HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

The assessment of upper secondary school students' oral skills in a face-to-face and a computer-based test

A case study in Finland

Essi Pohto
Master's thesis
English Philology
Department of Modern Languages
University of Helsinki
May 2019

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty Humanistinen tiedekunta	Laitos – Institution – Department Nykykielten laitos	
Tekijä – Författare – Author Essi Pohto		
Työn nimi – Arbetets titel – Title The assessment of upper secondary school students' oral skills in a face-to-face and a computer-based test: A case study in Finland		
Oppiaine – Läroämne – Subject Englantilainen filologia		
Työn laji – Arbetets art – Level Pro gradu -tutkielma	Aika – Datum – Month and year Toukokuu 2019	Sivumäärä– Sidoantal – Number of pages 79 + liitteet

Tiivistelmä – Referat – Abstract

Teknologiaa on viime vuosikymmeninä alettu yhä enenevässä määrin käyttää kielitaidon arvioinnissa. Tämä trendi on havaittavissa myös Suomessa, jossa ylioppilaskokeiden sähköistyminen on viime vuosina puhuttanut paljon. Kielten ylioppilaskokeisiin on kaavailtu kirjallisen kokeen oheen myös suullista osiota, joka olisi tarkoitus toteuttaa lähivuosina sähköisenä. Tämä tutkielma pyrkii ottamaan selvää siitä, onko lukiolaisten suullisessa tuotannossa eroja riippuen siitä, suorittavatko he kokeen kasvotusten arvioijan kanssa vai tietokoneen avulla. Tutkielma pyrkii lisäksi selvittämään, millaisia asenteita ja mielipiteitä lukiolaisilla on eri suoritusmuodoista.

Tutkimukseen osallistui 15 opiskelijaa yhdestä pääkaupunkiseudun lukiosta. He suorittivat ensin kasvokkain tehtävän kokeen ja kolmen viikon kuluttua tietokonepohjaisen kokeen. Tulosten verrattavuuden vuoksi kokeet suunniteltiin siten, että ne ovat mahdollisimman samanlaiset. Molemmat kokeet koostuivat neljästä eri tehtävästä. Opiskelijoiden suoritettua kokeet heitä pyydettiin vastaamaan kyselyyn, jonka tarkoituksena oli selvittää, kummasta koemuodosta he pitivät enemmän ja miksi ja kummassa he kokivat pystyvänsä antamaan paremman näytön suullisesta kielitaidostaan. Ryhmähaastatteluissa opiskelijoita pyydettiin täydentämään kyselyssä antamiaan vastauksia, minkä jälkeen keskustelu eteni suullisen kielitaidon testaamiseen ja kokeiden sähköistymiseen liittyviin kysymyksiin.

Opiskelijoiden suoritukset kahdessa eri koemuodossa litteroitiin, jonka jälkeen transkriptiot analysoitiin sekä kvantitatiivisin että kvalitatiivisin menetelmin. Suorituksissa verrattiin oikeakielisyyden (*accuracy*), kieliopillisen sekä sanastollisen kompleksisuuden (*complexity*) sekä sujuvuuden (*fluency*) näkökulmista. Kyselyvastaukset käytiin läpi kysymys kerrallaan ja haastatteluista litteroitiin ne osat, jotka olivat analyysin kannalta merkityksellisiä.

Tulokset osoittavat, että lukiolaisten suullisessa tuottamisessa on eroja eri koemuodoissa, mutta nämä erot jäivät pieniksi. Tutkimuksen pohjalta on kuitenkin mahdotonta päätellä, johtuvatko erot koemuodosta vai joistakin muista seikoista. Lukiolaiset tuottivat keskimäärin virheettömämpää englantia kasvokkain suoritettavassa kokeessa, ja osallistujat pitivät kyseisessä kokeessa vähemmän taukoja. Suuri osa opiskelijoista kuitenkin tuotti kompleksisempaa puhetta tietokonepohjaisessa kokeessa. Yksilöllisiä eroja oli tosin havaittavissa jokaisen mittarin kohdalla. Tutkimuksen tuloksista selvisi myös, että osallistujat suosivat kasvokkain suoritettavaa koemuotoa sen vuorovaikutuksellisuuden vuoksi. He pystyivät mielestään myös antamaan siinä paremman näytön suullisesta kielitaidostaan, kokivat sen olevan helpompi ja olivat vähemmän hermostuneita sitä suorittaessaan. Tätä aihetta tulee tutkia jatkossa enemmän erityisesti siksi, että suullisen kielitaidon testaamisella on suuret vaikutukset kielten ylioppilaskirjoituksiin, mikäli suullinen osio lisätään ja se suoritetaan tietokonekokeena.

Avainsanat – Nyckelord – Keywords
suullisen kielitaidon arviointi – kasvokkainen koe – tietokonepohjainen koe – kompleksisuus – oikeakielisyys – sujuvuus

Säilytyspaikka – Förvaringställe – Where deposited
Keskustakampuksen kirjasto

Muita tietoja – Övriga uppgifter – Additional information

Table of contents

List of tables

List of figures

List of abbreviations and acronyms

L2	Second language
SLA	Second language acquisition
CEFR	Common European Framework of Reference
CAF	Complexity, accuracy and fluency
OPI	Oral proficiency interview
SOPI	Simulated oral proficiency interview
COPI	Computerized oral proficiency interview
OPIc	Oral proficiency interview – computer
CALT	Computer-assisted language testing
CBT	Computer-based test
NCC	(Finnish) National Core Curriculum

1 Introduction

Despite its importance, the teaching and testing of speaking was, for a long time, secondary to the other three skills — listening, reading and writing (Bygate 2011: 412). One of the reasons that it was neglected for so long is the fact that the assessment of speaking skills poses several challenges. Unlike the other three skills, speaking is immediate, time conditioned, unpredictable and reciprocal, making it difficult to develop viable large-scale tests (Kenyon and Malone 2010: 2; Sawaki 2012: 433). According to Luoma (2004: 1), however, "[s]peaking skills are an important part of the curriculum in language teaching, and this makes them an important object of assessment as well".

In Finland, language assessors have, for a long time, attempted to incorporate an oral component to the final-school leaving examination (Huhta and Hildén 2016: 13). Although the introduction of an oral test section to the matriculation examination was initially proposed as early as 1958, it was only in 2010 that an optional advanced oral skills course was added to the foreign languages and second national language syllabus (Saleva 1997: 12; Ministry of Education 2006). It has not, to this day, been incorporated into the matriculation examination. The Finnish matriculation examination has, however, recently undergone a major transformation, whereby as of this year, all the examinations are delivered in electronic format. This change has brought with it yet another attempt to introduce an oral component which was initially meant to be realized by 2019 but was later postponed.

According to Hildén and Vähähyyppä (2016), the testing of oral skills is now possible due to the digitalization of the matriculation examination and the development of speech technology. However, as Chapelle and Douglas (2006: 3) point out, there are stakeholders who are concerned due to the risk that a computer-based language test may disadvantage learners, hampering their ability to demonstrate the full extent of their oral proficiency. This is also true in Finland. Although the main purpose of the matriculation examination is to act as a form of final summative assessment, it also has an important gatekeeper function, as the test results increasingly often determine higher education entry (Huhta and Hildén 2016: 6). The stakes are very high, and validity and reliability issues should therefore be

considered thoroughly. According to Huhta and Hildén (2016: 21) the question of how much technology changes the construct of language proficiency to be tested has been studied very little in Finland and will likely be studied in more detail in the near future.

L2 oral proficiency testing has become increasingly prevalent with technological advances (Isaacs 2016: 131), and this has subsequently led to the need to carefully define the construct of L2 proficiency. Nowadays, it is seen as having a complex, multicomponential structure which the measures of complexity, accuracy and fluency (CAF) are able to capture (Housen *et al*. 2012). These measures are, in fact, implied in the assessment rubrics of many influential testing institutions, including the Common European Framework of Reference (CEFR). Several studies have looked at how task types and conditions affect these particular measures, but there do not seem to be many which have focused on the impact of these measures in a testing context.

In addition to my interest in the potential conflict of interest regarding the assessment of speaking with a computer, I am interested in seeing whether testing mode can affect test-takers' speech with reference to CAF measures. Therefore, in this thesis, I examine the following research questions:

1. *What differences are there in upper secondary school students' oral performance when comparing a face-to-face test to a computer-based equivalent in terms of:*
    a. *complexity*
    b. *accuracy, and*
    c. *fluency?*
2. *What types of attitudes do students have regarding these two tests?*

Previous research focusing on test comparability has mostly indicated that, although test performance results are relatively equal, examinees generally have a strong preference for the face-to-face testing format (Kenyon and Malone 2010; Kiddle and Kormos 2011; Thompson *et al*. 2016). Establishing test comparability, not only in terms of scores, but also in terms of examinees' linguistic output is an important issue so as to ensure test validity. The issue of test-taker preferences is, however, also an important matter because, as Underhill (1987: 11–12) claims: "A good oral test

allows learners to be treated, and to behave, like ordinary human beings, so it seems right that we should do a market survey of what they think, expect and want".

This thesis is structured as follows: Chapter 2 introduces the theoretical framework, which focuses on the fields of language testing and, more specifically, the assessment of speaking. The measures of complexity, accuracy, and fluency, which are used to analyze the data, are discussed in detail with reference to their relevance to L2 oral proficiency. Then, before discussing relevant previous studies concerning test performance and test-taker preference comparability, computer-assisted language testing is addressed. The theoretical framework is concluded with a discussion of the relevance of this paper to the Finnish context. Chapter 3 contains a detailed description of the data, the data collection process and the methods used to analyze the data. Chapter 4 consists of the analysis in which findings regarding differences in CAF measures in the two tests as well as test-takers' attitudes towards the two test formats are discussed. Chapter 5 consists of the discussion, which entails reflections on the implications and limitations of this study as well as suggestions for further study. The thesis is concluded by Chapter 6, the conclusion.

2 Theoretical framework

The theoretical framework of this thesis centers around the field of language testing, with a particular focus on the testing of oral skills. In the first section, language ability is defined by looking into some of the most important language testing paradigms and prevailing language learning methods. The qualities of an ideal test are then discussed. The second section focuses on the nature of speaking and on the analysis of speech with the help of complexity, accuracy and fluency measures. The third section has as its focus the field of computer-assisted language testing, whereby the advantages and disadvantages of direct and semi-direct tests are analyzed. Relevant previous studies concerning the interchangeability of direct and semi-direct tests are then discussed in the fourth section in terms of comparability and test-taker preferences. Finally, as the Finnish context is pertinent to this study, the matriculation examination will be discussed in light of recent changes, such as its digitalization and the possible introduction of an oral component to language examinations.

2.1 Language testing

This section focuses on language testing by first going through the main language testing paradigms and the concurrent language learning methods. From a testing point of view, this is important so as to understand how the construct of language ability has changed over time. The second sub-section then looks into the question of what makes a test fair and thus seeks to answer the question of what a test designer ought to strive for in the design and development phases of a test.

2.1.1 Language testing paradigms

Over the years, the definition of language ability has been susceptible to paradigmatic swings in approaches to L2 teaching, and in turn, assessment. Starting with the Audiolingual Method to Communicative Language Teaching, up to the post-method era, the construct of language ability has continuously been witnessing changes, mainly in terms of the nature and method of designing test items. In the realm of language testing, Ellis (2003: 280–283) distinguishes between three major language assessment paradigms. These are: (1) the *psychometric* tradition in testing, (2) *integrative* language testing, and (3) *communicative* language testing. This section focuses on how these paradigms, together with the prevalent language learning theory of the time, affected the definition of language ability and the importance of learning to speak a foreign language.

According to Bygate (2012: 9), the teaching, and by extension testing, of speaking started to appear as a concern in its own right as late as the 1940s. Up until then, the Grammar-Translation Method had dominated the language teaching and testing scene, and the focus had consequently been on learning grammar rules and on translating into and out of the target language. Accuracy in writing was accentuated; however, oral skills were not seen as worthy of attention (Richards and Rodgers 2014: 6–7). Following World War II, the Audiolingual Method, which emphasized oral fluency and accuracy of phonology, began to emerge (Fulcher 2003b: 1). The theory of language underlying this method was *structural linguistics* which perceived language as a self-contained relational structure (Richards and Rodgers 2014: 62; Mitchell *et al*. 2013: 305). Language was broken into components such as pronunciation, grammar and vocabulary. Knowledge of these different elements was

tested in relation to the four language skills: listening, speaking, reading and writing (Lado 1961: 25–26; Bachman and Palmer 1996: 75; Ellis 2003: 280).

Moreover, drawing from behavioral psychology, language was regarded as "a system of habits of communication" (Lado 1961: 22), which meant that learning was seen as the mechanical formation of habits (Mitchell *et al.* 2013: 28). Exercises leading to memorization such as drills, pattern practice and model dialogues were favored (Richards and Rodgers 2014: 66–67). However, despite the emphasis on speaking and listening, the aim was not to foster spoken interaction (Bygate 2012: 9) but rather to acquire "a set of appropriate speech habits" (Mitchell *et al.* 2013: 28). The psychometric tradition in language testing, too, was strongly influenced by structuralism and behaviorism (Ellis 2003: 280). Psychometric language tests emphasized *objectivity*, and for this reason, tests often consisted of closed type questions (e.g. multiple choice). Pižorn and Huhta (2016: 241) point out that due to concerns about their reliability, the productive skills of speaking and writing were largely ignored in national assessments up until the 1970s. Statistical procedures were used to determine the validity and reliability[1] of tests (Ellis 2003: 280).

Similar statistical procedures were used in determining the reliability of integrative language tests. However, unlike the theory underlying psychometric testing, integrative language tests were influenced by a unitary rather than a multidimensional view of language (Ellis 2003: 281). In the 1970s, Oller developed his so-called *unitary hypothesis*, which challenged the multicomponential view of language proficiency. Instead of regarding language ability as consisting of different components, he believed that language ability was made up of one single, indivisible competence. He believed all verbal activity to be based on an internalized expectancy grammar, with the help of which a language user can take advantage of the redundancy in language as well as general world knowledge in order to comprehend and produce language (Huhta 1993: 80–81; Saleva 1997: 20). During this period, discrete-point tests[2] were common because high correlations between separate tests focusing on different elements (e.g. grammar vs. vocabulary) were found. This was

[1] The concepts of validity and reliability will be discussed in more detail in the next section.

[2] Discrete-point testing works on the assumption that language can be broken down into its components and that these can be assessed one at a time. Examples of discrete-point test items in language testing include multiple choice, true/false, fill in the blank, and spelling (Hughes 1989: 16–17; Douglas 2010: 70–71).

seen as an indication of the tests measuring the same underlying factor. Although his hypothesis was briefly influential, studies conducted in the 1980s which adopted better developed methods proved Oller wrong and indicated that language ability does indeed consist of more than one skill (Huhta 1993: 81).

In the 1980s, language ability began to be seen as a means for communication. The communicative approach stresses the importance of contextual appropriacy and fluency in addition to grammatical accuracy (Bygate 2012: 10–11). Canale and Swain (1980) and later, extending on their model, Bachman (1990) and Bachman and Palmer (1996), proposed an influential model that regards language use as interaction between language users and their context. With regard to testing, Bachman and Palmer (1996: 67) claim that the object of measurement in a language test is *language ability* which they define as a "capacity […] to create and interpret discourse" that is made up of two components: *language knowledge* and *strategic competence*. In other words, unlike the prior testing traditions, communicative language testing emphasizes, not only knowledge of language, but also its use. Their model also shows that an individual's language use and test performance are affected by topical knowledge, affective schemata and personal characteristics (Bachman and Palmer 1996: 64–65).

Although Bachman (1990: 82) and Bachman and Palmer (1996: 75) agree that language ability is not made up of a unitary global trait underlying L2 performance, they disagree with the view that language ability is manifested through the skills of listening, speaking, reading and writing due to the fact that it fails to capture the full context of language use whereby these skills overlap[3]. A face-to-face conversation and a radio broadcast, for example, are different activities; however, both of them involve listening. Isaacs (2016: 132) adds that, due to the interactional nature of speech, it may not only be conceptually difficult but also artificial to separate speaking and listening. From a testing point of view, it is therefore important to consider whether one is testing a specific, stand-alone skill (e.g. speaking) or integrated skills (e.g. both speaking and listening), as all four of the skills are essential for communication.

[3] However, Douglas (2010: 19) claims that several researchers nowadays tend to agree that language ability is manifested through these four skills. In my opinion, this is reflected in the number of large-scale commercial tests which include separate test sections for the four skills.

Naturally, as with any model, the communicative language teaching and testing paradigm is not without its limitations. Communicative language teaching and testing has mainly focused on real-life tasks which emphasize *meaning*. In addition to comprehensible input, *focus on form* is required[4]. However, as Skehan (2003: 392) points out, second language acquisition (SLA) studies have shown that input alone does not lead to mastering a language. Furthermore, Bygate (2011: 412) claims that the more recent communicative language approach has highlighted the importance of speaking, but rather than viewing speech as a target skill to be learned, it has been regarded as a medium for learning other (non-)linguistic skills. Despite these limitations, communicative language teaching has, according to Harding (2014: 187), "become the unremarkable norm in test design" and it also forms a fundamental basis for the Common European Framework of Reference (CEFR), which is very influential in Europe and in Finland in particular.

