaneous speaking and it usually concerns interaction, such as taking part in a conversation (Council of Europe 2001, 73). In the rating of the language performances, spoken production skills are represented in sample 1 and spoken interaction skills are represented in sample 2. Functional competence, on the other hand, consists of two generic qualitative factors, fluency and propositional precision (Council of Europe 2001, 128). Fluency tests the learner's ability to articulate and maintain the conversation despite difficulties, while propositional precision measures the learner's ability to make the meaning of the message clear. Furthermore, the test takers also evaluated their comprehensive language proficiency according to the global scale (Council of Europe 2001, 24). The global scale takes a holistic view of language use and includes all four language skills: reading, writing, listening and speaking. Table 3 below shows the results of the self-assessment in the four categories: fluency, propositional precision, spoken production and spoken interaction.

Table 3. Results of the self-assessment.

Test taker	Fluency	Propositional precision	Spoken production	Spoken interaction
TT1	C1	C1	B1	B2
TT2	B1	C1	B2	B2
TT3	B2	C1	B1	B1
TT4	C1	C1	B1	B1
TT5	B1	B2	B2	B2
TT6	B1	C1	B1	B1
TT7	B1	C1	C1	B2

The results show that the test takers were quite optimistic in their evaluation. None of the test takers evaluated their skills as Basic User (A1-A2) in any of the categories. The test takers were most optimistic in the evaluation of propositional precision, where six out of seven test takers chose level Proficient User (C1). Fluency and propositional precision were evaluated by responding to can-do statements (see section 3.2.2). The can-do statements for the self-evaluation can be found in Appendix 2. As an example, the statement for level Proficient User (C1) in propositional precision is:

> I can give opinions and statements precisely and express my certainty/ uncertainty, belief/doubt, etc.

In three categories (fluency, spoken production and spoken interaction) the majority of the test takers evaluated their skills as Independent User (B1-B2). In spoken interaction all seven test takers chose level Independent User (B1-B2). For example, the level B1 description for spoken interaction is:

> I can interact quite fluently and spontaneously, even with native speakers. I can take an active part in discussion in familiar contexts, presenting and supporting my views.

Overall, most of the group's self-assessments (68%) in all four categories concentrated on level Independent User (B1-B2). A summary of the results of the self-assessment is presented in figure 4a and table 4b below.

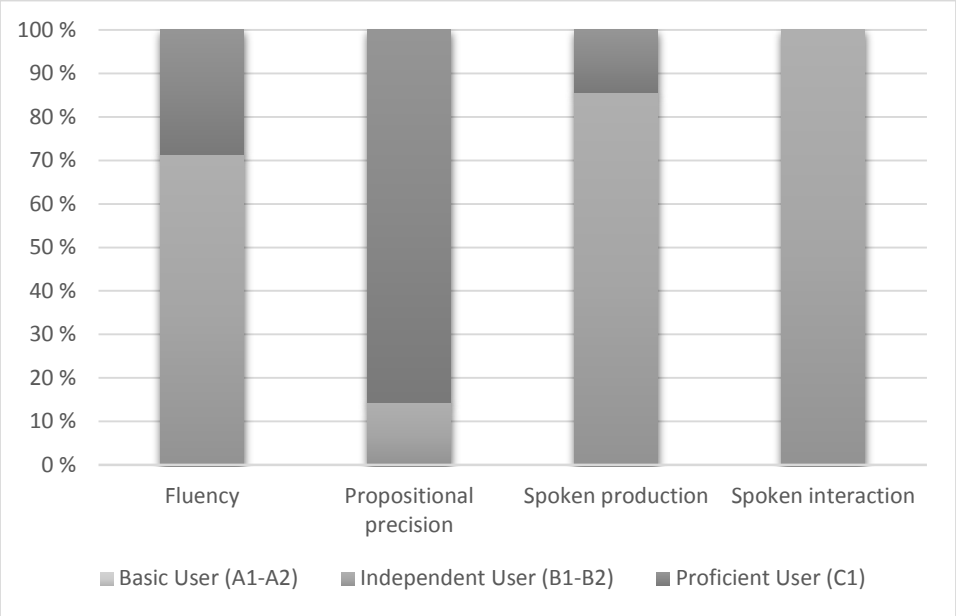

Figure 4a. Summary of the self-assessment results.

Table 4b. Summary of the self-assessment results in percentages (number of assessments).

Self-assessment	Fluency	Propositional precision	Spoken production	Spoken interaction	Overall
Basic User (A1-A2)	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Independent User (B1-B2)	86% (6)	100% (7)	71% (5)	14% (1)	68% (19)
Proficient User (C1)	14% (1)	0% (0)	29% (2)	86% (6)	32% (9)
Total	100% (7)	100% (7)	100% (7)	100% (7)	100% (28)

As a comparison, the self-assessment on comprehensive language skills against the global scale is presented in table 5 below. The self-assessment against the global scale was more moderate than in the other categories. All seven test takers chose level Independent User (B1-B2). The level descriptions for the global scale can be found in Appendix 2.

Table 5. Self-assessment in comprehensive language skills.

Test taker	Global scale
TT1	B1
TT2	B2
TT3	B1
TT4	B2
TT5	B1
TT6	B1
TT7	B2

4.2. External ratings

In this section the results of the external rating are presented and analysed. A total of 14 samples were rated, two samples by each of the seven test takers. Each sample was rated by two student raters and one expert rater in fluency and propositional precision separately. The results are presented in percentages and *rated samples*. A *rated sample* represents a sample that has been rated either in fluency or propositional precision. Therefore a *rated sample* can be, for example, "sample 1 fluency" by test taker TT1; or "sample 1 precision" by test taker TT1. Firstly, the results of the student ratings and the expert rating are presented. Secondly, the consistency of the external ratings is considered.

4.2.1 Student ratings

The performances of the test takers were rated by eight student raters in fluency and propositional precision. The samples were divided between the student raters so that each student rater evaluated both samples of either three test takers (TT1, TT2 and TT3) or four test takers (TT4, TT5, TT6 and TT7). This method was used to minimise the workload of each student rater and to prevent fatigue. Each student rater evaluated the samples either in fluency or in propositional precision. Therefore each sample was rated by four different student raters, two students rating in fluency and two students rating in propositional precision. Firstly, the evaluations of the student raters are presented and secondly, the consistency of the student ratings is analysed.

The results of the student ratings are presented in table 6a below. The colour codes in table 6b show, which samples were rated by which student raters. For instance, sample 2 fluency by TT4 was rated by student raters SR5 and SR6.

Table 6a. Results of the student ratings.

Test taker	Sample 1 fluency		Sample 1 precision		Sample 2 fluency		Sample 2 precision	
TT1	A1	A2	A1	A2	A1	A2	A2	A2
TT2	B2	B2	C1	B2	A2	B2	B2	B2
TT3	A2	B2	B1	B2	B1	B1	B1	B1
TT4	B1	A2	B2	B2	B1	B1	B1	B2
TT5	B2	B2	B2	B2	B2	B2	B1	B2
TT6	B1	A2	A2	B1	B1	B1	A2	A2
TT7	B2	B2	C1	B1	B2	A2	C1	B1

Table 6b. Colour codes for the student raters.

SR1	SR2	SR3	SR4	SR5	SR6	SR7	SR8

The total number of the rated samples was 56 (28 in fluency and 28 in propositional precision). A summary of the student rater evaluations can be seen in figure 7. Overall the student raters evaluated two-thirds (37 rated samples) of the test takers' performances as Independent User (B1-B2). Only 3 rated samples (about 5%) received the rating Proficient User (C1) and little less than one-third (16 rated samples) of the performances were rated as Basic User (A1-A2).

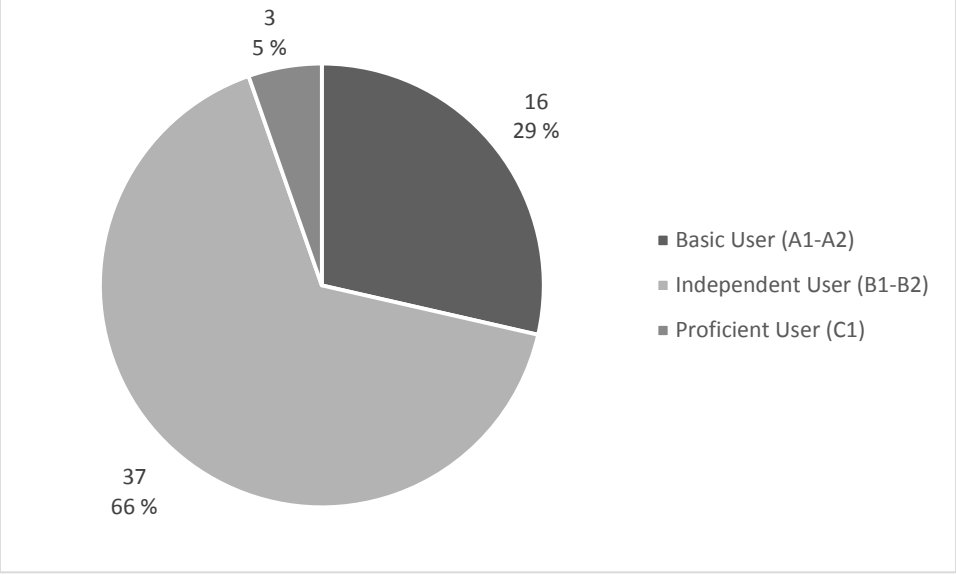

Figure 7. Overall results of the student ratings in rated samples and percentages.

The student ratings were analysed according to functional competence (fluency and propositional precision) and speaking skills (spoken production and spoken interaction). The results of the analysis are presented in figure 8a and table 8b below.

The analysis of the student ratings according to functional competence shows that the performances were rated slightly better in propositional precision than in fluency. The most noteworthy difference can be seen in level Proficient User (C1) ratings. Indeed, in propositional precision three samples were labelled as Proficient User (C1), whereas in fluency none of the performances were rated as Proficient User (C1). In contrast, the ratings on samples 1 and 2 were almost identical. In both samples about two-thirds of the student ratings were on level Independent User (B1-B2).

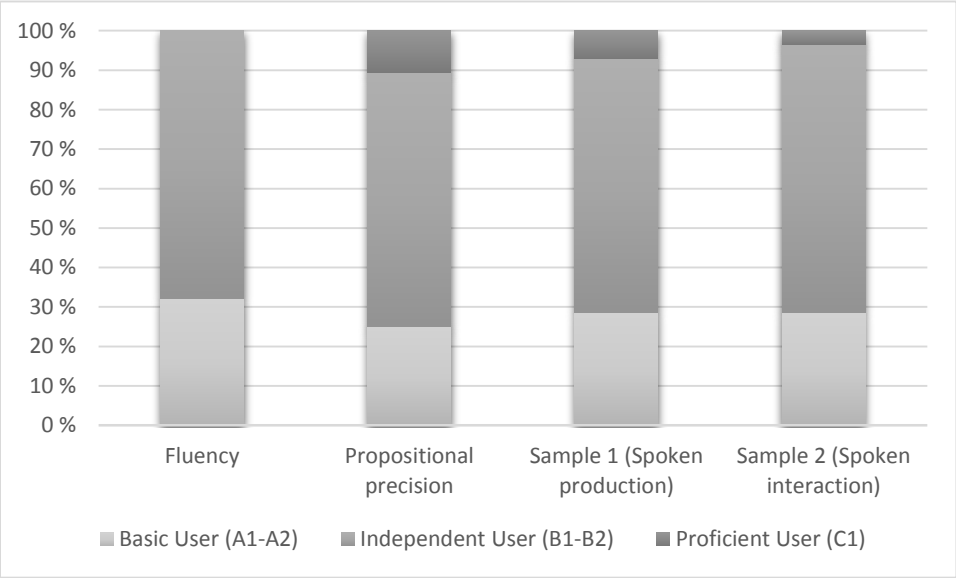

Figure 8a. Student rating according to functional competence and speaking skills.

Table 8b. Student rating according to functional competence as speaking skills in percentages (rated samples).

Student rating	Fluency	Propositional precision	Sample 1 (Spoken production)	Sample 2 (Spoken interaction)
Basic User (A1-A2)	32% (9)	25% (7)	29% (8)	29% (8)
Independent User (B1-B2)	68% (19)	64% (18)	64% (18)	67% (19)
Proficient User (C1)	0% (0)	11% (3)	7% (2)	4% (1)
Total	100% (28)	100% (28)	100% (28)	100% (28)

4.2.1.1 Consistency of student ratings

Each performance was rated by four different student raters. How well did the different ratings correlate with each other? The analysis is made according to the detailed division of the reference level labels: A1-A2-B1-B2-C1. Because each sample was rated in fluency and propositional precision separately, the total number of rated samples is 46. In the analysis, the two student ratings of each rated sample are compared to each other, and therefore the total number of *rated sample pairs* was 28 (14 pairs in fluency and 14 pairs in propositional precision). The results of the student rating consistency are presented in figure 9a and table 9b below. According to the analysis the consistency of the student ratings was fairly good. Nearly half (46%) of the student ratings on the test takers' performances were identical (for example B1 and B1). One level difference (for example B2 and C1) occurred in 10 (36%) rated sample pairs. The largest deviation in the ratings was two levels apart (for example A2 and B2). In five of the samples the rating differed with two levels. These samples will be discussed in detail in section 4.1.4.

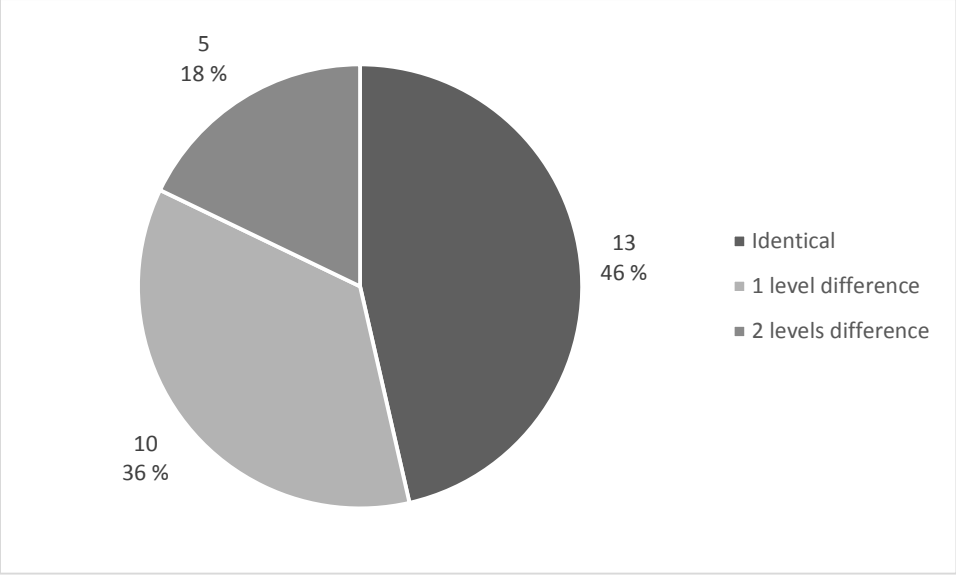

Figure 9a. Consistency of the student ratings in rated samples and percentages.

Table 9b. Consistency of the student ratings.

Consistency of the student ratings	Rated sample pairs (percentages)
Identical	13 (46%)
1 level difference	10 (36%)
2 levels difference	5 (18%)
Total	28 (100%)

4.2.2 Expert rating

The expert rater evaluated the performances of the seven test takers in both fluency and propositional precision. The results of the expert rating are presented in table 10 below.

Table 10. Results of the expert rating.

Test taker	Sample 1 fluency	Sample 1 precision	Sample 2 fluency	Sample 2 precision
TT1	B1	A2	B1	A2
TT2	C1	B2	C1	C1
TT3	B2	C1	C1	C1
TT4	B2	C1	B2	B2
TT5	B2	C1	C1	C1
TT6	B1	B1	B2	B1
TT7	B2	B1	B1	B1

The total number of rated samples was 28 (14 in fluency and 14 in propositional precision). The expert rater evaluated the majority of the performances as Independent User (B1-B2). The distribution of the overall expert rating according to the proficiency levels is presented in table 11 below.

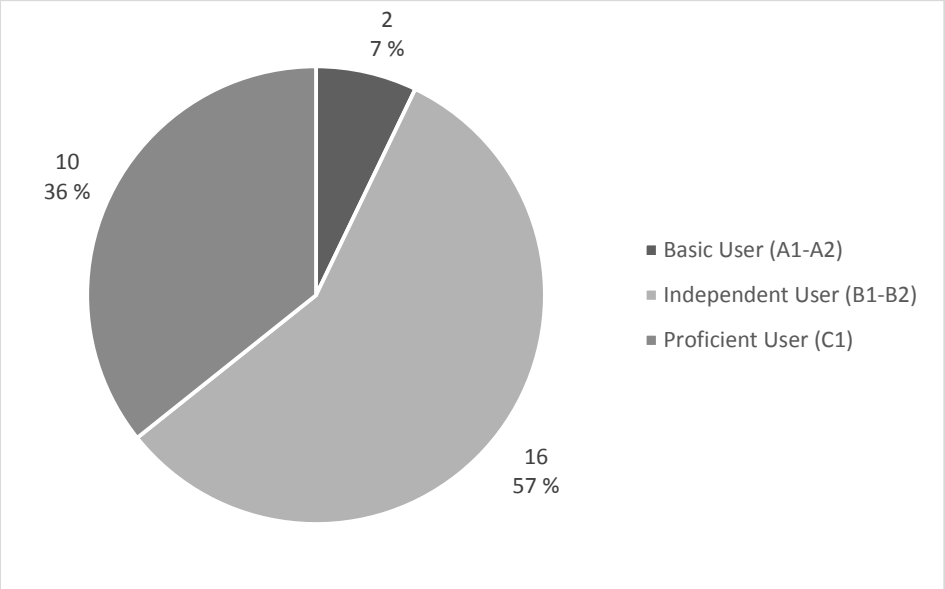

Figure 11. Overall expert ratings in rated samples and percentages.

The expert rating was analysed according to functional competence (fluency and propositional precision) and speaking skills (spoken production and spoken interaction). The results of the analysis are presented in figure 12a and table 12b below. In fluency the expert rater evaluated most performances (71%) as Independent User (B1-B2) and none of the samples were rated as Basic User (A1-A2). In propositional precision two of the samples (14%) were rated Basic User (A1-A2). Six of the samples received the rating Independent User (B1-B2) and six samples were rated as Proficient User (C1). To sum up, according to the expert rating there were more high-level performances in propositional precision than in fluency. In speaking skills sample 2, (spoken interactional skills) was rated slightly better than sample 1 (spoken production skills). Sample 2 was rated six times as Proficient User (C1), whereas sample 1 received the rating Proficient User (C1) only four times.

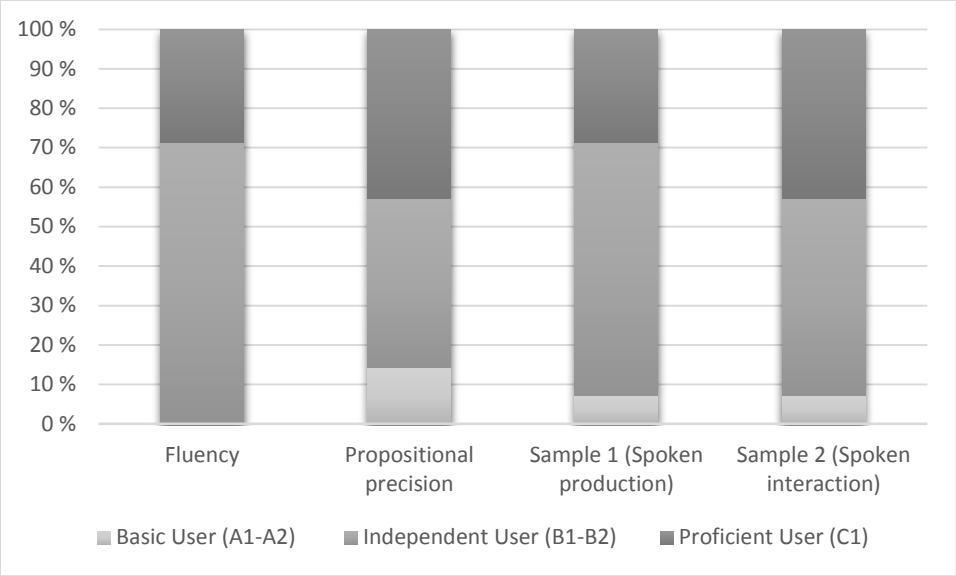

Figure 12a. Expert rating according to functional competence and speaking skills.

Table 12b. Expert rating according to functional competence and speaking skills in percentages (rated samples).

Expert rating	**Fluency**	**Propositional precision**	**Sample 1 (Spoken production)**	**Sample 2 (Spoken interaction)**
Basic User (A1-A2)	0% (0)	14% (2)	7% (1)	7% (1)
Independent User (B1-B2)	71% (10)	43% (6)	64% (9)	50% (7)
Proficient User (C1)	29% (4)	43% (6)	29% (4)	43% (6)
Total	100% (14)	100 % (14)	100 % (14)	100 % (14)

4.2.3 Consistency of external ratings

For judging the consistency of the external ratings the student ratings and the expert rating are analysed. In each sample the two ratings of the student raters and the expert rating are compared. The total number of rated samples was 28 (14 in fluency and 14 in propositional precision). The results of the analysis are presented in figure 13a and table 13b. The analysis shows that the consistency of the external ratings was weak because only in 3 out of 28 rated samples (11%) were identical. In majority of the samples (75%) the expert rating was more positive than either of the student ratings. Only in four rated samples (14%) the expert rater evaluated the performances on a lower

level than either of the student raters. It is noteworthy that the expert rater consistently evaluated the performances more positively than the student raters. Besides, even in the cases where the expert rater evaluated the performance more negative than one of the student rater, still the expert rating was either identical or more positive than the evaluation of the other student rater evaluating the sample.