2.1.2 Test qualities

When designing and developing a language test, the most important aspect to consider is "the use for which it is intended" (Bachman and Palmer 1996: 17). In defining a test's use, one needs to take into account such contextual factors as the test-takers, the stakes involved, and consequently, the implications that the test will have for the test-takers' future. According to Bachman and Palmer (1996: 17), a test's most important quality is therefore its usefulness, which they define as the sum of six test qualities, namely reliability, construct validity, authenticity, interactiveness, impact, and practicality. They stress that these qualities should not be seen as being mutually exclusive, but rather as complementary qualities that need to be balanced, while keeping in mind the test's purpose (cf. Morrow (1986) who argues that striving to design an authentic test ultimately compromises the test's reliability).

Reliability relates to the consistency of measurement, or in other words, to the extent to which a test provides an accurate measure of the abilities it is intended to measure (Bachman and Palmer 1996: 19; Douglas 2010: 10). In practice, this implies that if a test-taker were to take the same or a similar test meant to be used interchangeably

[4] Focus on form essentially entails the need to point a learner's attention to the structural dimensions of language.

with the first one, s/he should achieve equivalent results. However, as Douglas (2010: 10) points out, all tests are, to some degree, inconsistent in their measurement. This can be the result of either test-related (e.g. unclear instructions, unfamiliar test tasks) or examinee-related factors (e.g. test-takers experiencing fatigue or anxiety). Test designers therefore need to recognize that it is not possible to eliminate inconsistencies entirely (Bachman and Palmer 1996: 20), but that they have an ethical responsibility to design tests that are as accurate as possible in order to give examinees as fair a measurement of their abilities as possible (Douglas 2010: 10).

A fair exam is also one which is valid. A test has often been defined as valid when it measures what it sets out to measure (Lado 1961: 321); however, this definition has been deemed too vague because, as Weir (2005: 13) points out, validity is a multifaceted concept. *Construct validity*, and validity more generally, refer to the extent to which test designers can draw meaningful and appropriate conclusions about a test-taker's abilities based on their test score (Bachman and Palmer 1996: 21; Douglas 2010: 10). Bachman (1990: 161) claims that to ensure the validity and reliability of a test, the objectives in test design and development are twofold: (1) minimize measurement error effects and (2) maximize the effects of the language abilities to be measured. Another dimension of validity, the importance of which is often left unacknowledged, is that of *face validity*, i.e. the extent to which a test appeals to test-takers and users (Bachman and Palmer 1996: 42).

Furthermore, *authenticity* is an important quality to achieve in the language testing domain, considering that most language tests aim to say something about a test-taker's ability to use language in a variety of real-life situations. As Bachman and Palmer (1996: 23) point out, "[t]he primary purpose of a language test is to provide a measure that we can interpret as an indicator of an individual's language ability". Authenticity is thus the degree of correspondence of test task characteristics with target language use in natural situations outside the language test in a real-life context (Douglas 2010: 24).

Interactiveness refers to the interaction that occurs between the test task(s) and the test-taker; it is a function of the extent and type of involvement of the test-taker's language ability, topical knowledge, and affective schemata in accomplishing a test task. These can be engaged by a test task in diverse ways (Bachman and Palmer

1996: 25). The fifth quality, *impact*, encompasses the effect of a test on society as well as on the individual taking the test (Bachman and Palmer 1996: 29). Particularly in high-stakes tests, it is important to consider the ethical dimension the test has due to its effect on the test-takers' lives as well as its possible washback effect. Due to such far-reaching consequences, a language test should aim to be as equal and fair as possible (Tossavainen 2016: 28). Finally, as the term *practicality* suggests, a test needs to be practical in terms of the time and resources required for its arrangement (Bachman and Palmer 1996: 35; Luoma 2004: 175).

In addition to finding a desired balance between the aforementioned six qualities and thus ensuring that the test is appropriate in terms of its use, one has to also define the ability/abilities one wants to measure. In a language test, this means defining language ability in a way that is appropriate for the testing situation (Bachman and Palmer 1996: 66). This is by no means an easy task because, as Huhta (1993: 78) points out, theories of language ability are rarely specific and detailed enough in order to be directly applied to problem solving in practice. Moreover, as demonstrated in the previous section, definitions of language ability have greatly varied in time. Douglas (2010: 2) also addresses the problem of measuring language ability due to its abstract nature:

> A language test is an instrument for measuring language ability. […] But what does it mean to say that we want to measure ability or quantity of language? In what sense can we actually measure a concept as abstract as language ability?

Douglas (2010: 9–10) claims that it is in fact not possible to measure language ability at all: it is only possible to observe and measure performance, and to make inferences about test-takers' language ability on the basis of their performance. However, in order to ensure a test's reliability and validity, one needs to be able to base one's views on some theoretical framework. Otherwise one runs the risk of testing language ability from too narrow a perspective, or perhaps not testing language ability at all (Huhta 1993: 78). Although a test developer is ethically obligated to strive to meet all the above-mentioned criteria, no test is perfect. This is because, as Bachman and Palmer (1996: 19) point out, evaluating a test's overall usefulness is in essence subjective due to test developers' value judgements. Furthermore, Huhta

(1993: 78) adds that a language test always reflects the test designer's attitudes and preconceptions about language — whether s/he is aware of it or not.

2.2 Assessing speaking

This next section attempts to define the construct of L2 speaking by analyzing the nature of speech. This is followed by a section which describes the measures of complexity, accuracy and fluency. These are important because they are seen to sufficiently and systematically capture the major dimensions of L2 proficiency, and also because these measures are used in the analysis of this thesis.

2.2.1 The nature of speech

Developing and testing a person's ability to speak a foreign language was, for a very long time, secondary with respect to the other three skills. This was mainly due to the fact that it was impossible to test reliably and difficult to realize in practice. Luoma (2004: 1) also adds that the assessment of speaking skills poses several challenges "because there are so many factors that influence our impression of how well someone can speak a language". Unlike writing, speaking is not as well established because it does not have equally strict norms that are codified and taught; speaking is thus more subject to variation. In the literature, the construct of L2 oral proficiency is typically operationalized by describing its characteristics and by contrasting it with the other productive skill of writing.

Chafe (1985: 105), in his description of language, discusses the dimensions of fragmentation versus integration and involvement versus detachment:

> The fact that writing is a slow, deliberate, editable process, whereas speaking is done on the fly, leads to a difference that I called the integrated quality of written language as opposed to the fragmented quality of spoken. The fact that writing is a lonely activity whereas speaking typically takes place in an environment of social interaction causes written language to have a detached quality that contrasts with the involvement of spoken language.

Fragmentation and involvement are enabled and constrained by the "presence" condition of speech, which essentially means that speech prototypically occurs in the presence of an interlocutor (Bygate 2011: 417). The presence of an interlocutor leads to two further co-occurring conditions. The *reciprocity condition* involves the

dimension of interpersonal interaction in conversation. This essentially means that the speaker adjusts his/her speech according to the interlocutor's knowledge and expectations, and that s/he facilitates the interlocutor's participation (Bygate 1987: 7–8; Bygate 2011: 417). Speech is typically interactive, and Luoma (2004: 170) adds that from the perspective of language testing, this is what makes speaking special. The second condition is the so-called *time-pressure condition*. The immediacy of the interlocutor causes speech to be produced under time pressure, which means that the speaker's time to plan is restricted. This has observable effects, such as occasional overt editing. The speaker is also likely to be concerned with the need to allow the interlocutor to have his/her turn to speak (Bygate 1987: 7–8; Bygate 2011: 417).

Due to the dimensions of fragmentation and involvement as well as the two aforementioned conditions, speech is characterized by distinctive features. Chafe (1985: 106–108) argues that speech is structured around idea units, which he defines as clauses that are usually about seven words long or shorter and that are enunciated with a single coherent intonation contour, preceded and followed by some form of hesitation. This of course implies that speakers rarely speak in full sentences (Luoma 2004: 12). From a grammatical point of view, idea units can be defined as clauses. However, there is a difference in how the clauses are structured when comparing standard written clauses to those that typically occur in speech. When talking, we tend to emphasize certain elements either at the beginning of a clause (topicalization) or at the end (tails) (Luoma 2004: 15). In a written text, (pre)modified noun phrases are more normal than in talk, which means that in speech we are more likely to add one piece of information at a time (Brown and Yule 1983: 7). Furthermore, idea units are normally connected by coordination, rather than subordination, and the most typical conjunctions are *and*, *but*, *or* and *that* (Chafe 1985: 111; Luoma 2004: 12). Paratactic (i.e. unsubordinated) forms are more common because in speech pauses, rhythm, and to an extent intonation, play a significant role in expressing what the speaker wants to say (Brown and Yule 1983: 4). A further reason is that a speaker strives to be understood, and thus attempts to make the cognitive processing load lighter both for him/herself as well as the interlocutor. The idea unit, however, is considered very abstract[5], and for this reason, Foster *et al*. (2000: 365) propose the

[5] This is due to the fact that it is considered an intonational rather than syntactic unit. In L2 speech, intonation and pausing phenomena are problematic due to the inconsistency of such features in

use of the so-called AS-unit, which is defined as "a single speaker's utterance consisting of an independent clause, or a subclausal unit, together with any subordinate clause(s) associated with either".

Moreover, speaking and writing tend to differ in terms of their lexical range and degree of complexity. Chafe (1985: 114) makes a distinction between three types of words in the English lexicon: (1) those predominantly used in writing, (2) those most frequently used in speech, and (3) those that are neutral with respect to this distinction. The use of generic or vague words (e.g. *stuff*, *the round thing*, *thingy*) is very common in colloquial spoken interaction as is the use of demonstrative pronouns (e.g. *over there*, *that one*), because what is referred to is usually clear from the context (Luoma 2004: 17; Brown and Yule 1983: 6–7). When talking, we are also likely to use interactive expressions (e.g. *uhuh*, *well*) and fillers (e.g. *um*, *er*). These do not add meaning, but they keep the conversation going, help the speaker buy time and they also give speech a sense of fluency. Errors, repetitions and repairs are also very common in speech. All of the aforementioned features of speech result in information not being as densely packed in spoken language as it is in written texts (Brown and Yule 1983: 7).

Although speech is organized in particular ways, one should remember that speech is language, meaning that it is characterized by the same semantic and syntactic rules as written language. Fulcher (2003b: 24) claims that some researchers have in fact suggested that the differences between spoken and written language are not as great as they are made out to be. This is due to the fact that speaking is often associated with casual conversation and therefore, unplanned speech. A speech or a presentation, for example, is highly structured and usually well-planned, and is therefore closer to written language than a conversation, for example. However, as Bygate (1987: 10) states, "[s]peech is not spoken writing", and we rarely want to sound like a book.

Recognizing that speech is structurally different from writing has important implications for L2 testing of speaking, because as Brown and Yule (1983: 26) point out, "[i]f native speakers typically produce short, phrase-sized chunks, it seems

---

non-native speakers' speech, as they do not always mark the beginning or ending of an idea/message (Foster *et al*. 2000: 359).

perverse to demand that foreign language learners should be expected to produce complete sentences". The same goes for the other features of spoken language, which differentiate it from the written medium. Luoma (2004: 16–17), for example, states that,

> [m]any rating scales for speaking include descriptions of vocabulary use, and at the highest levels these often talk about being able to express oneself precisely and providing evidence of the richness of one's lexicon. […] Well-chosen phrases can also make descriptions or stories vivid, and learners who can evoke the listener's feelings deserve to be credited for their ability. However, very 'simple' and 'ordinary' words are also very common in normal spoken discourse, and using these naturally in speech is likewise a marker of highly advanced speaking skills.

This implies that learners and examinees should be given credit for both the use of ordinary and complex vocabulary, and I would argue that test tasks should elicit both. Underhill (1987: 38) also argues for a mixture of elicitation techniques and tasks in which the test-taker is required to do different things with languages, as this adds to the task's authenticity. Another factor contributing to authenticity is the construction of appropriate conditions for language use (Bygate 2011: 417). In testing a foreign language, the time pressure condition is relatively easy to create. The replication of the reciprocity condition, on the other hand, often poses difficulties due to the asymmetrical relationship between the tester and test-taker (Bygate 2011: 418; Young and Milanovic 1992; van Lier 1989).

2.2.2 Measures of complexity, accuracy and fluency

There is no universally acknowledged model of L2 proficiency (Drackert 2015: 34). However, researchers nowadays agree that L2 proficiency has a multicomponential structure which can systematically be captured by the notions of complexity, accuracy and fluency (henceforth CAF). Housen *et al*. (2012: 3) add that,

> [e]mpirically, factor analyses have identified complexity, accuracy and fluency as distinct and competing areas of L2 performance, implying that all three must be considered if any general claims about learners' L2 performance and proficiency are to be made.

These three dimensions of L2 production are significant in L2 language testing because, as Drackert (2015: 38) points out, they are implied in several language frameworks and very often assessment scales and rubrics include these in some form or another. The aspects of spoken language taken into account by the Common

European Framework of Reference (CEFR), for example, are the range, accuracy, fluency and coherence of speech as well as interaction[6] (Council of Europe 2001: 26–29).

The CAF measures are a contested topic for a number of reasons. Housen *et al*. (2012: 8) call attention to the number of existing methods[7] used for measuring complexity, accuracy and fluency, and claim that this in part reflects the lack of consensus regarding the definition of the CAF constructs. In the case of oral production, there is also disagreement over how to define and identify the unit of analysis. Based on the literature review conducted for this thesis, it would seem that the *Analysis of Speech Unit* (AS-unit) is the most reliable unit and thus the one most commonly used in the analysis of speech data (Foster *et al*. 2000; Ellis and Barkhuizen 2005).

Complexity is defined as "the extent to which learners produce elaborated language" (Ellis and Barkhuizen 2005: 139). 'Elaborated' can, on the one hand, mean that language learners differ in terms of how willing they are to leave their linguistic comfort zone and use more complex structures and vocabulary as well as language that is not fully automated or internalized. On the other hand, it can be operationalized to mean language learners' preparedness to use a wide variety of structures, which is determined by whether they want to experiment with the language and thus take risks. Moreover, researchers typically make a distinction between two different yet interrelated aspects of complexity, namely those of cognitive and linguistic complexity. *Cognitive complexity* is a subjective concept, which relates to the relative difficulty with which individual learners process language elements when performing or learning in the L2. On the contrary, *linguistic complexity* is objective, as it is independent of the learner, and "refers to the intrinsic formal or semantic-functional properties of L2 elements […] or to properties of (sub-)systems of L2 elements" (Housen *et al*. 2012: 4).

The notion of *accuracy* is more straightforward. Accuracy relates to "how well the target language is produced in relation to the rule system of the target language" (Ellis and Barkhuizen 2005: 139). In other words, it is a measure for how

[6] The CEFR is closely connected to how language ability is defined within the communicative language testing paradigm, particularly in the Finnish testing context (Huhta and Hildén 2016: 20).
[7] For an extensive overview of CAF measures, refer to Ellis and Barkhuizen (2005).

"error-free" language is with respect to a norm (usually a native-speaker). However, this concept entails its own problems, such as the question of what norm to abide by and, when considering the communicative competence approach, whether non-native speaker norms are equally valid. Housen *et al*. (2012: 4) further argue that accuracy should also encompass the notions of 'appropriateness' and 'acceptability' because in some contexts, non-native usage may be fully acceptable.

Finally, *fluency* is perhaps the most difficult concept of the three to define. This is because a distinction can be made with an overarching broad meaning (roughly parallel to the notion of global proficiency) and a componential narrow meaning (as one skill or "type" of phenomenon of speech among many) (Koponen and Riggenbach 2000: 5)[8]. Within the field of SLA, fluency is defined according to the latter as "the production of language in real time without undue pausing or hesitation" (Ellis and Barkhuizen 2005: 139). The definition has thus been narrowed down to it being a phonological phenomenon (Housen *et al*. 2012). This definition of fluency has also been criticized, as being a proficient speaker does not mean speaking without any disfluencies. However, there are differences in the distribution of L1 and L2 disfluencies, as L2 speakers have been found to pause more often within clauses compared to L1 speakers (de Jong 2016: 205–206; Lennon 2000: 25). Skehan (2009: 512–513) further distinguishes three sub-dimensions within the notion of fluency: speed fluency, breakdown fluency and repair fluency. *Speed fluency* refers to the density and the rate at which speech is produced. *Breakdown fluency* involves the analysis of pausing in terms of their number, length and location. Finally, *repair fluency* encompasses features such as false starts, reformulations, replacements and repetitions.

In the literature, there is also disagreement regarding the question of whether L2 speakers can simultaneously focus on all three CAF measures in their oral production. Skehan (2009: 512) proposes that learners/test-takers may be forced to prioritize one (or two) over the other(s) because humans have a limited information processing capacity, which leads to a trade-off in attention allocation when a task is performed. Figure 1 shows how test-takers might adopt a conservative stance and

[8] Lennon (2000: 25) terms these (i) *higher-order* (broad) and (ii) *lower-order* (narrow) fluency.

prioritize form over meaning, and thus accuracy and/or complexity over fluency, when completing a task. However, the opposite is equally possible.