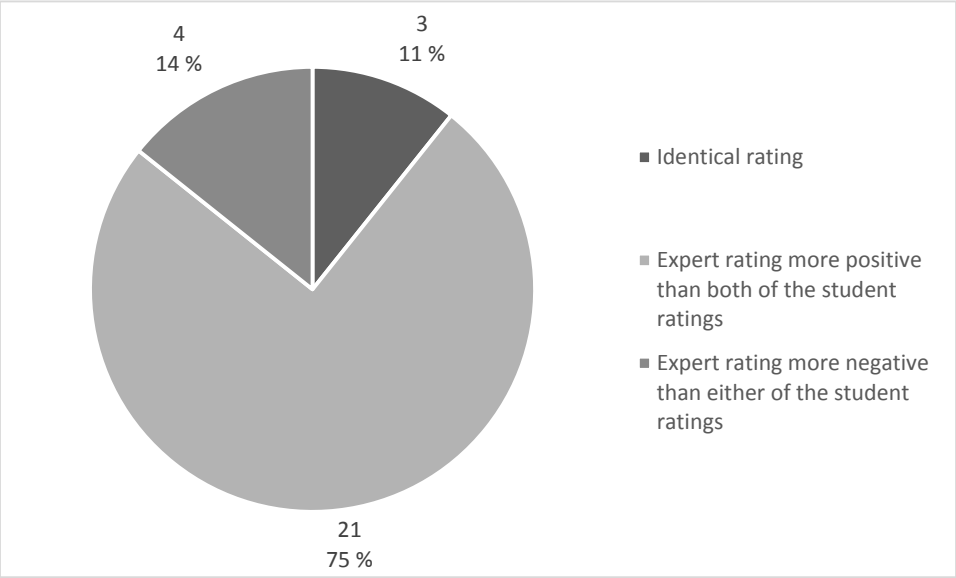

Figure 13a. Consistency of the external ratings in rated samples and percentages.

Table 13b. Consistency of the external ratings in percentages (rated samples).

Consistency of external rating	**Rated samples**
Identical rating	11% (3)
Expert rating more positive then both of the student ratings	75% (21)
Expert rating more negative than either of the student ratings	14% (4)
Total	100% (28)

4.3. Comparing self-assessment and external ratings

In this section the results of the external ratings are compared with the results of the self-assessment. The percentages and the number of rated samples according to the different evaluations are presented in table 14. The total number of rated samples varies between the evaluation methods according to the number of the raters and the rating criteria.

Table 14. Summary of comparison of different ratings in percentages (rated samples).

	Type of rating	Basic User (A1-A2)	Independent User (B1-B2)	Proficient User (C1)	Total
Fluency	**SR**	32% (9)	68% (19)	0% (0)	100% (28)
	ER	0% (0)	71% (10)	29% (4)	100% (14)
	SA	0% (0)	71% (5)	29% (2)	100% (7)
Propositional precision	**SR**	25% (7)	64% (18)	11% (3)	100% (28)
	ER	14% (2)	43% (6)	43% (6)	100% (14)
	SA	0% (0)	14% (1)	86% (6)	100% (7)
Sample 1 (spoken production)	**SR**	29% (8)	64% (18)	7% (2)	100% (28)
	ER	7% (1)	64% (9)	29% (4)	100% (14)
	SA	0% (0)	86% (6)	14% (1)	100% (7)
Sample 2 (spoken interaction)	**SR**	29% (8)	67% (19)	4% (1)	100% (28)
	ER	7% (1)	50% (7)	43% (6)	100% (14)
	SA	0% (0)	100% (7)	0% (0)	100% (7)

Table note: (SR = Student rating; ER = Expert rating; SA = Self-assessment)

Overall the different ratings were quite well in line with each other. As the figure 15 on the overall comparison of the different ratings shows, the majority of the performances were evaluated as Independent user (B1-B2) by all the different raters. The biggest difference in the evaluations was between the self-assessment and the student ratings. Indeed, the student rating was the most critical and the self-assessment was the most positive. In self-assessment none of the performances were rated as Basic User (A1-A2), while the student raters evaluated almost 30% of all the performances as Basic User (A1-A2). Furthermore, according to the self-assessment in comprehensive language proficiency against the global scale (see section 4.1), the test takers did not choose any Basic User (A1-A2) level labels.

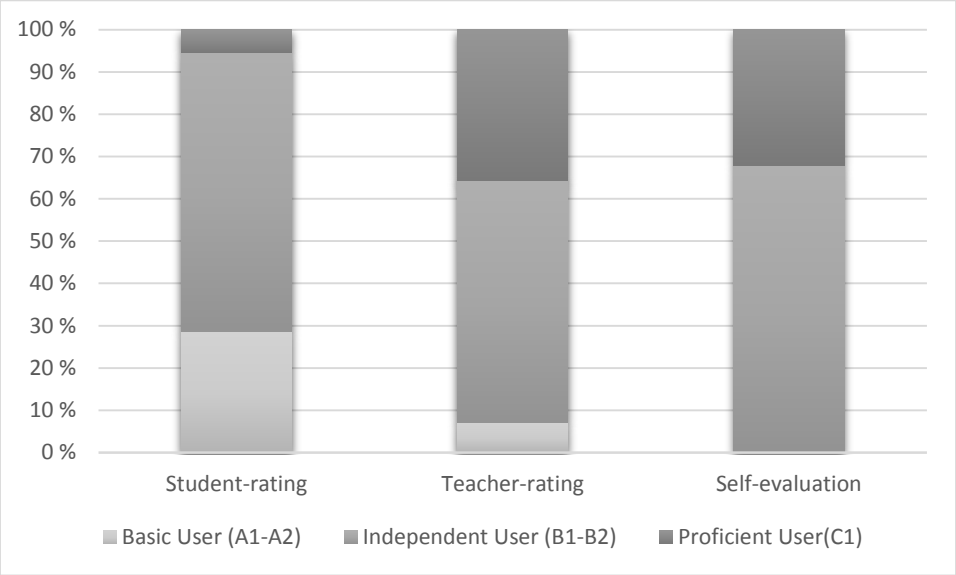

Figure 15. Overall comparison of the different ratings.

The analysis of the different ratings according to functional competence (fluency and propositional precision) can be seen in figure 16. In fluency all the raters evaluated the majority of the performances as Independent User (B1-B2). However, differences occur in the Basic User level (A1-A2) where student raters evaluated nearly 30% of the performances as Basic User (A1-A2), whereas the expert rater and the self-assessment rated none of the performances as Basic User (A1-A2). Overall, self-assessment was most in line with the expert rating with identical percentages. In contrast, the student rating was most critical in the rating of fluency. Indeed, it is noteworthy that in the rating of fluency the evaluations of the expert rater and the self-assessment were identical, while the student rating was clearly more critical. By contrast, the self-assessment was clearly most positively evaluated in propositional precision with more than 80% of the ratings on level Independent User (C1). As a whole, variation occurred among the different evaluations in propositional precision. Again, the student rating was most critical with the least performances rated as Independent User (C1) and a total of 25% of the performances rated on level Basic User (A1-A2).

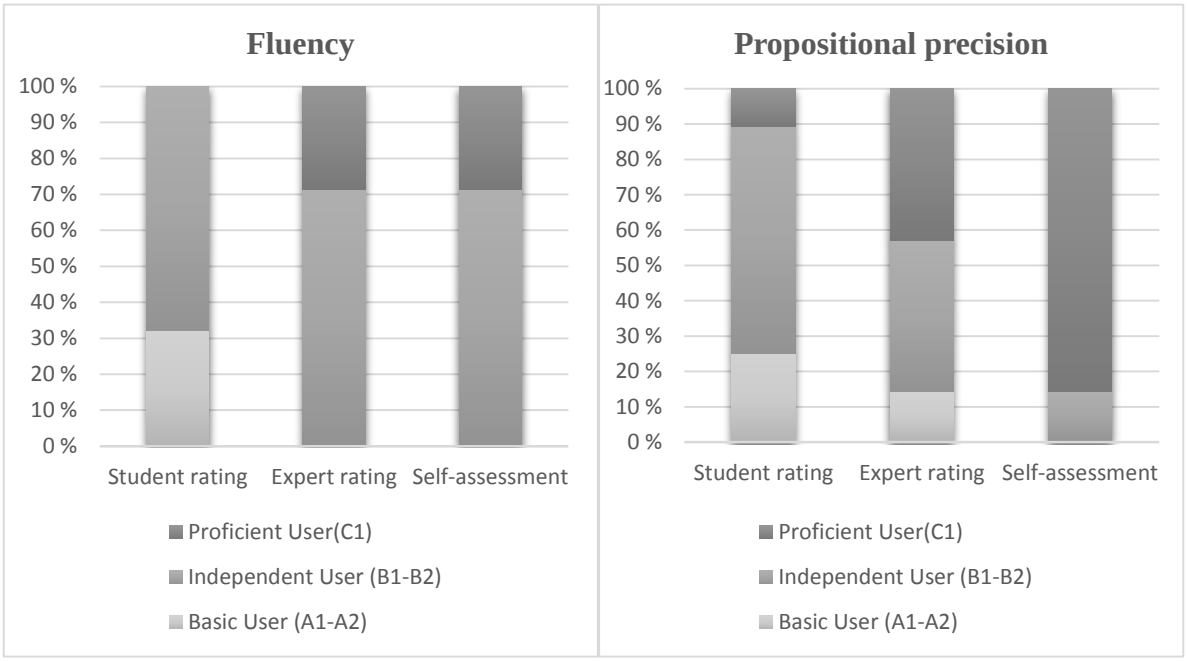

Figure 16. Comparing ratings according to functional competence (fluency and propositional precision).

The three different evaluations were also analysed according to speaking skills (spoken production and spoken interaction). Sample 1 represents spoken production skills and sample 2 represents spoken interaction skills. As figure 17 shows, both samples were rated most positively by the expert rater who evaluated more performances as Proficient User (C1) than the student raters or the self-assessment. To conclude, the student rating was most critical in the evaluation of speaking skills.