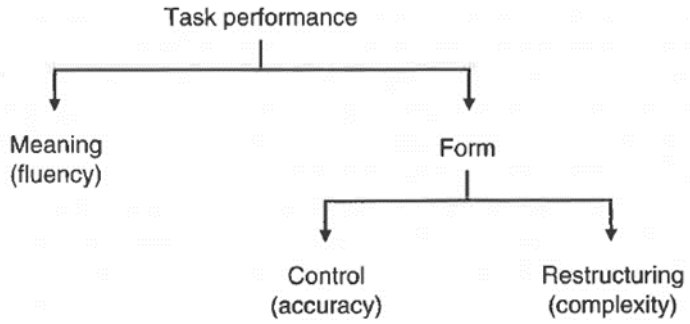


Figure 1 – *CAF measures mapped in terms of meaning and form (based on Skehan 1998)*

Ellis and Barkhuizen (2005: 141) claim that this 'trade-off' is caused by learners attempting to lighten the burden on their working memory by either prioritizing the content of the message or by adhering to linguistic norms. This model is called the *Limited Capacity Model*[9]. However, this model is contested by Robinson (2001), who claims that learners are able to draw on multiple attention pools simultaneously, implying that such trade-offs do not occur. He thus proposed the competing framework of the *Multiple Resources Attentional Model*. According to Housen *et al*. (2012: 6), empirical research so far does not incontestably support either model, implying that more research is required.

In this thesis, I am interested in seeing whether the testing mode (i.e. face-to-face and computer-based) affects CAF measures. As Housen *et al*. (2012: 3; 9) point out, complexity, accuracy and fluency may be manifested differently under different conditions of L2 use, and external factors such as a test-taker's personality or affective state (e.g. feelings of anxiety[10]) may affect performance. Task variables, too, have had an impact on CAF measures and thus the completion of a task.

2.2.3 Defining tasks

In any language test based on linguistic elicitation with tasks, it is important to define what is meant by *tasks*. In the literature, numerous definitions exist for tasks, and the

[9] This model has also been called the *Trade-off Hypothesis* (Skehan 2009: 512).
[10] Testing anxiety and anxiety related to speaking a foreign language should not go unacknowledged, as they have, to an extent, been found to affect test performance (Cassady 2010).

definition largely depends on the purpose it is used for[11]. Bygate *et al*. (2001: 11) offer an all-purpose definition:

> A task is an activity which requires learners to use language, with emphasis on meaning, to attain an objective.

In similar light, with a focus on speaking, Luoma (2004: 31) defines speaking tasks as:

> …activities that involve speakers in using language for the purpose of achieving a particular goal or objective in a particular speaking situation.

As one can see, these definitions are in line with the communicative language testing paradigm, which highlights the importance of *meaning* and the achievement of *communicative goals*. In fact, Chaloub-Deville (2001: 210) states that task-based pedagogy has shifted its attention away from the traditional focus on form to an approach promoting the importance of interaction and the achievement of communicative goals in addition to grammatical skills. Skehan (2003: 409) adds that SLA researchers have increasingly attempted to ensure sufficient focus on form within the task-based approach.

There are several different distinctions between speaking task types in the literature. Brown and Yule (1983: 107), for example, distinguish between four types of informational talk which differ in difficulty: description, instruction, storytelling and opinion-expressing/justification. A further distinction is proposed by Bygate (1987: 22–24), which is also based on information routines; he makes a distinction between tasks that are *expository*, i.e. description, narration, instruction and comparison tasks, and those that are *evaluative*, i.e. explanation, justification, prediction and decision tasks. Both Brown and Yule (1983) and Bygate (1987) state that speakers' use of language is different in each of these categories. In other words, if someone is good at telling a story, they are not necessarily equally good at justifying their opinion. For this reason and also to minimize the effect of test-takers' differing background knowledge on test performance, it is important to create tests which contain a series of tasks. Tasks should also differ in terms of the degree of cognitive load demanded (O'Sullivan 2008: 15).

[11] For an extensive yet non-exhaustive list of definitions of task, see Ellis (2003: 2–5).

Tasks can further be categorized according to how structured they are. A continuum can be imagined between tasks which are *open-ended* and those which are highly *structured*. Open-ended tasks allow room for different ways of fulfilling the task requirements, whereas structured tasks specify what the examinee should say (Luoma 2004: 48). Task qualities have also been found to have an effect on the CAF measures. Tasks which are based on concrete or familiar information increase accuracy and fluency, whereas tasks requiring information manipulation (e.g. narrative tasks) lead to higher complexity. Tasks that have a clear structure improve both accuracy and fluency. Interactive tasks, however, increase accuracy and complexity at the expense of fluency (Skehan 2009: 511–512). Dialogic tasks have thus been found to lead to greater attention to form (Ellis and Barkhuizen 2005: 143).

2.3 Computer-assisted language testing (CALT)

Over the last twenty to thirty years, computer technology has gained in importance within the field of language testing[12]. The term *computer-assisted language testing* (henceforth CALT) has been adopted to refer to language assessment that makes use of computer technology at the various stages of the testing process, including test design and development, delivery and the scoring and reporting of examinee test performance (Chapelle and Douglas 2006; Sawaki 2012; Suvorov and Hegelheimer 2014). The three main reasons for using computer technology in language testing are its efficiency, equivalency to paper-and-pencil tests[13] and its possibility for innovation (Suvorov and Hegelheimer 2014: 1). These advantages notwithstanding, in the case of speaking, there have been concerns as to whether it is possible to assess an examinees' ability to speak an L2 with technology validly and reliably enough. This question will be taken up in the following sub-section, in which both *direct* (i.e. face-to-face) and *semi-direct* (i.e. technology-mediated) tests of oral proficiency will be defined and discussed in terms of their advantages and disadvantages. This is followed by a sub-section discussing studies comparing direct and semi-direct tests.

[12] This is, for example, demonstrated by its increasing popularity among professional testing organizations (Qian 2009: 113). In fact, due to rapid technological advances, many global commercial language testing services (e.g. the Educational Testing Service (ETS), the Business Language Testing Service (BULATS), PTE Academic) offer computerized versions of their tests (Suvorov and Hegelheimer 2014: 6).

[13] In the field of language testing, paper-and-pencil tests have long been considered the "gold standard" (Suvorov and Hegelheimer 2014: 1).

2.3.1 Direct tests of oral proficiency

Clark (1979) provides the basis for distinguishing three different modes of speaking assessment, namely indirect[14], direct and semi-direct tests. Testing can be said to be *direct* when it requires test-takers to perform the skill that the test sets out to measure (Hughes 1989: 15). In the context of speaking, Clark (1979: 36) defines *direct* (also called *live*) *testing* as

> […] procedures in which the examinee is asked to engage in a face-to-face communicative exchange with one or more human interlocutors.

The human interlocutor can be a teacher, an interviewer or an examiner (Isaacs 2016: 133). In this format, the test-taker is therefore required to engage in face-to-face oral communication and to perform oral tasks which demonstrate his/her L2 oral proficiency. The direct face-to-face test setup is depicted in Figure 7 in Appendix 1.

An influential example of direct testing is the *Oral Proficiency Interview* (henceforth OPI), the first published test of speaking (Fulcher 2003b: 8). The OPI was designed in the 1950s to test the functional foreign language proficiency of those who worked for the Foreign Service Institute of the US Department of State. It was later adopted by the Educational Testing Service (ETS) and the American Council on Teaching of Foreign Languages (ACTFL) and was consequently widely spread into contexts of foreign language teaching and testing (O'Loughlin 2001: 4–5; Johnson 2001: 6–8). The OPI is a structured interview which consists of a warm-up phase meant to put the candidate at ease, followed by a level check and probes which determine the examinee's highest sustainable level of speaking proficiency, and finally a wind-down phase. The OPI uses a variety of question types and role-plays as elicitation devices (Johnson 2001: 11–13).

[14] *Indirect* tests of speaking, which belong to the "pre-communicative" era, refer to tests in which the abilities that underlie a particular skill (e.g. speaking) are tested (Qian 2009: 114; O'Loughlin 2001: 4). Lado (1961), for example, proposed testing pronunciation by asking examinees to listen to a series of words and identify the word which is pronounced differently; in other words, the test-taker is not required to speak. This is considered an outdated method which lacks in validity (O'Loughlin 2001: 4) and will therefore not be discussed further.

The OPI became the 'standard' after its introduction in the 1950s, making direct tests the most common mode for assessing oral proficiency[15] (Luoma 2004: 35). According to Underhill (1987: 31), the direct interview is "the […] most authentic type of oral test for normal purposes". However, oral proficiency interviews have been found to be a problematic form of assessment. This is mainly due to a power imbalance, or what van Lier (1989: 496) calls *asymmetry*, between the examiner and examinee. It leads to skewed turn sizes and distribution (Johnson 2001: 73), whereby the tester has a controlling and the test-taker a reactive role (Young and Milanovic 1992: 1). According to Ussama and Sinwongsuwat (2014), interview-based tests often result in what they term *institutional talk*. It does not represent natural conversation or provide students with a fair opportunity to demonstrate their interactional ability. For this reason, since the end of the 1980s, pair and group tasks have become a more frequent means of assessing oral proficiency[16] (Ussama and Sinwongsuwat 2014: 97). O'Loughlin (2001: 5) points out that a more structured, task-based approach to direct oral testing has become more popular.

Face-to-face tests are subject to further limitations. Firstly, unless the test is highly structured, the test may be administered in different ways to different people. *Inter-rater reliability* (i.e. that different raters rate a performance similarly) is therefore an important issue (Luoma 2004: 179). Brown (2003), for example, discovered that testers may adopt different elicitation techniques, resulting in different assessments of the same candidate, and one of the interviewers was also found to be consistently more lenient in comparison to the other tester. Finally, there is the issue of practicality, especially within the context of large-scale high-stakes tests. The administration of direct tests is resource intensive, as the tests must often be administered to candidates one by one. Not only does this take time, but it is also very expensive; it is thus neither cost-effective nor efficient (Qian 2009: 115).

[15] Luoma (2004: 35) adds that for the following thirty years the OPI format was not questioned and remained the norm.

[16] However, pair/group test formats also present challenges. Not only is meaning co-constructed during the performance, a participant's personality, level and communication style is also likely to have an influence on the pair's or group's performance (Luoma 2004: 37–38).

2.3.2 Semi-direct tests of oral proficiency

Issues related to practicality and reliability are, according to Fulcher (2003b: 2), the main drivers of research into *semi-direct* tests of speaking and automatic/machine scoring of speech[17]. According to Clark (1979: 36) semi-direct tests are organized

> […] by means of tape recordings, printed test booklets, or other 'non-human' elicitation procedures, rather than through face-to-face conversation with a live interlocutor.

This definition of semi-directs is partly outdated, as semi-direct tests nowadays are administered with the help of a computer. In a tape-mediated test, examinees typically listen to a recorded stimulus on the basis of which they are then asked to react. The test-takers then record their own response, which is finally listened to by the listener/assessor. Tape-mediated and computer-based tests differ in terms of the stimulus. In a computer-based test, the stimulus can be recorded, or it can manifest itself in the form of an avatar. Additionally, computer-based tests can also incorporate multimedia (pictures, videos, etc.) which can act as the stimulus. These two related formats are shown in Figures 8 and 9 in Appendix 1. Examples of semi-direct tests include the simulated oral proficiency interview (SOPI), developed in the 1970s, and the computer-based oral proficiency interview (COPI/OPIc)[18].

Semi-direct tests have a number of advantages compared to direct tests. They simplify the administration process because they allow for bulk and uniform administration (Kiddle and Kormos 2011: 344). Examinees can therefore be tested under identical conditions. As a result of technology, speech has gained a form of permanency, which frees the assessor from having to act as interlocutor and thus allows him/her to exclusively focus on the rating procedure. Moreover, due to functionalities available on computers, it is possible to incorporate tasks that integrate different modalities, which in turn enhances the authenticity of the test (Sawaki 2012: 428).

[17] Isaacs (2016: 133) argues that such technological developments have also made it possible to study speech using more precise scientific measurement.

[18] The COPI is a computer-adaptive test, i.e. a computer-based test which adapts to the test-taker's ability level with the help of an algorithm (Chapelle and Douglas 2006: 7). The OPIc, on the other hand, is an internet-delivered semi-direct test (Malone and Montee 2010: 981).

However, not all researchers are as convinced by these apparent benefits. A serious concern raised with regard to the use of technology in language assessment is that an examinee's performance on a CALT test may fail to reflect the same ability as that measured by a face-to-face interview (Chapelle and Douglas 2006: 42). This is because, as Mousavi (2009: 40) claims, "[c]omputers are not able to act as fully-fledged conversation partners." Underhill (1987: 35–36) adds that semi-direct tests lack authenticity due to the absence of interaction and non-verbal features of communication. Although an increasing amount of communication occurs in computer-mediate environments (Sawaki 2012: 430), real-life communication still typically takes place face-to-face, thus making the semi-direct test inferior (van Lier 1989). Moreover, in high-stakes, large-scale testing, there is always the question of test security and the need for vast, up-to-date item banks (Ockey 2009: 839). Finally, there is always a possibility that technical difficulties interfere with a test-taker's performance (Underhill 1987: 35–36; Kiddle and Kormos 2011: 344).

Perhaps the most obvious way of finding out whether a test-taker performs well on computer-based test is by comparing examinees' performance on two similar tests which differ only in terms of the mode of delivery (Chapelle and Douglas 2006: 43).

2.4 Previous studies comparing oral proficiency testing modes

Previous studies concerning the comparability of face-to-face and computer-based oral tests can be divided into two categories: those examining score comparability, on the one hand, and those analyzing linguistic/interactional comparability on the other. This division roughly coincides with that of quantitative and qualitative methods, although some studies make use of both methods. The predominantly quantitative studies investigate the concurrent validity and thus interchangeability of the two testing modes, whereas the qualitative studies have looked at a variety of features regarding the discourse produced in a direct and a semi-direct test.

2.4.1 Previous studies on score comparability and test-taker preferences

There are a number of studies which have compared direct and semi-direct tests in terms of score comparability. Most of the studies discussed in this section have found high correlations between the two testing modes. The results regarding test-taker

preferences are mixed, however, but participants have generally indicated a preference for the face-to-face test.

Kenyon and Tschirner (2000) report on a study which aims to compare test reliability and student performance in the German Speaking Test (a semi-direct tape-mediated test) and the ACTFL OPI. A randomly selected group of 20 university students took both tests which, together, counted for 15% of their final grade. Although the tests elicit speech in slightly different ways, the results indicate that final ratings agreed perfectly in 90% of cases (*ibid*.: 91). However, this agreement tended to only occur at higher proficiency levels. This may either arise from the fact that lower proficiency examinees' performance was qualitatively different in the two tests or it may be an indication of the SOPI being less reliable at lower proficiency levels (*ibid.*: 99). Furthermore, Kenyon and Tschirner (2000: 96) established that in terms of test reliability, the SOPI was slightly more reliable. Reliability on either test was found to significantly increase if the test performance was double-rated and arbitrated. Unfortunately, this study gives no indication of participant preference for one testing mode over another.

However, a year later Kenyon and Malabonga (2001) studied examinees' attitudes toward a tape-mediated (SOPI) and a computer adaptive test (COPI). Although the study mainly focused on the comparison of these two technologically-mediated tests, the undergraduate and graduate students taking Spanish (n=24) were additionally requested to comment on their experiences taking a face-to-face oral proficiency test (OPI) in addition to the two other tests. The data was collected by issuing two post-test questionnaires, one after each individual test and another concerning all three. The questions focused on aspects of difficulty, fairness, nervousness, clarity and accuracy (*ibid.*: 66–67). In order to avoid effects of task repetition, students were grouped so that they would complete the three tests in a different order.

The results indicate that the examinees' performance in the three tests was similar, but the SOPI/COPI ratings tended to be higher than the face-to-face interview rating. The OPI and COPI had a rank order correlation of .92, while that between the SOPI and OPI stood at .94 (Kenyon and Malabonga 2001: 70). Furthermore, although the students felt that they had the opportunity to adequately demonstrate their strengths and weaknesses in speaking Spanish in the two technologically-mediated tests, they

still felt that the face-to-face test better provided the opportunity to do so. Moreover, the examinees felt that the OPI gave a more accurate picture of their ability to speak Spanish in a real-life context; the technology-mediated tests were mainly seen as being one-sided. In other words, the majority felt that both of the technologically-mediated tests missed the conversational, interactional, and personal nature of a real-life conversation. In several of their comments, participants expressed doubt as to whether a technology-based test could ever capture these essential aspects of speaking (*ibid.*: 80.) This is demonstrated by the following comment:

> I don't feel technology based [*sic*] tests can come so close to having interaction with a person. "Speaking" with another person is the essence or key goal you want to capture with a foreign language testing [*sic*], and with technology you are not "speaking" with a person in either situation. (*ibid.*: 74)

Kenyon and Malabonga (2001: 81–82) thus conclude that "[o]ral communication remains a human phenomenon" and that this raises the question of what is being assessed when it comes to oral assessment. It all comes down to how speaking is defined as a construct. In other words, if interaction is considered an essential component of the construct, technology-mediated tests may not be the best option. Kenyon and Malabonga (2001: 82) add that "…it may be quite a while before [interactional competence] can be replicated with technology".

Surface *et al.* (2008) conducted a large-scale validation study of the web-based OPIc, and their findings were very similar to the aforementioned study. Among the many aims of their study was the comparison of the final ratings on the ACTFL OPI (telephonic interview) and OPIc, as well as participants' views of the two tests[19]. 99 employees from a Korean company participated in the study. The participants were divided into two groups according to the order in which they would complete the two tests. The second group completed the OPIc twice. This was due to the researchers additionally analyzing test-retest reliability. The participants were also requested to fill in pre- and post-assessment questionnaires.