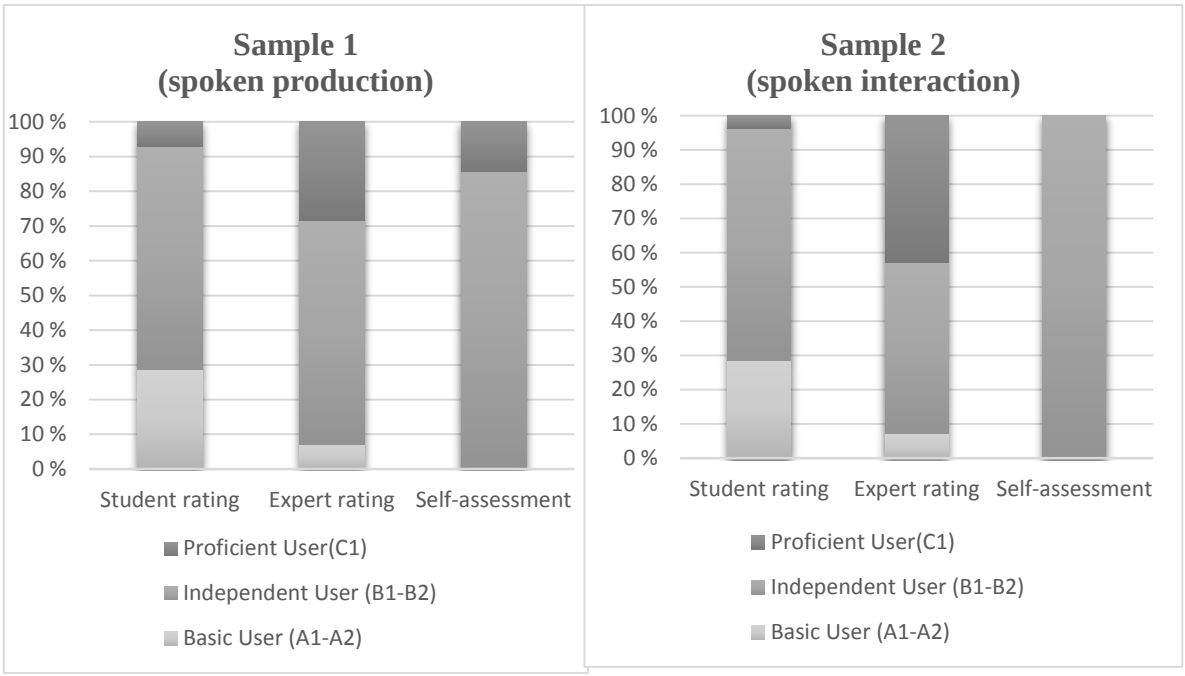

Figure 17. Comparing ratings according to speaking skills (spoken production and spoken interaction).

4.3.4 Who is a Proficient User?

The analysis of the results shows that the student raters were overall most critical in the evaluations while the expert rater evaluated the performances most positively. Overall, the expert rater evaluated 36% of the performances as Proficient User (C1) whereas student raters evaluated only 6% of the performances as Proficient User (C1). Which test takers' performances were rated as Proficient User (C1)? Table 18 shows which test takers received the rating of Proficient user (C1) in one or more rated samples. Interestingly, only one test taker's (TT2) performances were evaluated as Proficient User (C1) in all the different evaluation methods.

Table 18. Who is a Proficient User (C1)?

Test taker	Student raters	Expert rater	Self-assessment
TT1			x
TT2	x	x	x
TT3		x	x
TT4		x	x
TT5		x	
TT6			x
TT7	x		x

4.4 Some noteworthy observations

The following section presents some noteworthy observations in the evaluations of the performances. Seven examples have been chosen for a more detailed observation as shown in table 19a below. The examples contain performances from three test takers: TT7, TT2 and TT3. All eight student raters were involved in the rating of the chosen performances and the color codes for the student raters can be seen in table 19b. In the analysis of each observation the ratings of the two student raters and the expert rater on the rated sample are analysed together with the test taker's self-assessment in fluency or propositional precision, according to the sample. Also excerpts from the transcript as well as the raters' comments of the evaluation outcome are analysed. In addition, an evaluation by the researcher is presented. The excerpts are numbered and the key for the transcripts can be found in section 3.2.3.

Table.19a. Details on the noteworthy observations.

Observation	Test taker	Rated sample	Student rating 1	Student rating 2	Expert rating	Self-assessment
1	TT7	Sample 2 fluency	**B2**	**A2**	B1	B1
2	TT7	Sample 1 precision	**C1**	**B1**	B1	C1
3	TT7	Sample 2 precision	**C1**	**B1**	B1	C1
4	TT2	Sample 2 fluency	**A2**	**B2**	C1	B1
5	TT3	Sample 1 fluency	**A2**	**B2**	B2	B2
6	TT3	Sample 2 fluency	**B1**	**B1**	C1	B2
7	TT3	Sample 2 precision	**B1**	**B1**	C1	C1

Table 19b. Colour codes for the student raters.

SR1	SR2	SR3	SR4	SR5	SR6	SR7	SR8

Observation 1

The test taker TT7's performance in sample 2 fluency was rated by student raters SR5 and SR6 with a two level difference: SR5 granted level B2 (Independent User) and SR6 level A2 (Basic User) for the performance. Although the student raters had chosen a different label for the performance they both commented on similar criterion for the rating. Both student raters had observed features such as pauses, hesitations and intelligibility of the speech:

> "All speech was understandable, there were not many pauses and only little hesitation." (SR5)
>
> "Fairly many pauses and hesitation although the speech was understandable." (SR6)

Moreover, in the general comments on the ratings both student raters mentioned to have paid attention to similar features in the performances: pauses, hesitation and problems in pronunciation (SR5); the amount of pauses and hesitations, choice of vocabulary, language structures and intonation (SR6). In the transcript of the actual performance some of these features are clearly visible, such as hesitation:

(1) we have very .h very cold in (.) in winter and (0.3) and **eeh (.) a- e- (0.7) a-** (.) autumn so (TT7, sample 2)

In addition, long pauses occur occasionally, also in the middle of an utterance:

(2) well (0.3) we have some s- **(1.0)** well **(0.8)** eh (.) dangerous species (.) so to say .h eeh beasts (.) as wolves and **(0.5)** and (.) and bears (.) eh but- (TT7, sample 2)

(3) an:d the basic meaning of (.) of e- sauna **(0.5)** is **(0.4)** eeh **(0.9)** to: (.) em **(0.8)** to be a place to: relax (TT7, sample 2)

The expert rater evaluated the performance as B1 (Independent User) which differed from either of the student raters' evaluation. The expert rater commented on the sample:

> "Simple language which causes much hesitation but [the speaker] is able to keep going without having to start from the beginning."

The expert rater evaluated the samples both in fluency and propositional precision. For this reason the general comments on the ratings can be applied to both categories, fluency and propositional precision. The expert rater commented that the rating scales for fluency and propositional precision guided his evaluation process and that he was doing his best to observe the features that were mentioned in the rating scale descriptions. However, the use of vocabulary and the fluent flow of speech were features which the expert rater observed instinctively. The researcher's own rating on TT7's sample 2 fluency is B2 (Independent User) on the basis that the speech was diverse and rich in nuances. Spontaneous expression was partly very fluent although some hesitation was observable. However, many of the pauses seemed natural rather than expressing uncertainty, and they could be interpreted as for pausing to think of the exact expression to describe what is being said:

(4) and (0.6) .h my (.) hobbies are em **(0.8) artistic** (0.2) .h i (0.2) draw an:d eeh (0.5) (s) eeh also paint (0.3) (TT7, s2)

The test taker TT7's self-assessment on fluency was level B1 (Independent User) and it agreed with the expert rating.

Observations 2 and 3

The test taker TT7's performances in sample 1 propositional precision and sample 2 propositional precision were rated by student raters SR7 and SR8. SR7 rated both performances as C1 (Proficient User) and SR8 rated both performances as B1 (Independent User). SR7 thought the language use in both samples was fluent and variable, whereas SR8 commented on the lack of details and limited use of vocabulary. SR8 also mentioned that the test taker was able to express the main points of the message but that the argumentation was limited. In general comment about the ratings both student raters had observed similar features in the performances: SR7 reported to have paid attention to precision of articulation and grammatical fluency, while SR8 had considered the use of vocabulary, fluency of speech, versatility and language specific expressions. In the transcript of the sample, the lack of fluency is occasionally quite visible with pauses, hesitations and repetitions:

(5) eh the average of winter **temp- (0.4) temperature** is minus (0.3) ten (**1.0) (s) (.) eeh (0.6**) **degrees celsius degrees (0.4) so (1.1) hhm (s)** (TT7, sample 1)

Variability of the language and the lack of details was also mentioned by the student raters. In the following excerpt from sample 1 the test taker talks about Finnish nature but the narration is quite general and it does not include many details:

(6) so a nature var- (.) varies quite much (.) in different parts (.) of finland (0.3) .hh and eeh (0.2) eeh (0.3) we have (0.3) many (.) lakes here in finland (.) so eeh (.) d- (0.5) the special (0.2) landscape .h is eem (0.7) (s) filled (.) with (.) lakes .hh and (.) well (.) in some part of (.) finland .hh eem (0.2) rivers (0.4) (s) .hh (TT7, sample 1)

In contrast, when talking about the different seasons in Finland, a fairly detailed description on different winter activities is included:

(7) we (0.4) we enjoy all (0.4) kinds of activities .h which are related to: .h seasons (1.5) and em (0.6) (s) well (.) we have (0.9) especially (.) winter activities (.) skiing skating .hh eem (0.9) sledge (0.8) driving and eem (1.2) and a- (0.3) eem (0.5) well (0.8) building snow castles and (.) making snowman (0.9) (s) (.) and so on (0.5) .hh (KP, sample 1)

The expert rater evaluated both samples as B1 (Independent User) and agreed therefore with the rating of SR8. The expert rater commented on the simplicity of the language use, limited use of vocabulary and monotony. The researcher's evaluation for propositional precision in sample 1 is level B2 (Independent User) because the speech was clear and varied occasionally. In addition, there were more details in the description towards the end. Only few grammatical errors occurred but none of them interfered with the intelligibility of the message:

(8) in some part of (.) finland (*parts*) (TT7, sample 1)

(9) making snowman (*snowmen/a snowman*) (TT7, sample 1)

In sample 2 propositional precision the speech was clear but not as detailed as in sample 1, so the researcher's evaluation on the performance is B1 (Independent User). The test taker TT7's self-assessment on propositional precision was C1 (Proficient User) and it agrees with the rating by SR7.

Observation 4

The observations 4, 5 and 6 were rated in fluency by student raters SR1 and SR2. Overall in the evaluation of fluency SR1 commented to have considered the sentence structures, choice of words, hesitation and fluency of speech; whereas SR2 had regarded pauses, repair, hesitation, vocabulary and structures. Observation 4 was test taker TT2's performance in sample 2 fluency. SR1 rated the performance as A2 (Basic User), while SR2 rated the performance as B2 (Independent User). Also the raters' comments on the rating differed slightly. SR1 mentioned that the test taker used too simple vocabulary and had incorrect pronunciation, whereas SR2 commented on the fluency of speech and focused especially on hesitation. Some hesitation can indeed be noted also in the transcript of the performance:

(10) but we have a **lo- ah (1.1)** a bears that are not **hh eh (0.3)** is it called a brown bear (0.5) (TT2, sample 2)

The expert rater evaluated the performance more positively than either of the student raters: C1 (Proficient User). In the comments the expert rater noted some stammering in the beginning of the performance but after a while the speech was very natural, partly due to easy topics. The researcher rates the performance as B2 (Independent User) because the expression was quite clear and there were no long pauses and only very little hesitation, which appeared as if it was the test taker's natural way of speaking also in general. The self-assessment of TT2 on fluency was B1 (Independent User) and it differed from all the external ratings.

Observation 5

Observation 5 was TT3's performance in sample 1 fluency and it was rated by student raters SR1, who rated the performance as A2 (Basic User); and SR2, who rated the performance as B2 (Independent User). SR1 commented that the performance contained easy structures and had "lots of errors". SR2 also paid attention to some

problems in the formulation of speech but altogether considered the speech "really fluent". In the actual performance some hesitation and pausing can be noticed:

(11) we have a m- .h lot of **foo- (.) woods** an:d thousands of lakes .h a:nd **(0.9) (s) (0.4)** of course eem (.) we do have a (.) four seasons (TT3, sample 1)

(12) .hh a:nd (.) **it's (.) it's** a lot **(0.9) (s) yes: .h what else i (0.2) can i say (.)** about finland (0.2) (TT3, sample 1)

In addition, some occasional problems in the formulation can be detected:

(13) we have over eighty thousand (.) summer cottages (0.9) in eh (.) one hour (1.4) .h **how i *said* it (.) one hour h car ride (1.1) if i try to (1.1) by car (0.2) from tampere** (TT3, sample 1)

The expert rater evaluated the performance as B2 (Independent User) and commented that the speech seemed nervous in the beginning of the performance, which came across as lack of fluency. However, the expert rater considered the use of vocabulary very good. The researcher rated the performance as B2 (Independent User) because the speech was uninterrupted and sounded natural. In addition, it contained only flew pauses and repetitions. TT3's self-assessment on fluency was B2 (Independent User) and it was in line with the rating of SR2 and the expert rater.

Observation 6

The test taker TT3's performance in sample 2 fluency was rated by student raters SR1 and SR2. Both student raters evaluated the performance as B1 (Independent User). However, the expert rater evaluated the performance two levels higher, as C1 (Proficient User). The student raters commented that the speech was understandable although there were some errors. In general the student raters considered the speech fluent although some hesitation in the formulation of speech can be noticed:

(14) but .hh maybe one hundred years ago **its a- (0.7) it was a- (0.2) also (1.3) (s) about eh** hygenia (TT3, sample 2)

There was also an attempt to use an idiomatic expression (*melting pot*):

(15) i have many (0.5) favourite places but maybe: the number one is israel .hh a:nd i think it's eeh (0.8) it's a some kind of **a (1.1) [*smelting oven*]** (0.2) for the (.) whole (1.0) whole (0.5) world (0.5) the people are coming all (0.9) all around the world (TT3, sample 2)

Although the expert rater evaluated the performance as Proficient User (C1), he described the speech as "bouncing". However, due to the familiarity of the topics the narration was fluent and rich in nuances. The researcher rates the performance as B2 (Independent User) because there were only few pauses and hesitations, and the speaking was nuanced. The self-assessment of TT3 on fluency was B2 (Independent User).

Observation 7

Observation 7 was TT3's performance in sample 2 propositional precision and it was rated by student raters SR3 and SR4. Overall in the ratings SR3 concentrated in clarity of expression, intelligibility, complexity of ideas and complexity of structure. SR3 also mentioned not to have paid so much attention to linguistic errors. SR4 considered pronunciation of words, pauses and "searching for words", formulation of sentences, intelligibility of language and the extensiveness and details in the narration. Both student raters gave identical ratings for the sample: B1 (Independent User). In the comments SR3 mentioned that there were some problems with making the message understandable, especially when expressing more complicated ideas or with unfamiliar topics. Correspondingly, SR4 commented that the test taker was able to tell about personal topics fairly well, but the topics on the nature and sauna were slightly more challenging. Similarly to observation 6, the expert rater evaluated the performance two levels higher than the student raters: as C1 (Proficient User). The expert rater's comments of the rating were similar to those of the student raters, although the outcome of the evaluation was different: With familiar topics the narration was fluent but with more complicated topics some problems occurred. The researcher's rating for the performance is B1 (Independent User) because the main ideas are presented clearly but the narration lacks detail. In some occasions the test taker had to rely on other languages than English, as for the word *lynx*:

(16) .h oh (.) we do have eeh (0.2) **how i say it [*ilves*]** it's a: (.) big cat (.) its **a [*lynx lynx*] (0.3) in latin** (.) .h a:nd (0.5) (TT3, sample 2)

The test taker TT3's self-assessment on propositional precision was C1 (Proficient User) which agrees with the expert rater's evaluation.

5 Discussion

In this section the main results of the study are summarised and discussed. The section also raises issues about the limitations of the study, considers the appropriateness of the methods chosen for the study, and introduces interesting topics for further research.

The first research question concerned the similarities and differences in the external rating: *Q1 - How consistent is the evaluation among the external raters?* Firstly, the results reveal that the consistency of the student ratings was good. According to the analysis nearly half of the performances were rated identically by the student raters. By contrast, the consistency between the student ratings and the expert rating appears to be somewhat weaker. The results show that the expert rater evaluated the performances notably more positively than the student raters. The weak correlation in the external rating supports the idea that all evaluation is subjective.

A closer analysis on the performances that most caused deviation in the external rating suggests that in some cases there were differences in the interpretation of the rating scale descriptions. The comments on the ratings indicate that the student raters paid more attention to details, while the expert rater took a more general approach to the evaluation. The expert rater commented that he closely observed the rating scale descriptions in the evaluation although he admitted that it was difficult to exclude features that were not explicitly mentioned in the rating scales, such as grammaticality or the use of vocabulary. Nevertheless, it seems that the student raters were even more prone to include criteria that were not indicated in the rating scales. However, it should be noted that especially the description of the rating scale for propositional precision appears to give much room for different interpretations. Furthermore, also language professionals have criticised the inadequacies in the definitions of the level descriptions in CEFR (Figueras 2012, 483). Indeed, clear and understandable guidelines in evaluating language proficiency are essential (Underhill 1987, 23). It would be interesting to make a more detailed analysis on the different interpretations of the rating scale descriptions among the raters and that would indeed make a valid topic for further research.

The consistently critical tone in the student ratings is somewhat surprising. Is being critical a typical feature for raters with little or no experience in language evaluation or is the tendency more connected with the teacher identity that is still developing in the students? Although the students have acquired plenty of theoretical knowledge on testing language proficiency, they still lack experience of the actual

practice of evaluation. The expert rater, on the other hand, is able to be realistic about what can be expected from the language learners as a whole and put that into perspective with the criteria of the rating scales. The deviation in the ratings can also be explained by personal differences. Each rater is an individual and their personality plays a part in how critical a view they take in the evaluation. The positive evaluation of the expert rater could therefore indicate that his personality is more relaxed in general.

The second research question concerned the correspondence between the self-assessment and external rating: *Q2 - To what extent does the test takers' self-assessment correlate with the external rating?* The results suggest that the correlation between the self-assessment and the external rating was fairly good although some deviations did occur. Overall the self-assessment correlated best with the expert rating although the self-assessment was slightly more optimistic in general. The evaluations according to functional competence (fluency and propositional precision) were diverse. In the evaluation of fluency the correlation was especially good since the results of the self-assessment and the expert rating were identical. However, the evaluations varied the most in the rating of propositional precision where the student raters evaluated the performances most critical, while the self-assessment was most optimistic. On the other hand, the expert rating appeared to be more positive than the self-assessment in the analysis of the different ratings according to speaking skills (spoken production and spoken interaction), whereas the student rating still remained most critical.

The fairly good correlation between self-assessment and external rating in this study is in line with the results by Leblay (2013, 230) and Ross (1998, 16), who likewise found perceptible correlation between self-assessment and external rating. In the light of the results of this study and previous research, self-assessment appears to be a reliable method of evaluation. Indeed, self-assessment can contribute to language learning by increasing learning motivation (Bullock 2011, 119), but it could also be used for learning outcome evaluation. Furthermore, it could be suggested that self-assessment could be used as a supportive method for judging language performances. Although it is not necessarily recommended that self-assessment should replace external rating (Little 2005, 335), self-assessment could still contribute to providing a broader understanding of the language proficiency of the learner by adding one more perspective to the whole (Underhill 1987, 22). If self-assessment would, however, be used to evaluate the language proficiency level of a student, extra attention should be given to selecting or formulating the rating scales so that they would be as clear and

easy to understand as possible. As discussed earlier, the interpretation of the rating scales can be challenging, even for language professionals. A trained language teacher, however, is in a better position to apply the scales realistically than a language learner with no training or knowledge in working with scale descriptions. On the other hand, this view has been challenged by Leblay (2013, 46) who argues that despite good intentions, not all language professionals have appropriate training for carrying out proficiency evaluations. Nevertheless, if self-assessments were included in the evaluation process of, for example, the final grade of a course, they should be used with caution to avoid the effects of over-rating one's skills for their own benefit (Douglas 2010, 75; Ross 2009, 3).

Furthermore, the findings of this study support the results by Brantmeier et al. (2012, 153) who argued that self-assessment works well with advanced students. The goals of self-assessment, such as enhancing learning (Ross 2009, 3) and contributing to learner autonomy (Douglas 2010, 75; Little 2005, 322), are especially suitable for adult learners who are likely to possess internal motivation to learn the language, such as potential achievements in the workplace with the help of better language skills.

5.1 Simulated language test

The language simulation turned out to be a useful method for collecting data for the study. The computer technology enabled the raters to listen to the audio recordings of the simulation performances multiple times independently of time or place. The simulation also seemed to be a practical method for testing speaking skills in a workplace context. The third research question was related to the simulation test: *Q3 – How convenient is the language simulation concept for testing and assessing oral proficiency?* The benefits of the simulation in testing speaking skills were numerous. The simulation enabled a reliable and consistent way for testing language proficiency (El Hmoudova 2013, 411). One of the biggest advantages of the simulation, like computer assisted language testing (CALT) in general, is that it is easy to access and can be delivered to a large number of test takers at the same time, independent of time and place (Jamieson et al. 2013, 290), provided that the necessary equipment are available. The simulation can also be repeated as many times as needed, which improves the reliability of the test. A further aspect for language testing is fairness among the test takers (Jamieson et al. 2013, 290). In relation to test fairness, both

Bachman (1991, 101) and Okada (2010, 1648) emphasise the importance of the interlocutor for the test taker's performance. Indeed, since the simulation is identical for all test takers, it is not impartial towards the test takers and the fairness of the test is better than in the commonly used speaking tests where the test takers work in pairs or with a trained partner. Furthermore, the simulation made it possible to create an environment where the test takers could apply the skills they have learned to a new and unknown communication situation (Crosling and Ward 2002, 54). The topics of the tasks and the nature of the instructions allowed language use in different proficiency levels and therefore the test takers were able to use language according to their own skills.

Learners usually have positive attitudes towards working with computers (Álvarez and Laborda 2011, E136) and it can even increase motivation for learning (Lee 2000, under the headline "Why use CALL?"). Unfortunately this study did not include the test takers' opinions on the simulation. Indeed, it would have been interesting to include a short questionnaire or an open interview on the test takers' experiences on the simulation as a language testing tool. One of the objectives of the simulation concept is to create a relaxed and 'untest-like' situation for testing language proficiency (Haataja 2010, 187-189). According to Tuuna-Kyllönen's (2011) study on the use of the LangPerform concept, adjusting to working and interacting with a computer could indeed be a potential source of distraction. Furthermore, Brown (1997, 48) argued that working with computers can cause anxiety. Indeed, it would have been interesting to find out how the test takers experienced the testing situation.