Surface *et al.* (2008: 23) found the correlations between the two test formats (in both administrations) to be statistically significant and to indicate a strong positive

[19] The study conducted by Surface *et al.* (2008) consists of two separate validation studies. Due to its superior relevance, this paper only considers the first one.

relationship. The two tests were additionally found to measure the same construct (*ibid.*: 25). Furthermore, they found that although attitudes toward the OPIc were generally positive, test-takers preferred the OPI to the OPIc and felt that the OPI offered a better opportunity to demonstrate their language abilities. On the post-assessment survey 44 of the participants indicated that they were better able to demonstrate their oral proficiency in the OPI. On the other hand, 10 claimed that they were better able to do so in the OPIc. 15 of the participants indicated that they liked both assessment formats (*ibid*: 28). The participants' comments shed some light on the question of why they prefer one test mode to another:

> In the OPI, you can ask to slow down a bit when you are not feeling sure about the questions at hand but with OPIc, since you are talking with a computer you cannot make such a request. (*ibid.*: 73)
>
> The avatar made me feel uncomfortable. (*ibid.*: 74)
>
> Inevitably speaking to a computer causes less tension than talking with someone on the phone. (*ibid.*: 73)

Although both of the studies provided evidence for the validity and reliability of the OPIc, Surface *et al*. (2008) highlight the need for further research and the development of testing instruments (e.g. the quality of the avatar).

Kiddle and Kormos (2011) studied the effect of mode of response on a semi-direct test of oral proficiency in a Chilean university context, where there is a dire need for alternatives for face-to-face oral proficiency tests for geographical reasons. They constructed two versions of a speaking test, in which the test content was identical. Both tests consisted of three tasks, all of them containing video input (*ibid.*: 345). A total of 42 participants took part in the study; the test-takers were split so that half of them participated in the face-to-face and the other half in the computer-based test first. After a period of three weeks the participants took the other test (*ibid*.: 347).

The results suggest that the testing mode does not have a significant effect on the scores and the many-facet Rasch analysis shows no significant difference in test difficulty. However, there was some variation in performance, as 33% of the participants were awarded a different band score in the two tests (*ibid*.: 352). The mean scores for fluency and delivery on the two tests were similar, but the examinees' task achievement was assessed slightly higher on the CBT, whereas

grammar, vocabulary and pronunciation were rated higher on the face-to-face test (*ibid*: 349–350). Kiddle and Kormos (2011: 353) add that

> …in the face-to-face version of the test, students might have paid more attention to the accuracy of performance at the expense of the content of the task as evidenced by the slightly higher vocabulary and grammar scores in the face-to-face versions and by the mean values of the item in the questionnaire that enquired into concerns about mistakes.

This would suggest that the testing mode does in fact have an effect on test-takers' linguistic output. Furthermore, although the participants' perceptions were generally found to be favorable concerning both tests, the majority of candidates still indicated a strong preference for the face-to-face test which may suggest that they "do not fully accept the computer-administered test as equivalent to the face-to-face test" (*ibid*.: 354).

More recently, Thompson *et al*. (2016) conducted an analogous study with 154 Spanish language learners of varying proficiency levels, with a focus not only on student preference, but also on their achievement in the two tests. The OPI was organized as a telephone-mediated interview and the computer-based exam used was the OPIc. They issued a pre-test background information questionnaire, asking about language experience, goals and familiarity with the two test setups, and a post-test questionnaire asking for a self-evaluation and attitudes regarding the two test formats (*ibid.*: 80). Interestingly enough, students generally performed better on the OPIc, but tended to prefer the more traditional OPI. Those who preferred the telephone-mediated format justified their opinions by claiming that it felt more natural, as they had the possibility of receiving feedback, the provision of discussion topics was superior, and they had more time, since they were not cut off:

> I prefer the OPI. The conversation came more naturally. Plus, the fact that the interviewer was able to relate each question to my previous response allowed an easier transition from topic to topic. This also allowed me to think for myself which vocabulary I wanted to use and which direction I'd like to take the conversation. The experience with the OPI seemed more natural and realistic, which I liked. (*ibid.*: 86)

The minority who preferred the computer-based test gave reasons such as flexibility, the possibility to repeat questions by the push of a button and, as the following excerpt indicates, feeling less anxious due to lack of real personal interaction:

> The OPIc. I don't know why people would be more uncomfortable speaking to an avatar than to a real person; it seems counterintuitive. Speaking with a real person is incredibly nerve-wracking […] The OPIc was always very clear, and I didn't feel worried about what the other person would think of my answers or how they would respond. (*ibid.*: 87)

Thompson *et al.* (2016: 90) additionally claim that the results of their study clearly indicate that students' preferences are affected by "personal characteristics and preferred interpersonal style" when selecting an examination format. This is clear from the examinees' comments.

Mousavi (2009) discovered that the 30 participants in his small-scale study had very positive perceptions regarding digitally delivered proficiency tests and that they actually felt more comfortable when completing the semi-direct test. The test-takers commented that the CBT was also less threatening and that it could create a positive washback effect:

> I think this test has a good potential to be used even for practicing for language exams. This could also be a good attraction for students who feel intimidated talking to 'real' people especially for exams when they already feel nervous about doing an exam.

Mousavi (2009: 45) acknowledges the important fact that besides the effect of testing mode, test-takers individual differences may contribute to the reactions to the tests. Factors such as different levels of proficiency, language and cultural background and differing computer familiarity may also contribute to these reactions. Qian (2009: 123) also makes a similar point, claiming that test-taker attitudes may be influenced by a number of different factors ranging from test quality, the stakes of the test to cultural traditions and test-takers' personalities.

2.4.2 Previous studies on linguistic or interactional comparability

The previous studies focusing on linguistic/interactional comparability, on the other hand, show that the tests may not necessarily tap the same type of skill. Shohamy (1994: 102) argues that high correlations between tests are important, but this does not suffice as evidence for their exchangeability. In order to determine concurrent validity between tests, it is necessary to examine the language samples elicited by them. In her study, Shohamy (1994) thus compared the OPI and the SOPI in terms of the language they elicit. The two tests were found to have high concurrent validity. However, after performing a number of different qualitative analyses Shohamy found

there to be differences in the elicitation tasks in terms of the number and types of functions and topics as well as in the language samples obtained regarding communicative functions and discourse features.

There are significant differences in genre between the two tests, as the OPI is a conversational interview, whereas the SOPI is a form of reporting (i.e. monologic). Shohamy therefore makes the observation that in the OPI examinees are more likely to be focused on the transmission of information, while in the SOPI test-takers are more concerned with linguistic accuracy, which is indicated by the more frequent use of self-correction (*ibid.*: 115–116). Due to the fact that in the SOPI examinees are more interested in simply completing the task rather than adapting their speech to an unseen interlocutor, a prioritization of form over meaning occurs. Indicative of this is the more concise language, the increase in lexical density and the fact that the speech is more formal and cohesive than in the OPI (*ibid.*: 117).

The opposite happens when students complete the OPI, as the process is more dynamic because the language and topic are congruent to the feedback obtained from the interviewer. In other words, there are "signs of contextualization" (*ibid.*: 117). Shohamy thus concludes that context alone appears to be more powerful than the elicitation tasks themselves and therefore, even if a test were to include varied functions in the elicitation tasks such as those in the SOPI, there is no guarantee that it would lead to the production of varied language. Conclusively, "[t]he context of the test, either 'face-to-face' or 'tape-mediated', can affect or even dictate the type of language that is produced" (*ibid.*: 118).

O'Loughlin (1997) studied the lexical density of the responses of test-takers in two test formats, live and tape-mediated, in the *access: test*. Both tests contain four different tasks: description, narrative, discussion and role play. The results of his study indicate that the tape-mediated test elicits a significantly higher level of lexical density (*ibid.*: 169). According to O'Loughlin (1994: 171), this is in line with Shohamy's (1994) findings, although the differences — albeit being statistically significant — were not as large as in her study. He additionally points out that out of the four tasks, the role play had the lowest level of lexical density, which is most clearly explained by the fact that out of the four it is the most interactive task. In addition, open tasks such as the description task and the interview, in which

"candidates are not constrained by any stimulus material and may therefore be able to display a greater range of their lexical resources", are more likely to result in higher lexical density (*ibid*.: 172). He concludes that the two tests cannot safely be substituted for one another.

The same year, Luoma (1997) conducted a triangulation study in her licentiate thesis in which she analyzed test-taker and assessor perspectives[20], test discourse as well as the assessment process with regard to the speaking component of the Intermediate level English test[21]. A total of 37 candidates and 2 assessors took part in the study. The participants completed both the face-to-face and tape-mediated versions of the test. A wide variety of quantitative and qualitative data was collected by means of test scores, audio/video recordings, transcripts, test-taker and rater questionnaire feedback and interviews. These were analyzed to determine the degree of comparability between the two tests.

The results indicated that the tests correlated strongly, but qualitatively both similarities and differences were found. Although the tests consisted of different tasks, the candidates used similar words and grammatical structures in the two tests. Linguistic differences were mainly found in test-takers' use of hesitation markers and discourse particles, with hesitation markers being used more in the tape-mediated test (*ibid*. 99) and discourse particles more frequently in the face-to-face test (*ibid*. 98). However, the differences mainly relate to the nature of the two tests. In other words, the tape-mediated test conducted in the language laboratory was highly structured, which meant that answers were relatively predictable, whereas the face-to-face format was more flexible and allowed the examinee to more freely influence what was measured in the test due to its interactive nature (Luoma 1997: 129–130). Moreover, test-takers indicated a preference for the face-to-face test. Luoma (1997: 131) concludes that

> [t]he overall conclusion from the three studies was that while there was plenty of similarity between the tests, the constructs behind the two tests were to an extent different. Whether the tests were comparable is thus not easy to judge. Especially if comparability is interpreted to mean

[20] The assessor perspectives and the assessment process are not discussed in this thesis.
[21] This test forms part of the National Certificate of Language Proficiency in Finland. These language proficiency tests are targeted towards adults who require evidence of their language skills in practical situations.

> exchangeability, the yes/no decision depends on how important the differences are to the one asking the question.

Luoma (1997: 131–132) argues that in order to choose between the two formats, issues of practicality and face validity should be taken into consideration. Moreover, in the face-to-face test it is important to make sure that there is enough structuring so as to treat test-takers equally and to accordingly ensure test comparability.

All of the aforementioned previous studies have compared direct and semi-direct tests. The results appear somewhat mixed as to the question of test comparability and/or interchangeability. Furthermore, although there are individual differences, participants have generally indicated a strong preference for the face-to-face test due to its interactive nature. However, computer-based tests are potentially more interactive than tape-mediated ones, and in fact, Kenyon and Malabonga's (2001) study showed that participants tended to favor the COPI over the SOPI. However, it is not possible to draw any definitive conclusions from these studies because they differ not only in terms of scale but also in terms of methodology. Many of the studies' findings are based on a comparison of tests which differ in their format and the type(s) of tasks used in each test. The types of analyses carried out are also not consistent. Moreover, it is important to note that these studies have been conducted in various geographical locations and most of them are situated in a post-secondary institutional context. My study is situated in a Finnish secondary school context, so generalizations cannot be made in this respect either. To the best of my knowledge, no previous studies have analyzed the complexity, accuracy and fluency of oral performance on direct and semi-direct tests.

2.5 The Finnish matriculation examination

According to Atjonen (2015, as cited in Pollari 2016: 188), the testing tradition in Finland differs from the Anglo-American one. This is demonstrated by, for example, the lack of teacher accountability and centralized, standardized high-stakes examinations. In Finland, school assessment is for the most part low-stakes and tests are normally designed by the teacher. The only external high-stakes examination in the Finnish context is the matriculation examination, which is the official, nation-wide school-leaving examination. As is typical for national examinations, the Finnish matriculation examination, too, is centralized, standardized and based on a

national curriculum. It is designed outside the school by an examination board (Pižorn and Huhta 2016: 239).

The matriculation examination, which started off as a university entrance examination in 1852, became the final upper secondary school examination in 1919 (Pollari 2016: 188). According to Huhta and Hildén (2016: 11), the form of the tasks in the language tests as well as how they were carried out remained unaltered for the first one hundred years of its existence. The written, translation-based exam, which focused on grammar and drew influence from the pedagogy of classical languages, maintained its status up until the 1960s. In fact, the way that language skills have been defined in language tests at the end of upper secondary school has been influenced by the communicative and pedagogic emphases and values of a particular time period (Huhta and Hildén 2016: 8).

A listening comprehension part was officially introduced in 1974, giving the exam a three-part structure. The other two parts consist of reading comprehension tasks and writing a short composition. The foreign language test has become more versatile, and the current exam contains a variety of different task types. Testing students' productive skills has also consistently gained in importance (Matriculation Examination Board).

However, the lack of an oral component in the matriculation examination has puzzled language assessors in Finland for a long time (Huhta and Hildén 2016: 13). The introduction of an oral test section to the matriculation examination was proposed as early as 1958. The newly established Federation of Foreign Language Teachers in Finland (SUKOL) was put in charge of the project. However, the idea was rejected on the grounds of the change being too "sudden" and not planned well enough (Saleva 1994: 278). The issue was brought up again in 1988, and the National Board of Education appointed a group responsible for charting the possibilities of organizing an oral skills test in connection with the matriculation examination and for investigating the necessary measures for its implementation. The resulting report stressed the need to improve conditions for teaching oral skills (Ministry of Education 2006).

Between 1990 and 1993 experiments were conducted in four municipal upper secondary schools with the aim of investigating whether it would be feasible to teach and assess oral skills. A related PhD-study attempted to find out whether an oral test could be organized in the form of a mass language laboratory test (Saleva 1997: 11–12). Although Saleva found the language laboratory test to be a valid and a reliable option and thus argued in its favor, after the experiments, it was decided that an oral component would not be added to the examination. The working group did, however, make some suggestions. They were in favor of an optional oral test, the assessment of which would be based on the teachers' continuous monitoring of oral language skills (Ministry of Education 2006: 9). In 2006, the Ministry of Education again undertook an investigation into advancing oral skills testing, and as a result, in 2010, an optional advanced oral skills course was incorporated into the foreign languages and second national language syllabuses (Finnish National Board of Education 2010).

The National Core Curriculum for general upper secondary schools (Finnish National Board of Education 2016: 114) (henceforth NCC) accentuates that "[t]he instruction of languages is based on a broad definition of text, according to which a text refers to both spoken and written language". Furthermore, although there is variation in emphasis between courses, both oral and written interaction should be practiced in versatile ways (NCC 2016: 117). The aims of the English A-syllabus specialization course, *Speak and influence* (ENA8), as stated in the NCC (2016: 119) are as follows:

> The students advance their skills in producing language orally, understanding spoken language, and building dialogue. They strengthen their fluency of speech and practise oral production that requires preparation. The themes dealt with in the compulsory courses are revised or complemented according to the students' needs.

Pollari (2016: 190) adds that despite the fact that all of the English courses have separate themes, they comprise all areas of both oral and written language skills. Therefore, course assessment ought not to exclusively focus on just one area (e.g. grammar or writing) but should include them all. This implies that speaking as a skill should be promoted in all of the courses. In fact, as Huhta and Hildén (2016: 10) point out, the common goals of all languages emphasize the encouragement to use the language in many ways in different contexts.

Pollari (2016: 184), who examined the expectations and experiences of 142 second- and third-year students from an upper secondary school towards the English matriculation examination, found that, "[a]lthough the test did not seem to cause excessive washback, it caused significant stress and anxiety". However, she did find that "teaching to the test" seemed to increase as the exam approached (Pollari 2016: 196). Furthermore, students were relatively critical of the English exam's validity; one of the main reasons for this was the apparent lack of an oral skills test component (Pollari 2016: 200):

> In my opinion, the test was good but to my mind an oral test should be part of the package because oral communication is important.
>
> The test is deficient in the sense that it doesn't measure the student's ability to communicate orally in English.

In addition to the students criticizing the lack of oral skills testing, they considered "too detailed knowledge related to grammatical exceptions or rare vocabulary […] irrelevant for real-life communication skills." (Pollari 2016: 205).

Moreover, like many high-stakes tests, the Finnish matriculation examination is well-known for its washback effect (Saleva 1994: 227), which is going to increase in the near future. The student selection process for higher education is being renewed with the aim of reducing gap years and advancing the start of further studies. Matriculation examination results are meant to increasingly determine entry to higher education. As of this year, the matriculation examination is fully digitalized. Moreover, the digitalization of language tests opens up new possibilities: with the help of speech recognition software, it is possible to automatize part of the assessment process by electronically evaluating some of test-takers' responses. The development of such a speech recognition instrument is underway as part of a cross-disciplinary project called DigiTala. The aim of the project is to introduce speech recognition to the second national language (Finnish or Swedish) matriculation examinations, after which it is meant to be extended to the assessment of foreign languages. The assessment is meant to be carried out by both the instrument and human evaluators (DigiTala 2016; Huhta and Hildén 2016: 13–14).

3 Data and methods

In the following section the data collection process and the methods are described in detail. The data, which consists of the informants' performance results on the two tests as well as their responses to an online post-test questionnaire and focus group interview, is presented first. This entails a description of the informants and the two tests and their tasks. This is followed by an explanation of the methods which have been chosen to answer the two main research questions. These include the recording and subsequent transcription of the test responses and their analysis in terms of the CAF paradigm. More specifically, the complexity, accuracy and fluency measures that have been selected for the purposes of this paper are presented and justified. This section concludes with a description of the analysis adopted for the questionnaires and focus group interviews.

3.1 Data

The data consists of 15 upper secondary school students' oral test performance in the face-to-face and computer-based tests in the form of detailed transcripts as well as their responses to an online post-test questionnaire and successive semi-structured focus group interviews.

3.1.1 The informants

The informants who participated in this study are students from an upper secondary school located in the Helsinki metropolitan area. At the time of data collection, the informants were taking one of the following courses: ENA1, ENA5 or ENA7. The first two courses are mandatory for students completing the English A-syllabus, whereas ENA7, *Sustainable way of living*, is a national specialization course and therefore optional (Finnish National Core Curriculum 2016).

There were two groups taking ENA1, a third group taking ENA5 and a fourth group taking ENA7. From each group, three to four students were chosen on a volunteer basis. The teacher was asked to send the students and their parents a form of informed consent (see Appendix 2) ahead of time, in which information was provided about the aims and structure of the study. If the informants were under the age of 18, they were asked to get permission from their parents / legal guardian.

3.1.2 Designing the tests

The data was obtained by having the informants complete two oral tests: a face-to-face and a computer-based test. The two tests were planned and designed simultaneously so as to ensure comparability. An important aspect concerning test design that had to be considered early on was whether to make the two tests identical or whether to construct the tests using the two modes to their fullest potential. This is because, as Chapelle and Douglas (2006: 39) point out, "…computer technology expands the test developer's options for constructing language test tasks". As a test designer, one must therefore deliberate whether to take advantage of functionalities (e.g. video input) available only on computers. In the same way, one has to also consider whether to make the face-to-face test highly interactive (i.e. dialogic), a feature thus far not replicable on computers.

Moreover, Mousavi (2009: 37) emphasizes that when a test is delivered through a new medium, the medium becomes a factor that likely affects the nature and quality of test-takers' oral production. As a result of potential test method effects, test developers are justifiably concerned about the use of computers having an influence on what is being measured in a test. Having taken all of this into account and due to the fact that I am interested in seeing whether the testing mode alone is in fact a factor that affects the way the informants speak, I decided to design two tests that are similar in terms of both content and task types.

Furthermore, in computer-based testing, where usability problems may constitute a threat to construct validity, interface development and design are extremely important (Fulcher 2003a: 384). Therefore, before collecting the data, the two tests underwent two rounds of piloting. Altogether seven students from two separate ENA5 courses volunteered and completed the face-to-face and the computer-based test in succession. The students were then asked for feedback: comments concerning the tasks, instructions and overall experience were collected with the help of a questionnaire and a small-scale interview. The feedback, which mostly entailed suggestions for improvement regarding the tasks, was incorporated into the final test design.

The final face-to-face test was designed so that each of the four tasks (including the instructions and the prompts) was presented on a separate piece of paper. The tasks were then administered one by one. As the test administrator, I read the instructions out loud and then handed the paper over to the test-taker, giving him/her time to go over the instructions once more and familiarize him/herself with the task before starting to speak. I then took on the role of interlocutor. The preparation time was not limited, but the test-takers were told that, due to time restrictions, the total test time could not exceed 20 minutes.

The computer-based test was designed on Moodle (version 3.2.2+), an open source platform and a modular system based on plugins that can be adapted in the creation and administration of computer-based language tests (https://moodle.org/). I downloaded Moodle onto my own server, which requires a web server with PHP and a database. Before designing the tests, I created an admin account, which authorizes me to make changes to the platform content. I then created the course, *Testing speaking skills* (TSS), which contains four tasks similar to the face-to-face test. The instructions were provided on the Moodle page, but this time, I did not read them out loud. Instead, general instructions were given at the beginning of the test. In this testing situation, I was both administrator and invigilator. Again, the preparation time was not limited, but the testing time was restricted to a maximum of 20 minutes. In both of the tests, the instructions were given in Finnish, so as to avoid any misunderstandings during the completion of the tests.

There are both advantages and disadvantages to using Moodle for testing purposes. On the plus side, it is relatively user-friendly and is therefore commonly used in educational settings. It is also flexible and highly customizable. Also, depending on what one intends to test[22], the question bank offers the possibility of designing a variety of test task types, ranging from selected to constructed response items (Suvorov and Hegelheimer 2014; Douglas 2010: 60). One of the disadvantages, however, is the fact that Moodle is typically not used for administering high-stakes tests, meaning that I could not construct an authentic high-stakes test setting such as that of the matriculation examination. According to Suvorov and Hegelheimer (2014:

[22] This is an important question to consider because, as Fulcher (2003a: 388) states, the choice of authoring software depends to a great extent on the type of items in the test as well as on the functionality needed in the final product.

9), because it is "designed for teaching and learning purposes in a variety of educational settings", Moodle is typically only used for low- or medium-stakes assessment purposes.

3.1.3 Designing the tasks

I chose four different task types in order to provide test-takers with the possibility to demonstrate their abilities (see Appendices 3 and 4). The task types include:

1. a narrative task,
2. a comparison task,
3. an instruction task and
4. a role play / simulation task.

The first task type, the narrative task, is typically based on picture sequences, where the picture content largely determines what will be said (Luoma 2004: 144). In both of the tests, this task consists of a cartoon strip which the test-takers are asked to observe. They're then required to narrate the story it depicts. The comic strip in the face-to-face test has two characters, a man and his son, who are out on a ride when the car breaks down. In addition to recounting the events, test-takers are asked to describe the characters and to mention what they think is wrong with the car. In the computer-based test, the comic strip shows the same man and son as well as a third character, the mother, who has prepared dinner. She asks her husband to get their son, who is in his room reading a book. The boy joins his mother at the table, but the father has stayed behind to read his son's book. In addition to describing the plot and the characters, the examinees were asked to reflect on the type of book the father and son are reading.

The second task consists of two pictures that the test-takers are asked to compare. This type of task was chosen because, as Luoma (2004: 147–148) claims, compare and contrast tasks entail a greater cognitive load than description tasks for their completion due to the fact that test-takers are required to analyze and discuss similarities and differences, which in turn requires the use of comparative forms and complex grammatical structures. In the face-to-face task, the two pictures show a male student working alone and a group of students working together. The informants were asked to name advantages and disadvantages of working alone vs. in a group and to justify their opinion regarding their preference of working style. In the computer-based test, the pictures depict two different school settings: one is located

in a developing country and is highly teacher-centered, while the other one is in a developed country where technology is used in the classroom. In addition to again pointing out the differences, the informants were asked to list pros and cons of using technology in teaching vs. not using it. Therefore, in addition to comparing and contrasting, they had to justify their opinion(s). I had originally embedded two short videos to be compared in the computer-based test, but when piloting the task, it became obvious that it was too demanding. The students claimed that it was difficult to recall the information provided in the videos and to then compare and contrast them. Testing short-term memory was considered to result in construct-irrelevant variance, so I opted for pictures instead.

The third task is an instruction task, the main purpose of which is "getting the message across and making sure that it has been understood" (Luoma 2004: 146). In the face-to-face test, test-takers were given a map of Helsinki and they were supposed to instruct their friend, who is visiting from England, on how to get from place A to place B. I pretended to be the friend in this task. In the computer-based version of the test, the examinees were provided with the same map, but this time their friend had sent them a text message claiming to be lost and needing help getting from place to place. The test-takers were meant to leave him/her a voice message with instructions. In the version that was piloted, the computer version of the task required the examinees to give dog care instructions. However, most of the students claimed that this was difficult because they did not own a dog. The task was left out on the grounds that it was unfair towards some of the test-takers due to differences in topical knowledge.

The fourth and last task is a simulation task, in which students pretend to take part in a job interview. According to Luoma (2004: 151) "[r]ole-plays simulate different kinds of communication situations that the target group of the test could plausibly meet outside the test". A job interview seemed like a natural choice, given that the test-takers were at an age where they start looking for their first summer or part-time jobs. The examinees were first asked to choose between three different job advertisements, all of which I designed based on what I thought would interest adolescents. I had additionally planned the questions I would ask during the job interview simulation task so as to make sure that the task was as standardized as

possible. In the face-to-face test, I then acted as the interviewer, whereas in the computer-mediated test, an avatar, Kate, did the interviewing. The avatar was designed through a web-based educational tool, *Voki*, that allows users to create and customize a character, which can then be embedded into Moodle. Unfortunately, the avatar cannot be made interactive.

From the above task overview, one can see that the first three tasks are largely monologic. In the face-to-face test, I tried to intervene as little as possible and mostly resorted to backchannelling. However, I did help the test-takers if they were facing difficulties moving on. The last task in both tests is dialogic because I wanted to capture the interactive nature of speech in as far as it is possible to realize with both formats. The reason I did not want to design the face-to-face test exclusively in the form of an interview is that, in test situations, interviews typically lead to an imbalance in the amount of speech produced by the interviewer and the interviewee (Young and Milanovic 1992; Johnson 2001).

3.1.4 The administration of the two tests

All of the informants completed the tests in the same order. They were first asked to complete the face-to-face test, and then three weeks later, they took the computer-based test. The three weeks were regarded as long enough to avoid test "practice effects", or in other words, the consequence of test-takers' performance improving simply as a result of them gaining experience in taking the same or a similar test (Dörnyei 2007: 53). This could not be entirely avoided, however. The choice of having part of the examinees complete the direct test before the semi-direct test and the other half of the informants taking the semi-direct test before the direct test was considered, as this could, to an extent, have minimized the consequences of the practice effects. However, the choice of collecting the data this way was chosen, so as to prevent the examinees from discussing the test design with one another, and thus affecting their performance the second time they completed the test[23].

Both of the tests were recorded with a recording device. The length of the face-to-face test varied from informant to informant, lasting between 10 and 18

[23] Previous studies (e.g. Malabonga *et al*. 2005; Kiddle and Kormos 2011; Thompson *et al*. 2016) have not found the order in which the test have been completed to have a significant effect on test performance.

minutes. The computer-based tests lasted between 8 and 20 minutes. One potential source of construct-irrelevant variance may arise from test-takers' differences in performance on different days. In other words, the test-taker may perform better or worse depending on the time of day (Pollari 2016: 187) and on his/her physiological state (e.g. tiredness). These types of variables were controlled for to the greatest extent possible, and the tests were hence conducted in the same room, a small, relatively quiet conference room in the school, and at the same time of day.

3.1.5 Designing the post-test questionnaires and the group interview questions

Once the two tests had been completed, informants were asked to fill in an online post-test questionnaire (see Appendix 5). The questionnaires for each group were designed using Google Forms. The questionnaires were administered directly after the computer-based test, and they included both multiple choice and open-ended short-answer questions. The questionnaires were administered in Finnish because this was seen as giving the informants the opportunity to best express themselves.

The aim of the questionnaire was to gain insight into the test-takers' preferences regarding the two tests. The test-takers were first generally asked to state which of the two tests they preferred. The questions that followed were more specific. They were asked to state which of the two tests allowed them to demonstrate their oral language skills better and why. They were also asked whether they felt that the tests tested something other than oral language skills, and if this was the case, what the test tested in their opinion. Finally, the informants were asked to state whether one of the two tests was experienced as being more difficult than the other and whether one of the two made them feel more nervous, and if this was the case, which of the two. The participants were asked to complete the questionnaires before taking part in the interview.

The interviews were scheduled a few days after the completion of the tests, giving the test-takers time to reflect on their performance in the two tests. The interviews were organized as focus group interviews, with the groups being divided according to the course (e.g. ENA1) the informants were taking at the time. The focus group format was deemed appropriate as it "is based on the collective experience of group brainstorming, that is, participants thinking together, inspiring and challenging each

other, and reacting to the emerging issues and points" (Dörnyei 2007: 144). In such an interview, the interviewer acts as a moderator. The interviews were conducted in Finnish.

The aim of the focus group interview was to gain a better understanding of the students' attitudes towards the two tests. The answers given in the interview were therefore meant to complement the answers given in the questionnaires. The participants were additionally asked about their attitudes towards teaching and testing speaking as well as their feelings towards the digitalization of the matriculation examination, more generally. The interview questions were therefore divided into three themes: test-taker preferences towards the two tests in the current study, the testing of oral skills in general and finally, the digitalization of tests and examinations (see Appendix 6). The interviews varied in length, lasting between 15:22 and 37:01 minutes.

3.1.6 Transcribing the data

Once the informants had completed both of the tests, responded to the post-test questionnaire and taken part in the interview, I began transcribing the data from the two tests and the interviews. The data was transcribed using Atlas (version 4.1.0), a tool for the creation of complex annotations on video and audio resources (https://atlasti.com/). It can be used as an aide in the transcription process, as it allows for easy navigating of the audio file.

The transcripts are relatively detailed, as they have to take into account phenomena pertinent to complexity, accuracy and fluency. The detail of the transcripts was also meant to help with the interpretation of unclear utterances in the analysis. One aspect which was particularly important was pronunciation. According to Cameron (2001: 41), "[t]he issue of spelling is especially pertinent where the informants whose speech is to be transcribed are speakers of a nonstandard variety". All lexemes which were pronounced in a nonstandard manner[24] were transcribed phonetically using International Phonetic Alphabet (IPA) conventions. The transcription conventions used for transcribing the informants' oral test performance and interviews are shown in Appendix 7.

[24] The Cambridge online dictionary which provides the IPA transcriptions and standard pronunciation models for both standard US English and RP (UK) was consulted.

3.2 Methods

The methods consist of the tagging of the transcribed data in terms of complexity, accuracy and fluency measures, the comparison of these in the two tests, as well as the processing of the test-takers' responses in the online post-test questionnaire and the focus group interview.

3.2.1 Data tags in terms of complexity, accuracy and fluency

Following the transcription of the test-takers' oral performance, the data was manually tagged in terms of the measures of complexity, accuracy and fluency. The measures used in this thesis are shown in Table 1. The data obtained from the transcripts is predominantly quantitative.

CAF	***How it is measured***[25]
Complexity	- amount of subordination - type-token ratio
Accuracy	- percentage of error-free clauses - errors per 100 words
Fluency	- percentage of silent pauses - hesitation: false starts, repetitions, reformulations and replacements

Table 1 – *CAF measures used for analyzing the data*

Considering that complexity is commonly operationalized as the production of elaborated language, the measures chosen to account for differences in complexity are (i) the amount of subordination and (ii) the type-token ratio. The first measure, subordination, as defined by Biber *et al*. (2002: 223), is "one clause […] embedded as part of another clause". The unit of analysis used for the amount of subordination is the so-called *c-unit*[26] (communication unit), a syntactic unit defined as (a) one simple independent finite clause or (b) an independent finite clause plus one or more dependent finite or non-finite clauses (Foster and Skehan 1999: 229). The transcripts were tagged in terms of finite and non-finite clauses, after which the number of total clauses and c-units was determined. For this particular analysis, I referred to Biber *et al*.'s (2002) taxonomy of clauses. The amount of subordination was then calculated

[25] In my analysis, I disregarded unclear utterances.

[26] The c-unit has been defined in numerous ways in the literature. This particular definition is very similar to the definition of the AS-unit discussed earlier as defined by Foster *et al.* (2000: 365). The only difference is that the c-unit does not take into account ellipted elements (e.g. sub-clausal units).

by dividing the number of clauses by the number of c-units to yield a figure giving some indication of subordination per communication unit. The greater the value (≥1), the more subordination there is, and hence the more complex the language.

Moreover, because Skehan (2009: 514) claims that lexis is a form of complexity which ought not to go unacknowledged if a complete picture of test-takers' L2 speech performance is to be achieved, I decided to include a measure of lexical complexity: the *type-token ratio*. This is a calculation of the total number of different lexemes (i.e. a base word and all of its inflections) used (*types*) divided by the total number of lexemes in the entire stretch of speech (*tokens*). The closer the ratio is to one, the greater the lexical richness. The ratio was obtained by listing the types into an Excel sheet and calculating their frequencies. Only lexemes recognized as standard English words[27] and which were complete were considered in the analysis.

The measures chosen for accuracy determine the extent to which the informants' speech in the two tests deviates from the rule system of the target language. The measures used were (i) the percentage of error-free clauses and (ii) the number of errors per 100 words. These particular measures were chosen because they "serve as general measures of accuracy" (Ellis and Barkhuizen 2005: 151) and have been widely used. The first measure is calculated as the number of error-free clauses divided by the total number of independent clauses and subordinate clauses, multiplied by a hundred. The second measure is calculated as the number of total errors divided by the total number of words produced, divided by a hundred.

Before the calculations could be carried out, errors were identified in the transcripts. The errors were categorized into mistakes relative to phonology, morphology, syntax and lexis. Only such errors were considered that are "indisputably inappropriate" or simply "nonexistent in English" (Skehan and Foster 1997: 195). Errors were again validated by consulting the Cambridge online dictionary and corpora (BNC and COCA). However, self-corrections were not considered, as this would have seemed unfair; the informant did, after all, realize having made a mistake and proceeded to correct it.

[27] What to count as standard English words was determined by cross-referencing unclear cases with dictionary entries (Cambridge online dictionary) and corpora (the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA)).

Finally, fluency, which is defined as speech produced under time pressure without excessive pausing and hesitation, was measured by means of (i) the temporal variable of *percentage of silent pauses* and (ii) by analyzing *hesitation phenomena*. The first of these, the percentage of silent pauses[28], provides a measure of silence during a task. This was calculated as the percentage of silence in a test-taker's performance. The length of pauses (in seconds) was marked at the time of transcription. For practical reasons, only pauses beyond the threshold of 2 seconds were considered. The pauses were then summed up and divided by the total performance time so as to correct the result for length of speaking performance. For the extent of pausing to be comparable between the two test formats, any speech produced by me in the face-to-face test or the avatar in the computer-based test was subtracted. The result was finally multiplied by a hundred to obtain the percentage of total pausing during test performance.