A further issue that would have been interesting to discuss with the test takers is the question of authenticity of the simulation. Authentic language material plays a central role in good workplace language assessment because it increases the validity of the test (Tratnik 2008, 7; Chapelle 2003, 28). However, often merely the sense of authenticity is enough, even if the situations as such would not be truly authentic. Did the simulation create a sense of authenticity and did the test takers think that the situations in the test resembled real-life events? In order to create an authentic testing environment in the simulation test, the tasks were designed to resemble real-world situations and native speakers were included as actors. Moreover, involving native speaker input in a speaking test could often be extremely challenging and expensive without the help of computer technology.

The simulation is especially suitable for testing communicative language

performances. On the other hand, exactly this type of language use can be challenging to implement on computer (Brown 1997, 45). Indeed, true interaction is never fully possible when working with a computer. For example, the technology does not allow asking questions or using other communication strategies which would be essential for real interaction to take place. However, if technology ever develops to allow the point where such interaction is possible, the computer simulations could be used on a whole new level. Indeed, it would be interesting to find out how the developing technology could be used to promote spoken language testing and learning in the future. The technical barriers of the simulation can also concern the implementation of the test. Because of some technical problems with the computers as well as the program itself, a number of performances could not be included in this study. Some of the problems were probably caused because the program was still in a beta stage at the time of the study. However, other problems, such as dysfunctional computer microphones, could have been avoided with more careful preparation.

The simulation test was designed mindful of the fact that it could also be used for other purposes outside of this study, for example in the workplace context. Being aware of employees' language skills is in the interest of the employers because linguistic competence plays an important role in the financial success of the organisations (Hellekjær 2007, 6; PIMLICO 2011, 10). Indeed, investing in language skills can be a valuable asset when competing in the market (ELAN 2006, 57). In addition, good language skills can be the decisive factor, for example, in promotion or recruitment processes (Lehtonen and Karjalainen 2008, 495). The simulation meets the needs of effective workplace language assessment (Tratnik 2008, 7) as it is, unlike oral language assessment in general (Álvarez and Laborda 2011, E136), easy, fast and cost-effective to administer. Although the process of making the simulation can be somewhat laborious and time-consuming, a ready-made simulation could be easily put into use, for example, with a one-time licence fee. Indeed, the simulation can be used multiple times and for different groups, even simultaneously, if needed. Furthermore, the simulation concept is a complete package as it includes, in addition to the speaking test itself, an easy access to the test results through an online language lab which enables, for instance, external rating and giving relevant and accurate feedback on the performances. The LangPerform concept is very flexible and each test can be tailored according to the specific needs of the organisation. Indeed, such workplace specific language training methods are needed (Lockwood 2012, 111). Moreover, the use of the

simulation is not only restricted to language testing but it could also be used to support language learning and teaching, or mapping the needs for further language training. Furthermore, it would be interesting to use the simulation concept, for instance, to investigate which language needs the graduating students will have in the workplace.

5.2 Rating of samples

Using CEFR as the framework for the study was a reliable choice because CEFR has been successfully used in language testing in general as well as in assessing workplace language skills (O'Loughlin 2008, 77). In addition, CEFR provides clear level descriptions which are empirically developed (Figueras 2012, 480). In the workplace clear and understandable communication is required to "get the job done" (Kankaanranta and Louhiala-Salminen 2010, 205). This view was taken into consideration already in the design of the simulation test. More explicitly it contributed to choosing functional competence (fluency and propositional precision) as the main criteria for the evaluations in the study. Indeed, Amos (2012, 457) mentions fluency and precise expression among the features that are essential for clear communication of the message. Therefore, the rating scales for functional competence were seen to represent the practical approach to language use that the workplace context demands.

The choice for using functional competence for measuring oral proficiency in this test could be criticised because it does not take a holistic view of language competence, but instead only concentrates on some aspects of the language leaving others aside. However, the rating scales for functional competence were chosen to guide the raters' attention away from linguistic correctness and towards more important aspects of workplace communication, such as the above mentioned clarity of communication and the content of the message (also Kankaanranta and Louhiala-Salminen 2010, 207). It should also be noted that although the linguistic ability of the test takers was not at the centre of attention, it was, however, being judged in an indirect way. Indeed, this argument can be supported by Littlewood's (1992, 41–42) four level model for language production. The model displays the extent to which precise expression of the message and fluent outcome of the speech can only be attained if the speaker has enough knowledge of the right word forms and sentence structures of the language. Without automated language ability (Fulcher 2003, 24), spontaneous speech in real time (Biber and Quirk 1999, 1048) could not be achieved. The comments of the

raters in their evaluations suggest that the choice of the rating scales was successful. Although some raters mentioned that they included features, such as the use of vocabulary or grammaticality, in their rating criteria, the focus of the evaluation was clearly on the 'intelligibility' and 'main points of the message', as expressed by the rater's own words. A natural way to continue with the study would be to examine, for example, how the statistical differences of the fluency features are displayed in the actual language performances, and how they correlate with the external rating. Some work has already been initiated in the process of this study but was not included in the analysis; for example, features of fluency are included in the markings of the data transcription.

The raters in this study were asked to comment on the sample length in judging functional competence. The comments were overall positive, for example: "Suitable length."; "The length was suitable, perhaps even a shorter sample (1 min 45 sec) would have been enough." Some comments mentioned that the sample length was fine but also recognised the benefits of a longer sample. Indeed, in a longer sample the test taker would have been able to speak more relaxed after having time to 'warm up' first. Moreover, it would have been interesting to add a task in the simulation with a longer speaking time, such as a free speech section, and in this way to see whether the length of the task really had an effect on the language performance. On the other hand, the short samples had the benefit that the total duration of the language test remained reasonable.

The simulation tested general language ability in a workplace environment and two different contexts of spoken language use were included in the study: formal and informal communication. This was done because the context affects our language use (Bachman 1991, 82–83) and learning to communicate in a flexible way in a given context is important (Louhiala-Salminen and Kankaanranta 2011, 247). For judging language use in these two contexts the rating scales for spoken skills (spoken production and spoken interaction) were chosen. Spoken interaction tested spontaneous communication skills. Indeed, informal conversations and work-related discussions are not only among the most common forms of oral communication in the workplace (Crosling and Ward 2002, 53) but they are also considered the most challenging ones (Charles 2007, 272). Spoken production was tested in a formal presentation task, and indeed, formal language use is also a part of everyday communication in the workplace (Crosling and Ward 2002, 41; Moslehifar and Ibrahim 2012, 530). In this study the two

contexts of language use provided an interesting viewpoint for the analysis of the data. However, a more thorough examination on language use in different contexts would make an interesting topic for further research. For example, the data of this study could provide material for the research on the use of vocabulary in formal and informal communication situations.

6 Conclusion

The objectives of this pilot study have been twofold. Firstly, the study has examined the evaluation of oral proficiency by comparing the results of self-assessment and external rating in oral performances. Secondly, the study has considered how a computer-based language simulation test based on LangPerform concept could contribute to testing and assessing oral language skills in the workplace context. Because of the binary nature of the study both the analysis of the results as well as the procedures and methods for gathering data have been equally important. The study has revealed a fairly good correlation between self-assessment and external rating. However, the consistency of the evaluations by the external raters has appeared to be somewhat weaker. In the light of the results it has been argued that all evaluation is subjective and the interpretation of the rating scale descriptions play a central role in the evaluation process. Furthermore, the study has also suggested that self-assessment could be used as an additional method for evaluating language performances, especially among advanced language learners.

The study has also found that the computer-based language simulation test could be a practical tool for testing language skills in the workplace context. The simulation test has appeared to enable authentic language input and to increase the fairness and reliability in oral proficiency testing. In addition, the test seems to be fast and easy to use. However, technical barriers, the lack of genuine interaction and time-consuming creation of the simulation have been noted as downsides of the test. Nevertheless, the simulation has been found to be suitable for testing speaking skills in the workplace context because of its flexibility, reasonable expenses and easily accessible evaluation and assessment methods.

This pilot study has been a small-scale study. The number of participants has been limited and therefore the results of the study should be interpreted as indicating tendencies rather than suggesting universal principles. For this reason any generalisations on the basis of the results of this study should be done with careful consideration. It would be interesting to conduct a larger-scale study with more participants on similar topics. Hopefully the study could provide useful information and inspirational ideas for further research on the use of language simulations and/or self-assessment in the field of language proficiency testing.

References

Amos, Paran. 2012. “Language Skills: Questions for Teaching and Learning.” *ELT Journal* 66, 4: 450–450.

Alderson, J. Charles. 2007. “The CEFR and the Need for More Research.” *The Modern Language Journal* 91: 659–663.

Alderson, J. Charles and Ari Huhta. 2005. “The development of a suite of computer-based diagnostic tests based on the Common European Framework.” *Language Testing* 22, 3: 301–320.

Álvarez, Miguel Fernández and Jesús García Laborda. 2011. “Teachers' interest in a computer EFL university entrance examination.” *British Journal Of Educational Technology*, 42, 6: E136–E140.

Bachman, Lyle F. 1991. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Biber, Douglas and Randolph Quirk. 1999. *Longman grammar of spoken and written English*. London: Longman.

Birjandi, Parviz and Saeideh Ahangari. 2008. “Effects of Task Repetition on the Fluency, Complexity and Accuracy of Iranian EFL Learners' Oral Discourse.” *Asian EFL Journal* 10, 3: 28–52.

Brantmeier, Cindy, Robert Vanderplank and Michael Strube. 2012. “What about me?: Individual self-assessment by skill and level of language instruction.” *System* 40, 1: 144–160.

Brown J. D. 1997. “Computers in Language Testing. Language Learning and Techonlogy.” *Language Learning & Technology* 1, 1: 44–59. Available from http://llt.msu.edu/vol1num1/brown/default.html [Accessed 9 July 2014]

Bullock, Deborah. 2011. "Learner Self-Assessment: An Investigation Into Teachers’ Beliefs." *ELT Journal: English Language Teachers Journal* 65, 2: 114–125.

Bygate, Martin. 1999. “Quality of language and purpose of task: patterns of learners’ language on two oral communication tasks.” *Language Teaching Research*, 3, 3: 185–214.

Chafe, Wallace and Deborah Tannen. 1987. “The Relation Between Written and Spoken Language.” *Annual Review of Anthropology* 16: 383–407.

Chapelle, Carol. 2003. *English language learning and technology: lectures on applied linguistics in the age of information and communication technology*. Amsterdam: John Benjamins Pub.

Charles, Marja-Liisa. 1998. “Europe: Oral Business Communication.” *Business Communication Quarterly* 61: 85–93.

Charles, Marja-Liisa. 2007. "Language Matters in Global Communication." *Journal of Business Communication* 44, 3: 260–282.