The second fluency measure, hesitation phenomena, includes false starts, repetitions, reformulations and replacements. False starts refer to incomplete utterances, which may occasionally be followed by reformulations, i.e. words, phrases or clauses that are repeated with minor modification. Repetitions are words, phrases or clauses that are repeated without any modification, whereas replacements constitute lexical items that are immediately replaced by some other lexical items (Foster and Skehan 1999: 230). These hesitation indices were marked in the transcripts, counted and finally divided by the total number of words uttered during performance to enable comparability.

3.2.2 The post-test questionnaire and the focus group interview

The post-test questionnaire data is mainly quantitative. All of the respondents' answers for each multiple-choice question were first quantified and then represented visually in the form of pie charts. The short-answer questions, on the other hand, were first translated into English and then categorized in terms of test preference. The focus group interview answers were then used to elaborate on the responses provided in the questionnaire. The interviews were partially transcribed, so that relevant comments could be selected for the discussion of participant preferences,

[28] In this thesis, pause length was not measured in terms of a test-taker's mean duration of pauses. Filled pauses were also left out of the analysis.

teaching and testing oral skills and the digitalization of tests. The interviews were conducted in Finnish, so the relevant extracts had to be translated into English. This part of the analysis is qualitative.

4 Analysis

The analysis section is divided into two main parts. The first section, which is itself divided into three sub-sections, focuses on the differences in terms of complexity, accuracy and fluency in the face-to-face and computer-based tests. Each CAF measure is discussed in turn, starting with complexity, followed by accuracy and finally concluding with fluency. The results will be complemented by extracts from the analysis so as to demonstrate how the analysis was carried out and how the numbers were obtained. The second main section then concentrates on the results of the post-test questionnaire and the focus group interviews and is meant to shed light on test-taker preferences.

4.1 Differences in complexity

Complexity was measured by looking at (i) the *amount of subordination* and (ii) the *type-token ratio*. The first measure is meant to give an indication of grammatical complexity, whereas the second measure attempts to demonstrate differences in lexical richness.

The first table, Table 2, makes comparisons in terms of the amount of subordination between the face-to-face and computer-based test formats. This was determined by dividing the number of clauses by the number of c-units to yield a figure giving some indication of subordination per communication unit. The minimum value of such a figure is 1.00, which would mean that every c-unit is represented by one single clause. In other words, the greater the value, the more subordination there is and the more complex the language. The test on which informants demonstrated more subordination has been indicated in green.

	Amount of subordination	
	Face-to-face	*Computer-based*
Informant 1	1.68	1.54
Informant 2	1.71	1.63
Informant 3	1.51	1.98
Informant 4	1.81	1.88
Informant 5	1.79	1.89
Informant 6	1.80	1.65
Informant 7	1.59	1.58
Informant 8	1.75	1.81
Informant 9	1.77	1.91
Informant 10	1.75	2.01
Informant 11	1.54	1.93
Informant 12	1.65	1.54
Informant 13	1.67	1.97
Informant 14	1.71	1.65
Informant 15	1.59	1.98

Table 2 – *Comparison of subordination in the two tests*

As Table 2 indicates, all of the informants recorded values between the minimum of 1.00 and a figure of 2.01. This means that the average c-unit contains one subordinating clause. Furthermore, nine out of fifteen (60%) informants have a higher clause to c-unit ratio in the computer-based test. This means that the speech of these informants can be considered more complex in terms of subordination in the computer-based test compared to their speech in the face-to-face test. Six out of fifteen (40%) of the informants, on the other hand, performed better in terms of this measure in the face-to-face test. However, as one can see, the differences in these measures are relatively small for all of the participants.

The following extracts demonstrate how the amount of subordination was determined. Due to space restrictions, only parts of the analysis are shown. During the process of the analysis, each task was first dealt with separately, after which the totals for all of the tasks on each test were summed up and the results were obtained. This was mainly done in order to reduce the possibility of error, but also to make the analysis process easier. As stated earlier, the transcription conventions used can be found in Appendix 7. Furthermore, in this particular analysis, square brackets were used to indicate grammatical ellipsis that I proceeded to fill in, in order to make the analysis clearer. The following two extracts are from the performance of informant 1 (S1) on task 1 of the face-to-face and subsequently computer-based tests. Task 1 is the description task in which test-takers were asked to describe the events in a comic strip.

Independent finite clause
Dependent finite clause
Dependent non-finite clause

FACE-TO-FACE

Task 1

S1: uh: we have a man and a (.) boy (.) [who are] driving a car (.) and the man is driving the car (.) and the (.) boy is [sitting] behind (2) they stop and (.) the man is trying to figure out what the problem is (.) by (.) checking out the engine

EP: (2) mhmm

S1: he's also looking under the /tə/ car (3) then (.) the man can't (.) figure out the (.) problem (.) and the boy leaves (4) uh (2) then (3) the boy leaves with a (.) scooter (.) uh then (2) .t the man (2) starts to (.) move the car (.) with his legs

EP: (3) what kind of problem do you think there is with the car

S1: uh the engine is not working

C-units[29]: 11

Clauses: 16

As one can see, the oral performance above is made up of 16 clauses and 11 c-units. Below is the same candidate's oral performance in the computer-based test.

[29] C-unit: (a) one simple independent finite clause or (b) an independent finite clause plus one or more dependent finite or non-finite clauses

COMPUTER-BASED

Task 1

S1: so (.) in the story there are (.) uh (2) a wife and a husband (.) who are eating (.) and their boy is missing (.) then the (2) little bit FAT (.) h: father (2) with /wɪt/ (.) mo- uh moustaches /ˈməstɑːʃɪz/ (.) goes to (.) boy's room and [he] tries to (.) <get the boy to:> (.) eat h: (.) then the boy (.) leaves the room and the (.) dad stays and [he] watches (.) what book is (.) uh: (2) .t the boy reading (.) then the boy is (.) at the (2) table (.) uh [he is] ready to start to eat (.) with the his mom and then (.) uh they're (.) confused because (2) they're thinking where (.) could (.) his dad be (.) then (.) they see that his dad is (.) reading the same book as he was [reading]

C-units: 11

Clauses: 21

The number of c-units in this performance is identical to the one in the face-to-face test. However, the number of total clauses is greater (16 vs. 21). Therefore, there is more subordination in the computer-based version of Task 1. Interestingly enough, this is not in line with informant 1's performance overall. This is due to the fact that he demonstrated a greater amount of subordination in the face-to-face test when all four tasks are taken into consideration. This implies that the task itself can determine the success of performance on a test.

As opposed to the previous complexity measure, the type-token ratio gives an indication of the variety of lexis in the two tests. The higher the type-token ratio is, the more complex and the higher the degree of lexical variation is in terms of test performance. For this analysis, I went through the transcripts by hand and recorded all the individual lemmas (i.e. a base word and all of its inflections), or *types*, in an Excel sheet. Therefore, compound nouns (e.g. *department store*, *railways station*, etc.), phrasal and prepositional verbs (e.g. *figure out*, *take care of*, etc.) as well as certain idioms (e.g. *to kill a bird with one stone*) were calculated as one single type. The same is true for singular and plural nouns as well as verbs with their different inflectional forms. These forms were determined with the help of the Cambridge online dictionary and the two major standard English corpora: the BNC and COCA. Non-existent forms, or in other words, lexemes which could not be found, were left out of the analysis. After these had been listed in the Excel file, the frequencies (i.e. *tokens*) were recorded and finally, the ratio was determined by dividing the types by

the tokens, the result of which was multiplied by a hundred in order to get a percentage.

Table 3 indicates differences with regard to the type-token ratio in the two tests. The test on which informants demonstrated a higher type-token ratio has been indicated in green.

	Type-token ratio	
	Face-to-face	*Computer-based*
Informant 1	35.0%	46.1%
Informant 2	29.5%	32.1%
Informant 3	36.4%	34.5%
Informant 4	30.5%	30.1%
Informant 5	33.7%	34.9%
Informant 6	30.1%	32.3%
Informant 7	27.3%	33.8%
Informant 8	22.8%	19.6%
Informant 9	28.9%	27.9%
Informant 10	23.9%	24.0%
Informant 11	33.7%	33.6%
Informant 12	33.1%	25.9%
Informant 13	34.5%	36.7%
Informant 14	39.7%	39.9%
Informant 15	37.8%	32.5%

Table 3 – *Comparison of the type-token ratio in the two tests*

As can be seen, the informants' performance in the two tests in terms of their type-token ratio is rather equal. Eight of the informants (~53%) have a higher type-token ratio in the computer-based test and seven (~47%) in the face-to-face test. Again, as was the case with the c-unit analysis, the differences between the

informants' performances are relatively small, with most differences being just one or two percentage points. In fact, the only participants with greater differences are informants 1 (35% vs. 46.1%) and 12 (33.1% vs. 25.9%).

4.2 Differences in accuracy

This second sub-section focuses on accuracy. As discussed earlier, the measures chosen to determine the informants' degree of accuracy in the two tests are (i) *the percentage of error-free clauses* and (ii) *the number of errors per 100 words*. The results for these two measures are shown in Table 4 below. The values which indicate a higher percentage of error-free clauses or, inversely, a greater number of errors per 100 words are indicated in green.

	Percentage of error-free clauses		**Errors per 100 words**	
	Face-to-face	*Computer-based*	*Face-to-face*	*Computer-based*
Informant 1	73.8%	61.7%	0.00041	0.00077
Informant 2	75.5%	69.6%	0.00052	0.00067
Informant 3	63.1%	73.6%	0.00062	0.00049
Informant 4*	72.1%	69.8%	0.00073	0.00068
Informant 5	90.6%	86.5%	0.00011	0.00026
Informant 6	85.4%	84.3%	0.00027	0.00037
Informant 7*	82.1%	75.6%	0.00038	0.00032
Informant 8*	68.7%	60.2%	0.00088	0.00082
Informant 9*	80.3%	79.7%	0.00041	0.00035
Informant 10	86.0%	83.2%	0.00027	0.00032
Informant 11	64.7%	71.6%	0.00073	0.00067
Informant 12	54.7%	41.4%	0.00113	0.00156
Informant 13	38.3%	22.9%	0.00194	0.00273
Informant 14	42.7%	43.9%	0.00174	0.00172

Informant 15*	67.6%	69.9%	0.00076	0.00085

Table 4 – *Comparison of accuracy measures in the two tests*

From the percentage of error-free clauses one can see that 11 (~73%) informants have a higher percentage of error-free clauses in the face-to-face test. There were fewer phonological, morphological, syntactical and lexical errors in the direct test, which would suggest that the informants strive to be more accurate when speaking to a human interlocutor. Four of the informants (~27%), on the other hand, made fewer mistakes in the computer-based test.

However, when comparing these results with some of those for the accuracy measure of errors per 100 words, the results appear somewhat contradictory. If one looks at the results of the second measure of accuracy, one can see that eight participants (~53%) made more mistakes within a stretch of 100 words in the computer-based test, while a relatively equal number, seven informants (~47%), did so in the face-to-face test. However, as is the case with the complexity measures, most of the differences between the two test performances are relatively small. The number of errors per 100 words was mainly under the value of 0.00100. Only three of the test-takers, namely informants 12, 13 and 14 had a value greater than 0.00100.

Due to the differences in results provided by the two accuracy measures, the results are somewhat inconclusive. Informants 4, 7, 8 and 9 (marked with an asterisk) appear to have a higher percentage of error-free clauses in the face-to-face test, on the one hand, but on the other hand, they also have more errors occurring over the stretch of 100 words. Contrarily, informant 15 has a higher accuracy percentage in the computer-based test, but according to the other accuracy measure, he makes more errors within 100 words. One of the reasons behind this discrepancy could be the nature of the measures used. In other words, the measure which takes into account error-free clauses, may be more unreliable, as the analysis only takes into account fully-formed clauses, disregarding sub-clausal units, for example. The distribution of errors is another possible reason. In other words, it is possible that test-takers are able to produce a longer stretch of error-free speech, leading to a greater number of error-free clauses, after which they proceed to make several mistakes in a row. The first measure would inherently not be able to pick up on the degree to which the

errors are skewed. However, more qualitative analysis would be required to confirm this second hypothesis.

There were a number of different error types. Common grammatical mistakes were, for example, the misuse of articles, the wrong use of pronouns, the use of the wrong verb tense or lack of subject-verb agreement and mistakes in terms of word order. Some examples from the data are provided below. Examples of such errors have been marked in red. Green is used to indicate self-corrections, which were not counted as errors.

S5 (face-to-face): (5) we can (.) go there (.) by (2) ***the*** subway

S9 (face-to-face): I think ***I'm*** (.) ***pretty social person*** I'm (.) athletic (.) as well I like I like sports

S13 (face-to-face): … there is (2) one older man and
EP: mhmm
S13: ***kid*** (.) behind the car (2) then (.) ***car*** is stopped and (.) there is some wrong with /wɪt/ (.) ***engine***

S9 (computer-based): … the mom (.) has uhm (2) put ***his*** hair nicely and ***he's*** uh ***she's*** got a dress

S2 (face-to-face): (2) and when *the man* (.) c- (2) ***manage*** to fix (.) the engine /ɛnˈʤiːn/

S8 (computer-based): … it's a huge (.) like grocer- or department store (.) and: I think there ***is*** some really good *souvenirs* /ˌsuvəˈnɪːs/

S2 (face-to-face): uh and it- I think ***it's built*** /bɪldt/ (.) <during the /tə/ (.) nineteenth (.) century /ˈsɛntʊri/>

S8 (computer-based): … he <***founds***> his dad reading the same book /buːk/ >that he was reading

S12 (computer-based): hey (.) I saw your text mi- message and .h: (.) and tried to call but you didn't answer so I (.) ***left*** this voice message for YOU

S14 (computer-based): hey (.) you didn't pick up your phone (.) I hope (.) you didn't ***lost*** it (.) too

S6 (face-to-face): mmm (3) well uhm (.) <what kind of> (.) is it a (.) just basic summer camp or is it a (.) like special: some- do do you have like some kind of specialty as in: (.) ***it's like a*** (.) ***maybe football camp or something***

S7 (computer-based): … the (.) boy (.) goes to check (.) where >his dad is and then< (2) the (.) dad (.) is (.) reading the (.) book that ***the boy was earlier*** (.) ***reading***

In terms of pronunciation, the test-takers had the most problems with the (inter)dental fricatives (/θ/ and /ð/), and in some cases, the distinction between voiced and unvoiced plosives. The informants also had problems with the postalveolar fricative /ʃ/ in the sense that they tended to overcompensate its use and place it where it did not belong. The problems wit these particular phonemes did not come as a surprise, as they tend to cause the most problems for Finns (Sajavaara and Dufva 2001: 249–250). Examples are given below and the relevant errors have been highlighted in red.

S13 (face-to-face): (3) and I don't ***think /tɪŋk/*** I don't take (.) ***things /tɪŋs/*** (.) too serious

S11 (computer-based): (2) YES I have previously work on: grocery ***stores /ʃtɔrs/***

S8 (face-to-face): and: (5) .t maybe (.) <I'm kinda shy> but I also like people and I like to be with people (2) and: (9) and: (.) ***probably /ˈprɒpəpli/*** that I'm: (6) like (.) happy and like

The test-takers occasionally also made mistakes with regard to word choice. In some cases, this occurred when the test-takers had problems with remembering the correct word. Below are some further examples from the data. Again, relevant errors are marked in red and self-corrections in green.

S1 (face-to-face): and uh (.) you can get there by (3) taking a (2) {knocking on table} #raitiovaunu (.) °mikäs hitto se nyt sit on°# (.) a ***cable car*** #vaikka#

S2 (computer-based): you can buy ice cream there and (.) uh (.) sit on the bench and (.) ***only*** watch people

S12 (face-to-face): and then /dɛn/ (.) something goes wrong and the (.) <motor crashes /ˈkræsɪs/ and (.) and h-> (.) dad needs to ***prepare*** /priˈpɛr/ it
EP: mhmm
S12: (2) but (.) dad: doesn't know what's wrong so he <*checked*> *checks* under the /tə/ (.) car

S3 (computer-based): … <there*'s* famous finnish *paintings*> (.) uh and (.) it's (.) very (.) basic like uhm (2) tourist uhm (2) ***sightsee*** (.) uh and (.) from ateneum (.) we can (2) uh (.) >go to #tuomiokirkko# by walking< (.) we should pass (2) uh (.) >ateneuminkuja< (.) then /tɛn/ then walk straight to yliopistonkatu

4.3 Differences in fluency

The third and final feature which was considered in the comparison of the two test formats was fluency. The first of these measures (i) the *percentage of silent pauses* in the two tests, aims to establish the extent of breakdown fluency, whereas the second measure, (ii) *hesitation phenomena,* consisting of gives an indication of the informants' repair fluency. In the first part of the analysis, pauses are meant to provide an indication of the extent to which learners disengage from speaking in order to plan their spoken message. The measures of repair fluency, on the other hand, give an indication of the extent to which speakers adjust their message in the event that they recognize having made a mistake.

Table 5 below indicates the differences in the percentage of silent pauses in the two tests. Again, the test performance with a greater percentage of silent pauses is marked in green.

	Percentage of silent pauses	
	Face-to-face	*Computer-based*
Informant 1	42.9%	43.2%
Informant 2	46.8%	44.9%
Informant 3	18.2%	24%
Informant 4	24.4%	29.4%
Informant 5	32.6%	35.3%
Informant 6	15%	19.2%
Informant 7	18.6%	26.8%
Informant 8	24.6%	33.3%
Informant 9	15%	19.6%
Informant 10	10.9%	22.7%
Informant 11	34.9%	46.8%
Informant 12	16.2%	16.3%

Informant 13	34.4%	51.6%
Informant 14	35.4%	38.6%
Informant 15	33.5%	29.8%

Table 5 – *Comparison of the percentage of silent pauses in the two tests*

With the exception of two of the informants, namely informants 2 and 15, all of the participants paused more in the computer-based test. Overall, the differences in terms of silent pauses in the two tests are again relatively small. In the face-to-face tests, the percentage of silent pauses ranges from 10.9% to 46.8%. In the computer-based test, on the other hand, the range is 16.3% to 51.6%. There are some individual differences, however. While informant 12 has a difference of 0.1 percentage points, informant 13 has a difference of 17.2 percentage points.

The overall differences in silent pauses can in part be explained by the lack of a human interlocutor filling in pauses and reacting to the informants' speech in the computer-based test. The interactiveness of the face-to-face test should therefore be kept in mind. A second and very probable reason may simply be that the informants stop to search for words or, if they do not come up with the right expression, a way of paraphrasing what they want to say. A further factor that may contribute to the differences in breakdown fluency is the fact that the informants had to scroll up and down on the computer in order to see the whole task (i.e. instructions, prompts, etc.), whereas in the face-to-face test, the tasks were given to the test-taker one by one and everything could be seen on the paper at once. The act of scrolling may have the effect of adversely interfering with the test-taker's train of thought.

The second measure of fluency aims to shed light on the differences in hesitation phenomena, namely false starts, repetitions, reformulations and replacements. To see the distributions and frequencies of the individual hesitation markers in the two tests, see Appendix 8[30]. The total number of hesitation markers was standardized to the number of words uttered by the test-taker in each test. These figures are shown in Table 6. Green is again used to indicate the test performance which exhibited a greater use of hesitation markers.

[30] Table 7 shows the distributions and frequencies of the hesitation markers in the face-to-face test and Table 8 shows them in the computer-based test.

	Hesitation	
	Face-to-face (Total relative to total number of words uttered)	*Computer-based* (Total relative to total number of words uttered)
Informant 1	0.037702	0.047619
Informant 2	0.067633	0.040615
Informant 3	0.070362	0.089253
Informant 4	0.035294	0.049716
Informant 5	0.032967	0.035541
Informant 6	0.108796	0.077519
Informant 7	0.054726	0.045283
Informant 8	0.057461	0.087625
Informant 9	0.051223	0.037809
Informant 10	0.045894	0.042810
Informant 11	0.043728	0.053192
Informant 12	0.124451	0.104686
Informant 13	0.101617	0.083117
Informant 14	0.062837	0.057234
Informant 15	0.049587	0.098336

Table 6 – *Hesitation phenomena in the two tests*

As one can see by looking at Table 6, the difference between the two tests in terms of the number of hesitation markers is again minimal. Eight out of fifteen (~53%) of the informants exhibited more hesitation in the face-to-face test, whereas seven out of fifteen (~47%) hesitated more in their performance in the computer-based test. False starts were, by far, the most common type of hesitation marker in the two tests. Very often, the informants' false starts were followed by some form of reformulation, as they attempted to rephrase what they were trying to say. Repetitions and replacements occurred more rarely.

The following extracts are again meant to give an indication of how the analysis was carried out and how the different hesitation markers were categorized.

False starts
Reformulations
Repetitions
Replacements

S2 (face-to-face): and (.) there are also (3) uh some (3) they are not ships I d- I c- I can't remember the word (.) >but you uh< .t you can (.) go to: (.) korkeasaari (.) the zoo
EP: ok
S2: or: (.) suomenlinna uh which is the old: (2) I think they kept uhm (8) I oh: (.) sorry
EP: it's ok so we go there by boat
S2: yeah (.) the- (4) I can't remember the word #vanki# @ mmm .t (3) crimi- [uh]
EP: [mhmm]
S2: old criminals there

S12 (computer-based): … #noin# (5) >#ai niin pitää vielä sanoa hyviä ja huonoja puolia#< .h: well I think in the picture A (.) uh if you (.) are (.) if you are good /kʊt/ at learning by YOURSELF and (.) you- (.) don't probably like (2) <that /tæt/ (.) the /tə/> (.) when you are (.) with the partner learning maybe you don't learn that /dæt/ well so you have to (.) be yourself /ˈjɔrʃɛlf/ and (.) >think /tɪŋk/ about it yourself so maybe if you are like< (.) THAT (.) kind of a person that (.) that /tæt/ suits /suɪts/ you better (2) but if you are (2) if you are: (.) the kind of person that /tæt/ (.) learns m- much better from (.) a group /krup/ standpoint then /tɛn/ you (.) then the picture B suits /suɪts/ you better and of course if you are (.) better /ˈbɛθər/ with (.) computer: with learning (.) with computers than with (.) book and pen (.) then /tɛn/ the /tə/ picture B suits /suɪts/ better /ˈbɛθər/ as well

S14 (face-to-face): and when (.) suddenly the scar (.) car stop (2) dads goes to look at (.) *enginer* (2) and realize (.) it's overheated
EP: mhmm
S14: (3) uh (3) her son (.) uh isn't a patient /ˈpeɪsən/ one so .t (2) he go backs (.) he: goes back to home (.) and: (2) uh (3) (>starting to wo-<) start going her own way (.) in the school

Foster and Skehan (1999: 230) point out that hesitation phenomena are common in speech because they reflect moment-by-moment decisions that a speaker makes while speaking. The speaker wants his/her message to be clear, so s/he adjusts and improves his/her message within the constraints of real-time communication. In other words, both native and non-native speakers hesitate. However, there are individual differences and I too was able to pick up on this while conducting my analysis. Some of the informants hesitated and adjusted their message a lot more than others, irrespective of testing mode. In fact, Krashen, for example, notes that some speakers appear to monitor their linguistic output whenever possible, while others seem to not do so at all (Mitchell *et al*. 2013: 43).

4.4 Post-test questionnaire and focus group interview responses

This final part of the analysis focuses on interpreting the informants' responses to the post-test questionnaire administered electronically to each participant after completing the second test (i.e. the computer-based test) and the focus group interviews. Each of the questions from the questionnaire is first analyzed separately and complemented by extracts from the interviews. The responses from all four groups are shown in the form of pie charts. At the end of the section, there is an overview and summary of the findings as well as a short analysis of the two other themes discussed in the interviews: whether the informants consider testing oral skills important and what their attitudes are towards the digitalization of tests overall.

The first question which aimed to uncover the test-takers' overall preference regarding the two test formats is presented in Figure 2 below.

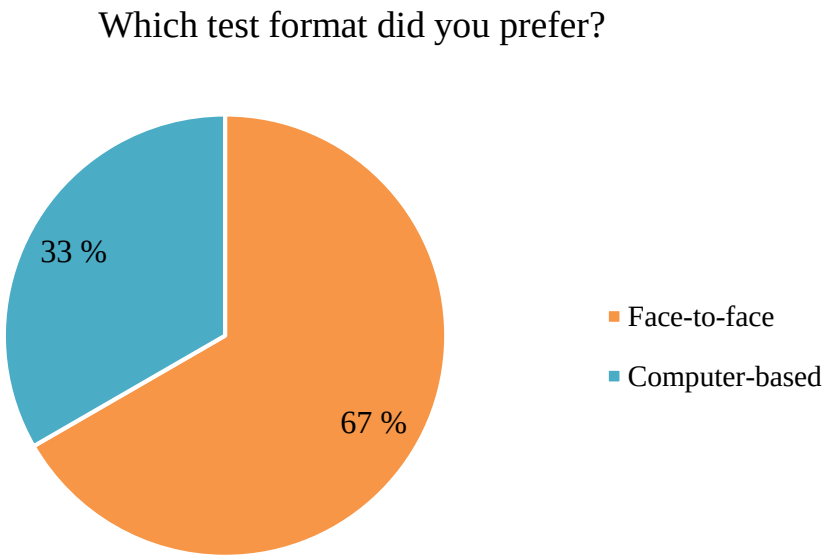


Figure 2 – *Question 1 from the questionnaire*

The overall majority of the test-takers indicated a preference for the face-to-face test. In fact, two thirds (*n*=10) claimed they preferred it to the computer-based test, while one third of the test-takers (*n*=5) showed a clear preference for the computer-based test.

Interestingly enough, when asked in which test they felt they were better able to demonstrate their oral skills, the number of informants indicating a preference for the face-to-face test increased. This is shown in Figure 3 below.

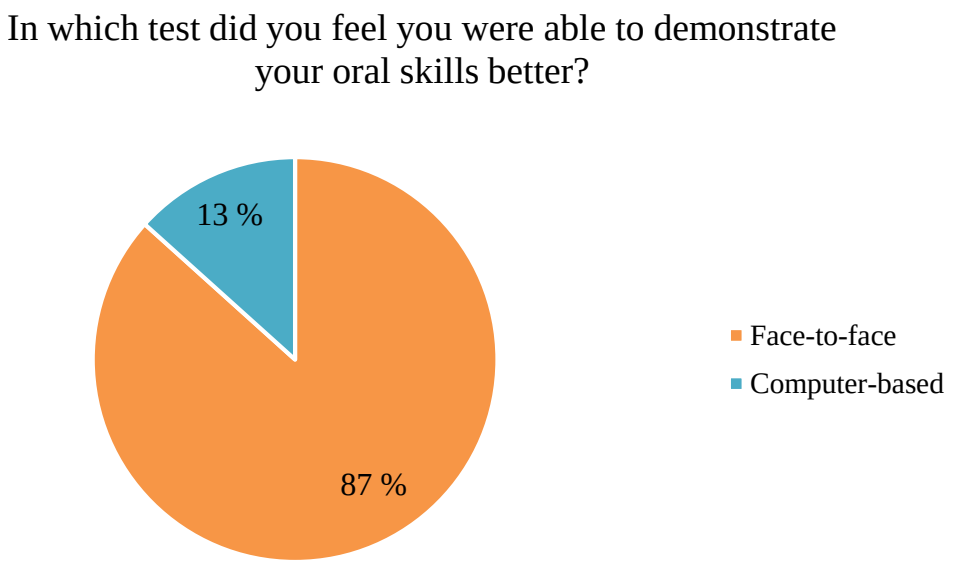


Figure 3 – *Question 2 from the questionnaire*

Only 13% of respondents (*n*=2) claimed they found the computer-based test more reliable. The reasons they gave were the following:

1) Informant 6 (Q[31]): It somehow felt easier speaking to a machine than a human being because there was less pressure.

2) Informant 10 (Q): I wasn't as nervous about the other person's reactions or about understanding what was being said.

In both cases, the informants felt less nervous or pressured when a human interlocutor was not present. In the interview, informant 2, despite indicating a clear preference towards the face-to-face test, points out that the computer-based test can be less face-threatening because you do not necessarily have to think about what you sound like or whether you pronounce every single word correctly. In the focus group interview, informant 10 added that she felt that her way of speaking was different in the two tests:

[31] The answers from the questionnaires are indicated by the letter Q next to the informant's number, whereas test-takers' comments from the interviews are indicated by the letter I.

3) Informant 10 (I): when you speak to another person you want to make your speech as understandable as possible […] but on the computer it's sort of like I just have to get this over and done with […] because it's not like the computer is going to like nod and follow what I'm saying.

She also felt that she did not have to focus as much on body language in the computer-based test.

However, the majority of respondents, 87% (*n*=13), were of the opinion that the face-to-face test offered them more and better opportunities to prove their ability to speak English. In the questionnaire and interview, they offered a number of reasons for this. Most of the responses stress the *naturalness* and *ease* of talking to another human being:

4) Informant 1 (Q): It feels more natural to speak to a human being than a machine.

5) Informant 3 (Q): In my opinion, it was easier to communicate and come up with things to say in the face-to-face test. As opposed to the face-to-face test, in the computer-based test my answers did not come as naturally.

6) Informant 4 (Q): It's easier to speak with a human being.

7) Informant 7 (Q): It's a lot more natural speaking face-to-face with another person than to a machine.

In the interview, informant 11 added that some of the tasks required an interlocutor by nature, because they were interactive:

8) Informant 11 (I): in the other tasks it's like more natural to speak to another human being, for example the map task […] that it like because the other person reacts to what you say but that you just babble on to the computer it just feels unnatural […] it might feel like I'm just talking to thin air

A further reason for the preference of the face-to-face test was that it was more interactive. Both the test-taker and the tester could pose questions and, as the second to last example below shows, this could be helpful in prompting to talk about something that one might not have thought about on one's own.

9) Informant 2 (Q): In the face-to-face test you could ask questions if you did not completely understand what you were being asked. In the computer-based test, this was not a possibility and therefore, if you did not understand something, the whole task could go wrong.

10) Informant 12 (Q): Because in the face-to-face test the other person could react to what you said, the answers were more indicative of my abilities.

11) Informant 14 (Q): In the face-to-face test, the tester could pose additional questions with the help of which I could talk about things that would not have necessarily come to mind.

12) Informant 15 (Q): In the face-to-face test there was a conversational feeling, which gives the test-taker (me) a different picture of the whole test. Although the task types in the computer-based test were similar, speaking to the screen felt, honestly speaking, slightly stupid.

Perhaps because the tester could offer prompts and help the examinee along if s/he got stuck, some of the informants claim that they were better able to recall vocabulary and they felt their speech was more fluent:

13) Informant 5 (Q): I could remember more words [in the face-to-face test].

14) Informant 8 (Q): I could recall vocabulary better in the face-to-face test, and in the computer-based test it felt as though you were talking in vain.

15) Informant 9 (Q): I felt like I could speak more fluently and easily in the face-to-face test. In the computer-based test my speech was more rigid.

16) Informant 13 (Q): More fluent speech

In the interview, informant 13 added that these types of tests were so new to him that the unfamiliarity of the tests could also have affected his performance. He also stated that:

17) Informant 13 (I): it might also have depended on the day but in theory it was more difficult to speak to the computer or (.) I don't know but I experienced more blackouts on that test

The next question in the questionnaire dealt with the question of whether the tests were comparable in terms of validity, i.e. whether they both exclusively tested the test-takers' ability to speak. The responses are shown below in Figure 4:

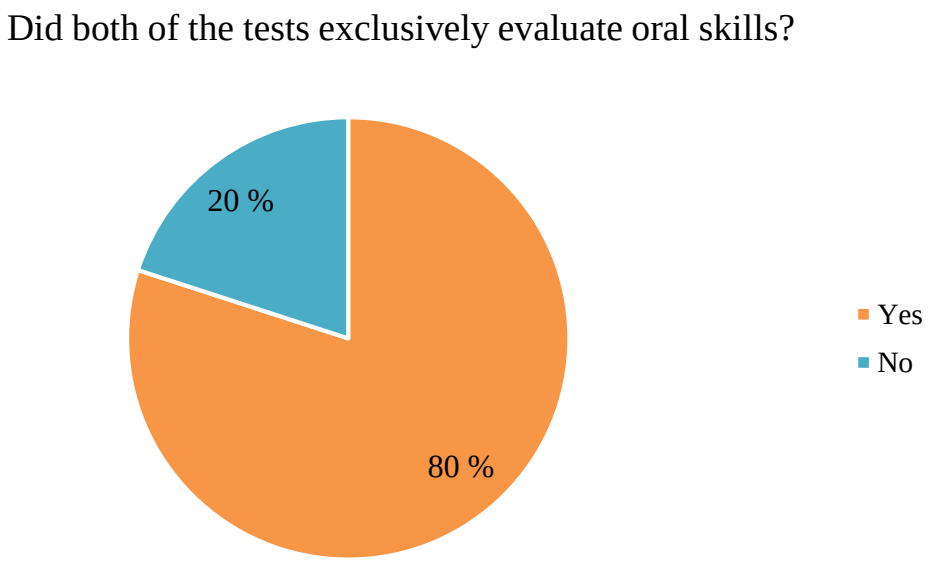


Figure 4 – *Question 3 from the questionnaire*

The majority of the informants, 80% (*n*=12) felt that both tests were equally valid, suggesting a high degree of comparability. However, the fact that 20% (*n*=3) disagreed, claiming that one of the tests or both tested something other than the ability to speak English, should not go overlooked. They were asked to elaborate on the skills they felt the tests additionally evaluated:

18) Informant 2 (Q): The ability to interact in the face-to-face test, e.g. how you react in a situation, whether you look the other person in the eyes, etc.

19) Informant 5 (Q): IT skills, the ability to deal with pressure and knowing vocabulary and syntax

20) Informant 10 (Q): On the map task you need to be familiar with culture and you have to know the cardinal points. I have poor orientation skills and I am not able to give but very simple directions. Otherwise there was no difference between the two tests.

In the first example, the informant felt that the face-to-face test tested interaction, implying that the computer-based test did not — or at least not to the same extent. This is interesting considering that the assessment of speaking should, in my opinion, test one's ability to interact as well. Speech is, after all, mostly interactive. The second example suggests that the computer-based test tested IT skills. In the interview, the informant elaborated that this was due to the fact that she accidentally shut the browser and was unable to return to the test without help. This is of course

an important aspect to consider in computer-based tests. Finally, the third comment raises an important issue in test design: The prompts have to be fair. Informant 10 clearly felt disadvantaged by the task, which required the test-taker to give the interlocutor directions with the help of a map.