Cumming, Alister. 2008. “Assessing Oral and Literate Abilities.” In *Encyclopaedia of Language and Education, Volume 7: Language Testing and Assessment*, ed. Elana Shohamy, 3–18. New York: Springer Science and Business Media.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching and Assessment.* [Internet] Cambridge: Cambridge University Press. Available from http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf. [Accessed 2 June 2014]

Crosling, Glenda and Ian Ward. 2002. “Oral communication: the workplace needs and uses of business graduate employees.” *English for Specific Purposes* 21, 1: 41–57.

Davies, Alan. 2001. “The logic of testing Languages for Specific Purposes.” *Language Testing* 18: 133–147.

De Jong, Nel, and Charles A. Perfetti. 2011. "Fluency Training In The ESL Classroom: An Experimental Study Of Fluency Development And Proceduralization." *Language Learning* 61, 2: 533–568.

Douglas, Dan. 2010. *Understanding Language Testing*. London: Hodder Education.

Edwards, Nathan. 2000. “Language for business: effective needs assessment, syllabus design and materials preparation in a practical ESP case study.” *English for Specific Purposes* 19, 3: 291–296.

Ehrenreich, Susanne. 2010. “English as a Business Lingua Franca in a German Multinational Corporation: Meeting the Challenge.” *Journal of Business Communication* 47: 408–431.

El Hmoudova, Dagmar. 2013. “The Impact of Learning Style Dimensions on Computer-based Key Language Competence Testing.” *Procedia - Social and Behavioral Sciences* 82, 3: 411–416.

ELAN. 2006. *Effects on the European Economy of Shortages of Foreign Language Skills in Enterprise*. [Internet] Brussels, European Commission. Available from http://ec.europa.eu/languages/policy/strategic-framework/documents/elan_en.pdf. [Accessed 2 June 2014]

Ellis, Rod. 2005. “Instructed Language Learning and Task-based Teaching.” In *Handbook of research in second language teaching and learning*, ed. Eli Hinkel, 713–729. Mahwah (N.J.): Lawrence Erlbaum.

Ellis, Rod. 2009. “The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production.” *Applied Linguistics* 30: 474−509.

European Commission. 2012. *Eurobarometer. Europeans and their languages*. Special Eurobarometer 386/Wave EB77.1. European Commission. Available from http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf. [Accessed 12 March 2014]

Evans, Stephen. 2012. “Designing email tasks for the Business English classroom: Implications from a study of Hong Kong’s key industries.” *English for Specific Purposes* 31, 3: 202–212.

Figueras, Neus. 2012. “The impact of the CEFR.” *ELT Journal* 66, 4, Special issue: 477–485.

Firth, Alan. 1995. "'Accounts' in Negotiation Discourse: A Single-Case Analysis." *Journal of Pragmatics* 23, 2: 199–226.

Firth, Alan. 2009. "Doing Not being a Foreign Language Learner: English as a Lingua Franca in the Workplace and (some) Implications for SLA." *IRAL* 47,1: 127–56.

Fulcher, Glenn. 2003. *Testing second language speaking.* Harlow: Longman.

Frost, Kellie, Catherine Elder and Gillian Wigglesworth. 2012. "Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances." *Language Testing* 29, 3: 345–369.

Green, Anthony. 2012. *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range*. Cambridge University Press. Partly available from: http://books.google.com. [Accessed 14 May 2014]

Haataja, Kim. 2010. "Das Konzept LangPerform: Entwicklung und Einsatz von Simulationsinstrumenten zur computermedialen Dokumentierung von (fremd-) sprachlichen Kompetenzen; innovativ und integrativ." *Jahrbuch Deutsch als Fremdsprache* 36: 183–199.

Haataja, Kim and Taina Wewer. 2013. "Elokuvapohjaisia tietokonesimulaatioita kokonaisvaltaisen kieli- ja kulttuurikasvatuksen tueksi - mutta miksi? LangPerform-simulaatiokonseptin soveltaminen Proficom-kehittämishankkeessa: Taustoista, tavoitteista ja rakenteista sekä tähänastisista käytännön kokemuksista". In: *Löytöretkillä toisessa maailmassa. Oppimispelit ja virtuaalimaailmat (Finnish National Board of Education)*, eds Lauri Pirkkalainen and Petri Lounaskorpi. Konnevesi. Available from http://konnevedenlukio.onedu.fi/verkkojulkaisut/zine/33/cover [Accessed 26 August 2014]

Haataja, Kim and Rainer E. Wicke. 2014."Deutschsprachige Immersionsprogramme in den USA und Kanada. Eine bedeutende Variante des Content and Language Integrated Learning in German (CLILiG)". *Deutsche Lehrer im Ausland*, 3/2014, Jahrgang 61: 216–223. Münster: Aschendorff.

Hamp-Lyons, Liz and Tom Lumley. 2001. "Assessing language for specific purposes" *Language Testing* 18: 127–132.

Handford, Michael and Petr Matous. 2011. "Lexicogrammar in the international construction industry: A corpus-based case study of Japanese–Hong-Kongese on-site interactions in English." *English for Specific Purposes* 30, 2: 87–100.

Hasan, Eva-Lisa. 2011. *L'influence de l'âge sur l'acquisition de la prononciation du français. Une étude pilote sur un test de prononciation utilisant le concept de LangPerform*. Master's Thesis, Tampere: University of Tampere. Available from http://urn.fi/urn:nbn:fi:uta-1–22049. [Accessed 16 February 2014]

Hellekjær, Glenn Ole. 2007. *Fokus På Språk. Fremmedspråk i norsk næringsliv - engelsk er ikke nok!* Available from http://www.hiof.no/neted/upload/attachment/site/group55/Fokusnr3.pdf. [Accessed 6 December 2012]

Holmes, Janet. 2000. “Doing collegiality and keeping control at work: small talk in government departments”. In *Small talk,* ed. Justine Coupland, 32–61. Harlow: Longman.

Housen, Alex and Folkert Kuiken. 2009. *“*Complexity, Accuracy, and Fluency in Second Language Acquisition.” *Applied Linguistics* 30, 4: 461–473.

Huhta, Marjatta. 2010. *Language and communication for professional purposes needs analysis methods in industry and business and their yield to stakeholders.* Doctoral dissertation. Espoo: Yliopistopaino. Avaliable from http://lib.tkk.fi/Diss/2010/isbn9789522482273/isbn9789522482273.pdf. [Accessed 6 January 2014]

Ilkankoski, Katja. 2012. *«Apprendre pour la vie ou pour une épreuve, telle est la question.» Analyse de la compétence langagière en français des futurs bacheliers par un test pilote issu du concept LangPerform.* Master's Thesis, University of Tampere. Available from http://urn.fi/urn:nbn:fi:uta-1–23006. [Accessed 16 February 2014]

Jacoby, Sally and Tim McNamara. 1999. “Locating Competence.” *English for Specific Purposes* 18, 3: 213–241.

James, Robert. 1996. “CALL and the speaking skill.” *System* 24, 1: 15–21.

Jamieson, Joan, Linxiao Wang, Jacqueline Church. 2013. “In-house or commercial speaking tests: Evaluating strengths for EAP placement.” *Journal of English for Academic Purposes* 12, 4: 288-298.

Jenks, Christopher Joseph. 2011. *Transcribing Talk and Interaction: Issues in the Representation of Communication Data*. Amsterdam: John Benjamins.

Kankaanranta, Anne and Leena Louhiala-Salminen. 2007. "Business Communication in BELF." *Business Communication Quarterly* 70, 1: 55–9.

Kankaanranta, Anne and Leena Louhiala-Salminen. 2010. "English? -Oh, it’s just work!": A study of BELF users’ perceptions.” *English for Specific Purposes* 29: 204–209.

Kassim, Hafizoah, and Fatimah Ali. 2010. “English communicative events and skills needed at the workplace: Feedback from the industry.” *English for Specific Purposes* 29, 3: 168–182.

Knower, Franklin H. 1944. "What Is A Speech Test?" *Quarterly Journal Of Speech* 30, 4: 485–495.

Leaper, David A. and Mehdi Riazi. 2014. "The Influence Of Prompt On Group Oral Tests." *Language Testing* 31,2: 177-204.

Leblay, Tarja. 2013.*”Voi ei, nää on tosi hyviä verrattuna muhun!”.* Doctoral dissertation. Jyväskylä: University of Jyväskylä. Available from http://urn.fi/URN:ISBN:978–951–39–5384–3. [Accessed 16 February 2014]

Lee, Kuang-wu. 2000. “English Teachers' Barriers to the Use of Computer Assisted Language Learning.” *The Internet TESL Journal*. Available from http://iteslj.org/Articles/Lee-CALLbarriers.html [Accessed 23 May 2014]

Lehtonen, Tuula and Sinikka Karjalainen. 2008. “University graduates’ workplace language needs as perceived by employers.” *System* 36, 3: 492–503.

Leppänen, Sirpa, Anne Pitkänen-Huhta, Tarja Nikula, Samu Kytölä, Timo Törmäkangas, Kari Nissinen, Leila Kääntä, Tiina Virkkula, Mikko Laitinen, Päivi Pahta, Heidi Koskela, Salla Lähdesmäki and Henna Jousmäki. 2009. *National Survey on the English Language in Finland: Uses, Meanings and Attitudes.* Jyväskylä: University of Jyväskylä. Available from http://www.helsinki.fi/varieng/series/volumes/05/evarieng-vol5.pdf. [Accessed 8 September 2013]

Little, David. 2005. "The Common European Framework and the European Language Portfolio: Involving Learners and Their Judgements in the Assessment Process." *Language Testing* 22, 3: 321–336.

Littlewood, William. 1992. *Teaching oral communication: a methodological framework.* Oxford: Blackwell.

Lockwood, Jane. 2012. “Developing an English for specific purpose curriculum for Asian call centres: How theory can inform practice.” *English for Specific Purposes* 31, 1: 14–24.

Louhiala-Salminen, Leena, Marja-Liisa Charles and Anne Kankaanranta. 2005. “English as a lingua franca in Nordic corporate mergers: Two case companies.” *English for Specific Purposes* 24: 401–421.

Louhiala-Salminen, Leena and Anne Kankaanranta. 2011. “Professional Communication in a Global Business Context: The Notion of Global Communicative Competence.” *IEEE Transactions on professional communication* 54, 3: 244–262.

Lumley, Tome. 1998. “Perceptions of Language-trained Raters and Occupational Experts in a Test of Occupational English Language Profficiency.” *English for Specific Purposes* 17, 4: 347–367.

Malabonga, Valerie, Dorry M. Kenyon, and Helen Carpenter. 2014. "Self-Assessment, Preparation and Response Time on a Computerized Oral Proficiency Test." *Language Testing* 22, 1: 59–92.

Mikkonen, Simo, Anna Veijola and Pasi Ihalainen. 2013. Vertais- ja itsearvioinnin käyttö yliopistollisessa historianopetuksessa: tapaustutkimus menetelmäopetuksen opetuskokeilusta. In *Kasvatus & Aika* 7, 3: 69–85. Available from http://www.kasvatus-ja-aika.fi/dokumentit/mikkonen_ja_kump_0609131006.pdf. [Accessed 22 May 2014]

Mochizuki, Naoko and Lourdes Ortega. 2008. “Balancing communication and grammar in beginning-level foreign language classrooms: A study of guided planning and relativization.” *Language Teaching Research* 12: 11–37.

Moslehifar, Mohammad Ali and Noor Aireen Ibrahim. 2012. “English Language Oral Communication Needs at the Workplace: Feedback from Human Resource Development (HRD) Trainees.” *Procedia - Social and Behavioral Sciences* 66, 7: 529–536.

Nedzinskaitė, Inga, Dana Švenčionienė and Daiva Zavistanavičienė. 2006. “Achievements in Language Learning through Students’ Self-assessment.” *Studies about Languages (Kalbu Studijos)* 8: 84–87. Available from http://www.kalbos.lt/zurnalai/08_numeris/12.pdf. [Accessed 30 June 2014]

Neeley, Tsedal B, Pamela J. Hinds and Catherine D. Cramton. 2012. “The (Un)Hidden Turmoil of Language in Global Collaboration.” *Organizational Dynamics* 41, 3: 236–244.

Nickerson, Catherine. 2005. “English as a lingua franca in international business contexts.” *English for Specific Purposes* 24, 4: 367–380.

Niemelä, Pia. 2012. *Kerro, kerro kuvastin: tietokoneavusteinen itse- ja vertaisarviointi.* Master's thesis. Tampere: University of Tampere. Available from http://urn.fi/urn:nbn:fi:uta-1–22956. [Accessed 17 April 2014]

Ockey, Gary J. 2009. "Developments And Challenges In The Use Of Computer-Based Testing For Assessing Second Language Ability." *Modern Language Journal* 93: 836–847.

Okada, Yusuke. 2010. “Role-play in oral proficiency interviews: Interactive footing and interactional competencies.” *Journal of Pragmatics* 42, 6: 1647–1668.

O'Loughlin, Kieran. 2008. “Assessment at the Workplace. In *Encyclopaedia of Language and Education, Volume 7: Language Testing and Assessment*, ed. Elana Shohamy, 69–80. New York: Springer Science and Business Media.

PIMLICO. 2011. *Report on Language Management Strategies and Best Practice in European SMEs*. [Internet] Brussels, DGEAC (European Commission). Available from http://ec.europa.eu/languages/languages-mean-business/files/pimlico-full-report_en.pdf. [Accessed 2 June 2014]

Pradas Macías, Macarena. 2006. "Probing Quality Criteria In Simultaneous Interpreting: The Role Of Silent Pauses In Fluency." *Interpreting: International Journal Of Research & Practice In Interpreting* 8, 1: 25–43.

Pullin, Patricia. 2010. "Small Talk, Rapport, and International Communicative Competence." *Journal of Business Communication* 47, 4: 455–76.

Roberts, Celia. 2005. 'English in the Workplace'. In *Handbook of research in second language teaching and learning*, ed. Eli Hinkel, 117–136. Mahwah (N.J.): Lawrence Erlbaum.

Ross, John A. 2006. “The Reliability, validity and utility of self-assessment.” *Practical Research, Assessment & Evaluation* 11, 10: 1–13. Available from http://pareonline.net/pdf/v11n10.pdf [Accessed 7 April 2014]

Ross, Steven. 1998. “Self-assessment in second language testing: a meta- analysis and analysis of experiential factors”. *Language Testing* 15, 1: 1–20.

Rossiter, Marian J. 2009. “Perceptions of L2 Fluency by Native and Non–native Speakers of English.” *Canadian Modern Language Review* 65, 3: 395–412.

Räisänen, Tiina. 2012. “Processes and practices of enregisterment of business English, participation and power in a multilingual workplace.” *Sociolinguistic Studies* 6, 2: 309–331.

Räisänen, Tiina. 2013. *Professional communicative repertoires and trajectories of socialization into global working life*. Doctoral dissertation. University of Jyväskylä. Available from https://jyx.jyu.fi/dspace/bitstream/handle/123456789/42565/978–951–39–5470-3.pdf?sequence=3. [Accessed 20 February 2014]

Stoynoff, Stephen. 2012. “Looking backward and forward at classroom-based language assessment.” *ELT Journal* 66, 4: 523–523.

Sweeney, Emma and Zhu Hua. 2010. "Accommodating Toward Your Audience. Do Native Speakers of English Know how to Accommodate their Communication Strategies Toward Nonnative Speakers of English?" *Journal of Business Communication* 47, 4: 477–504.

Tavakoli, Parvaneh and Pauline Foster. 2011. “Task Design and Second Language Performance: The Effect of Narrative Type on Learner Output.” *Language Learning* 61: 37–72.

Tuuna-Kyllönen, Tanja. 2011. Zur Beurteilung der mündlichen Sprachfertigkeit in der Endphase der gymnasialen Oberstufe: Der Test des Zentralamtes für Unterrichtswesen und die Computersimulation LangPerform im Vergleich. Master's Thesis, Tampere: University of Tampere. Available from http://tampub.uta.fi/bitstream/handle/10024/82720/gradu05215.pdf?sequence=1. [Accessed 16 February 2014]

Tratnik, Alenka. 2008. “Key Issues in Testing English for Specific Purposes.” *Scripta Manent* 4, 1: 3–13. Available from http://www.sdutsj.edus.si/ScriptaManent/2008_4_1/Tratnik.pdf [Accessed 13 May 2014]

Underhill, Nick. 1987. *Testing spoken language: A handbook of oral testing techniques.* Cambridge: Cambridge University Press.

Van Teijlingen, Edwin R. and Vanora Hundley. 2001. “The importance of pilot studies.” *Social Research Update* 35: (No page numbers). Available from http://sru.soc.surrey.ac.uk/SRU35.html [Accessed 15 May 2014]

Warren, Martin. 2013. “‘Just spoke to …’: The types and directionality of intertextuality in professional discourse.” *English for Specific Purposes* 32, 1: 12–24

Weisi, Hiwa and Mohammad Nabi Karimi 2013. “The Effect of Self-Assessment Among Iranian EFL Learners.” *Procedia - Social and Behavioral Sciences* 70, 2: 731–737.

Wewer, Taina. 2014. *Assessment of Young Learners’ English Proficiency in Bilingual Content Instruction CLIL*. Doctoral dissertation. Turku: University of Turku. Available from https://www.doria.fi/handle/10024/96838 [Accessed 16 August 2014]

Wigglesworth, Gillian. 2008. “Task and Performance Based Assessment.” In *Encyclopaedia of Language and Education, Volume 7: Language Testing and Assessment*, ed. Elana Shohamy, 111–123. New York: Springer Science and Business Media.

Wilson, Mary S., and Bernard J. Fox. 1982. "Computer-Administered Bilingual Language Assessment and Intervention." *Exceptional Children* 49, 2: 145–149.

Wood, David. 2006. “Uses and Functions of Formulaic Sequences in Second Language Speech: An Exploration of the Foundations of Fluency.” *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes* 63, 1: 13–33.

Appendix 1
Illustrative scales for spoken fluency and propositional precision (Council of Europe 2001, 129).

Spoken fluency	
C2	Can express him/herself at length with a natural, effortless, unhesitating flow. Pauses only to reflect on precisely the right words to express his/her thoughts or to find an appropriate example or explanation.
C1	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.
B2	Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech.
	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions, there are few noticeably long pauses. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without imposing strain on either party.
B1	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 keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.
A2	Can make him/herself understood in short contributions, even though pauses, false starts and reformulation are very evident.
	Can construct phrases on familiar topics with sufficient ease to handle short exchanges, despite very noticeable hesitation and false starts.
A1	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.

Propositional Precision	
C2	Can convey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of qualifying devices (e.g. adverbs expressing degree, clauses expressing limitations). Can give emphasis, differentiate and eliminate ambiguity
C1	Can qualify opinions and statements precisely in relation to degrees of, for example, certainty/ uncertainty, belief/doubt, likelihood, etc.
B2	Can passion detailed information reliably.
B1	Can explain the main points in an idea or problem with reasonable precision.
B1	Can convey simple, straightforward information of immediate relevance, getting across which point he/she feels is most important. Can express the main point he/she wants to make comprehensibly.
A2	Can communicate what he/she wants to say in a simple and direct exchange of limited information on familiar and routine matters, but in other situations he/she generally has to compromise the message.
A1	No description available

Appendix 2
Can do –statements in the self-assessment questionnaire for spoken fluency, propositional precision, spoken production, spoken interaction and global scale.

Spoken fluency	
C1	I can express myself fluently and spontaneously, almost effortlessly without long pauses. Only an unfamiliar subject can hinder a natural, smooth flow of language.
B2	I can communicate spontaneously and very fluently with a fairly even tempo in longer complex stretches of speech.
B1	I can communicate spontaneously and very fluently with a fairly even tempo in longer complex stretches of speech.
A2	I can make myself understood and handle short conversations on familiar topics.
A1	I can manage very short, isolated phrases.

Propositional precision	
C1	I can give opinions and statements precisely and express my certainty/ uncertainty, belief/doubt, etc.
B2	I can pass on detailed information reliably.
B1	I can explain the main points in an idea or problem and get across the point which I feel is most important.
A2	I can say what I want to say in a simple conversation on familiar and routine matters.
A1	-

Spoken production	
C1	I can present clear, detailed descriptions of complex subjects developing particular points and closing with an appropriate conclusion.
B2	I can present clear, detailed descriptions on many different subjects that I am interested about. I can explain a viewpoint giving the advantages and disadvantages of various options.
B1	I can connect phrases in a simple way in order to describe experiences and events, my dreams, hopes and ambitions. I can briefly give reasons and explanations for opinions and plans. I can tell a story, or the plot of a book or film and describe my reactions.
A2	I can describe in simple language my family and other people, living conditions, my educational background and my job.
A1	I can use simple phrases and sentences to describe where I live and people I know.

Spoken interaction	
C1	I can present clear, detailed descriptions of complex subjects developing particular points and closing with an appropriate conclusion.
B2	I can present clear, detailed descriptions on many different subjects that I am interested about. I can explain a viewpoint giving the advantages and disadvantages of various options.
B1	I can connect phrases in a simple way in order to describe experiences and events, my dreams, hopes and ambitions. I can briefly give reasons and explanations for opinions and plans. I can tell a story, or the plot of a book or film and describe my reactions.
A2	I can describe in simple language my family and other people, living conditions, my educational background and my job.
A1	I can use simple phrases and sentences to describe where I live and people I know.

Global scale	
C1	I can understand a wide range of demanding, longer texts, and recognise implicit meaning. I can express myself fluently and spontaneously without much obvious searching for expressions. I can use language flexibly and effectively for social, academic and professional purposes. I can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
B2	I can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in my field of specialisation. I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. I can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
B1	I can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. I can deal with most situations likely to arise whilst travelling in an area where the language is spoken. I can produce simple connected text on topics which are familiar or of personal interest. I can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
A2	I can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. I can describe in simple terms aspects of my background, immediate environment and matters in areas of immediate need.
A1	I can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. I can introduce myself and others and can ask and answer questions about personal details such as where I live, people I know and things I have. I can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.