During the interviews, the question of the clarity of instructions was brought up. When designing the tests, and in particular the computer-based one, I had to consider how I would make the instructions as clear as possible. From the participants' comments it became clear that they had, at times, struggled to figure out what was expected of them:

21) Informant 5 (I): they [the instructions] were really long or there were like many things (.) that I had to remember to talk about

22) Informant 8 (I): it took me a long time in the map task to figure out that (2) because there was the text message and then there were the instructions .h: I didn't realize I had to scroll down so much (.) I wondered for a long time (.) or at least it felt like a long time (.) where the map is

23) Informant 13 (I): well you kind of forget what was asked once you start speaking

24) Informant 14 (I): you can't see the whole thing like in the first one [face-to-face test] where you had the whole cartoon before you […] I started and then was like oh there are three more squares

The comments above also show that the computer has its limitations as a medium because the instructions and prompts could not all be made visible simultaneously. Furthermore, although there were differences, in the interviews one of the participants claimed to be surprised by the similarity of the two tests:

25) Informant 15 (I): I expected the computer-based test to be .h: radically different from the other one

In theory, the computer-based test could be made very different from the face-to-face test, as test designers could easily take advantage of the numerous functionalities available on computers. However, for the purpose of this study, this was not relevant, as it would have distorted the findings.

The next question in the questionnaire sought to find out, whether one of the two tests was experienced as being more difficult than the other. If the informants

answered “yes”, they were asked to state which one of the two tests they felt was more difficult. The responses to these two questions are shown in Figure 5 below:

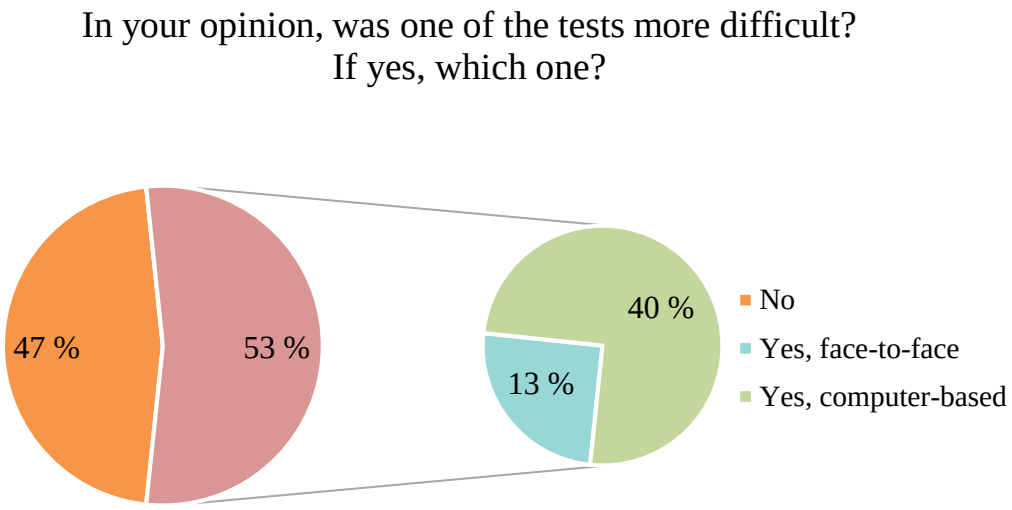


Figure 5 – *Question 4 from the questionnaire*

Figure 5 shows that just under half of the respondents, 47% (*n*=7), felt that there was no difference in the difficulty of the two tests. However, 53% (*n*=8) claim that one of the two tests was in fact more challenging. 13% were of the opinion that the face-to-face test was more difficult, while 40% stated they felt the computer-based test posed more of a challenge.

In the interview, one of the participants stated that some of the tasks might have contributed to one test format being more difficult than the other. In fact, the avatar seemed to cause unease:

26) Informant 6 (I): well for me I think it was maybe easier to speak [to the computer] when having to give directions and such but then when there was the interview I found it a little strange (.) because of the avatar

27) Informant 15 (I): only the last task in which you spoke to the avatar felt a little strange but maybe that’s just because it’s difficult to get used to something new

The responses to the fifth and final question are similar to those in the fourth. These are shown in Figure 6. The test-takers were asked to state whether one of the two test formats made them feel more nervous. Again 47% of the informants (*n*=7) felt that

there was no difference, whereas 53% (*n*=8) state that there was. 20% felt the that the face-to-face test was more nerve-wracking, whereas 33% claim it was the computer-based test that caused them to feel less at ease.

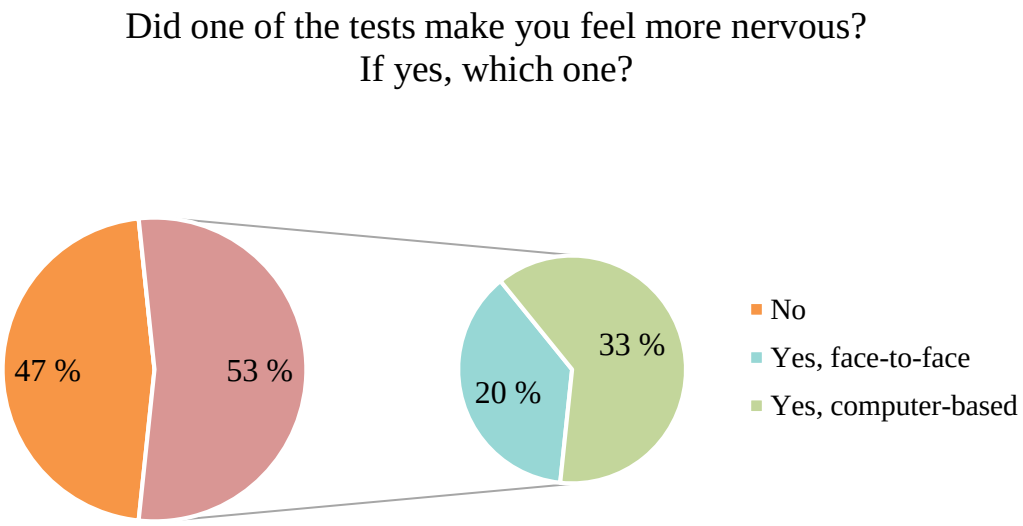


Figure 6 – *Question 5 from the questionnaire*

In the interviews, some of the participants commented that they felt more nervous when a real interlocutor was present because they felt that I was waiting for them to answer and complete the task, whereas the avatar did not comment on what they said, and they could therefore control the tempo of the test situation:

28) Informant 10 (I): its [the avatar's] presence like made me feel less nervous than when a real person was present […] because it won't react like (.) to what I say

29) Informant 11 (I): s/he [real person] is like constantly present so in that sense the artificial intelligence didn't like (.) no matter what you explained it automatically moved on to the next question, so it couldn't ask like a small follow-up question regarding what you had said

Some informants, on the other hand, felt more comfortable when speaking to another person:

30) Informant 13 (I): the fact that another person is present might be precisely what reduces my anxiety

It would seem, as Thompson *et al*. (2016) also pointed out, that personal characteristics and preferred interpersonal style have a great impact on the preference of testing mode. This has important implications for testing because, as Hildén (2000) discovered, outgoing and extroverted test-takers have a tendency to score higher on oral tests than shyer speakers. Therefore, test-takers who feel uneasy when speaking to another person may feel more comfortable when they are given the chance to complete an oral proficiency examination on a computer.

To summarize the findings from the questionnaire, it would seem that the majority of the respondents prefer the face-to-face test because they felt they were better able to demonstrate their oral skills in the test with a human interlocutor. This was mostly due to the fact that it felt more natural to speak to another human being. Although around half of the informants felt that there was no difference regarding the difficulty of the two tests, or a sense of nervousness arising from completing either one, of those who did indeed find there to be a difference in these aspects felt that the computer-based test was the one that made them feel less at ease.

The interviews further covered the topics of oral language teaching and testing as well as the digitalization of the matriculation examinations. The participants' arguments and opinions concerning these two topics are discussed in the concluding paragraphs of the analysis section.

When asked about their attitudes towards teaching and testing speaking, all of the participants were of the opinion that learning to *speak* a foreign language is the most important — if not the ultimate goal.

31) Informant 1 (I):iIt is an important part [of language testing] (.) if you can't speak it (.) the language is pretty much useless

32) Informant 2 (I): such a large part of communication occurs orally (.) it's not like you can discuss everything on the internet

33) Informant 10 (I): speech is a feature that people have in common

34) Informant 11 (I): well I would say that it is recommendable that you are able to speak at least a little […] it would be a little stupid if I can read and write but I can't produce any speech

From the comments above, it becomes clear that oral skills which, as pointed out by informant 10, are a human phenomenon, are vital for communication and should

therefore be strongly present in a foreign language syllabus. Furthermore, Sawaki (2012: 429–430) states that "[i]n today's highly computerized society, many language use activities take place in computer-mediated environments". Although this is true, informant 2 reminds us that not all communicative exchanges occur on the internet. Face-to-face interaction is still a very important feature of communication.

Furthermore, the vast majority of the participants agreed that the testing of oral skills should be compulsory in the matriculation examination.

35) Informant 8 (I): well if it's [speaking] emphasized in teaching then it should be tested

36) Informant 11 (I): it would be quite good [to test speaking] (3) because the matriculation examination brings your studies to an end, so it would be a valuable addition

37) Informant 14 (I): yeah (.) I think […] otherwise those who have like been on exchange or come from a bilingual family would be the only ones who take the test.

These comments are in line with Pollari's (2016) findings, who states that assessing speaking should somehow be incorporated into the matriculation examination. Informant 14 also makes an important point because she implies that it would be most fair if everyone were required to do the test so as not to cause anyone to be in a disadvantaged position.

Moreover, the participants unanimously agreed that there have not been any major changes in the number or types of oral tasks over the years. In fact, the informants feel there could be more variation. During the focus group interviews, all four focus groups mentioned that the most common task type is the A/B dialogue in which one ('A') works together with a partner ('B'). 'A' then translates his/her lines into English, while 'B' — who has the sentences in English — observes and corrects 'A' if necessary. The other type of tasks that came up in the interview are question and answer tasks and pronunciation exercises in which one identifies the correct phoneme (e.g. /p/ or /b/) from a tape containing minimal pairs. As some of the informants pointed out, however, the aim is often not that of learning to speak:

38) Informant 4 (I): these types of tasks are done because we're meant to learn vocabulary or grammar […] there aren't that many where you just learn to talk about something

39) Informant 14 (I): […] and then it usually just focuses on some grammar topic and the sentences aren't something you would normally use (2) we could have more task where we practice natural conversation

40) Informant 15 (I): the idea of that task [A/B] isn't to learn to speak (.) the main thing is that you understand and are able to translate

Some of the participants highlighted that they would like to practice more spontaneous speech by improvising, debating or giving speeches:

41) Informant 9 (I): it [giving a presentation] was probably the most effective oral skills test I have ever taken […] I personally enjoyed it

The final topic in the interview had to do with the digitalization of the matriculation examination. The participants were generally of the opinion that some subjects are easier and more suitable to be digitalized than others.

42) Informant 2 (I): well at least like essay question responses are in my opinion nice to write on a computer but then like math (.) chemistry where you have to write formulas it's tricky

However, once the discussion shifted back to oral skills testing, the participants brought up a number of concerns they had, one of which was the speech recognition instrument's ability to identify accents and speakers' individual differences in speaking styles.

43) Informant 2 (I): the fact that English can be spoken with so many different accents […] what does the computer consider correct

44) Informant 10 (I): it depends on what type of accent has been introduced to the computer and on whether it is specifically (.) British English that we are then meant to speak […] people have such different ways of speaking too (2) how will that be taken into account

45) Informant 13 (I): nervousness can be detected in speech […] will it affect my score

Informant 13 also brought up the question of how being nervous will affect his score if a computerized instrument were to assess his speech. Although the participants of this study raised a number of important questions regarding the fairness of computerized tests, particularly in the case of speaking, the bottom line seems to be

that the participants find it acceptable to organize an oral exam digitally, given that these aforementioned issues have thoroughly been considered and subsequently addressed in the design and development of the test.

5 Discussion

In this section, the main findings of the study and their implications are discussed. I start by summarizing the findings regarding complexity, accuracy and fluency in the two test situations, after which I consider the differences in the informants' preferences relative to the testing mode. I conclude this section by discussing the implications and limitations of the study.

5.1 Differences in complexity, accuracy and fluency

The main purpose of this study was to find out whether there are differences in the complexity, accuracy and fluency of Finnish upper secondary school students' speech when comparing their performance on two different test setups, namely a face-to-face test and a computer-based equivalent. The results indicate that there are indeed differences, but that these differences are relatively small in all cases. This would suggest that the testing mode may not have a significant effect on the informants' oral performance, and that the tests are, in fact, interchangeable.

The first measure considered the grammatical and lexical complexity of the informants' oral performance. The results indicate that, in terms of grammatical complexity, which considered the amount of subordination demonstrated in the two tests, the informants' speech was generally more complex in the computer-based test: 60% of the informants achieved a higher clause to c-unit ratio in the computer-based test. The second measure, the type-token ratio, gives an indication of the lexical complexity of the test-takers' oral performance. Slightly over half (~53%) of the participants demonstrated a higher type-token ratio in the computer-based test.

Based on these results, one could conclude that the test-takers prioritize complexity when taking a computer-based oral test. However, the increased complexity in this mode of testing could partially also be due to the fact that the computer-based test, which was very similar in structure and content to the face-to-face test, was completed after the face-to-face one. In other words, practice effects may have had

an impact on the result. In fact, while going through the transcripts during the course of the analysis, there were instances in which the informants recycled their ideas from the first testing event. Slight differences in the tasks may also have contributed to this difference. In other words, taking the first task of both tests as an example, it could be that the increased number of elements, in this case characters, in the comic strip prompt in the computer-based test could have led to higher complexity measures. This is because, as Ellis (2003: 120) points out, tasks which require manipulation of features (e.g. in terms of the number of elements in a task), may potentially lead to higher complexity. However, the possibility of this influence cannot be confirmed by the present study.

The second measure, accuracy, was measured by first calculating the percentage of error-free clauses and by then counting the errors per 100 words. The results of this part of the analysis proved to be somewhat contradictory due to the fact that the first measure made it seem like an overwhelming majority (73%) of informants had a higher percentage of error-free clauses in the face-to-face test, whereas the second measure indicated that, in fact, nearly half of the informants (~47%) actually made more grammatical, morphological, syntactical and lexical mistakes in the face-to-face test within a stretch of 100 words. The results for this part of the analysis remain inconclusive, as more qualitative research would be required to determine the distribution of errors. However, overall it would seem that test-takers were slightly more accuracy-oriented in the face-to-face test. This tendency could arise from the fact that a real human interlocutor was present. In fact, in the interviews some of the informants explicitly mentioned having prioritized accuracy in the face-to-face test.

Moreover, measuring accuracy is relatively challenging, because it is not only difficult to determine which target accent to use as a baseline for examining the test-taker's pronunciation, but it is also difficult to determine what constitutes target-like use of vocabulary. Hence, determining what to consider an error proved to be difficult. One aspect which likely poses difficulties in any oral test is pronunciation because in addition to facing difficulties in selecting an appropriate target accent, it is also difficult to determine which features of speech are systematic (e.g. due to a learner's inability to produce certain phonemes) and which are caused by external factors such as the testing mode. Also, if the informant is nervous, his/her speech

may at times be extremely unclear. This does not imply that s/he does not know how to speak the language or articulate well but may simply arise from the fact that s/he is anxious due to the testing situation or even simply due to having to speak in a foreign language. L2 speaking has, after all, been associated with anxiety (Horwitz *et al*. 2010: 106). When considering accuracy, it is therefore important to define what the standard is against which test-takers' speech is assessed. This is particularly important when the rating of pronunciation is not the responsibility of a human rater, but rather that of a machine.

The third measure, fluency, indicates that in terms of breakdown fluency (i.e. pausing), test-takers appear to be more fluent in the face-to-face test. The differences in the percentage of silent pauses indicate that approximately 87% of the informants had a higher percentage of silence in the computer-based test. In other words, only two out of the fifteen participants paused more throughout their oral performance in the face-to-face test. The repair fluency measures, on the other hand, which demonstrate the extent of hesitation in the form of false starts, reformulations, repetition and replacements in the two tests, indicate that slightly over half (~53%) of the informants hesitated more in their oral performance in the face-to-face test.

Furthermore, I think it is important to consider that the interactive nature of the face-to-face test may in part explain the more frequent occurrence of hesitation in the face-to-face test and the tendency to pause more in the computer-based one. In other words, in the face-to-face test there are more filled pauses due to the fact that the interlocutor can, after a moment of silence, jump in and help the speaker when s/he is unable to express him/herself. In the same way, hesitation may arise more often in the face-to-face test because it may well be that the test-taker repeats something or otherwise feels the need to repair his/her speech because there is an interlocutor present. As is typical in interaction, the interlocutor may, at times, interrupt the test-takers' turn. In the face-to-face test I tried to intervene as little as possible, but the little intervention that did occur could at least partially explain these tendencies. Moreover, analyzing fluency is by no means an easy task. Repair fluency, in particular, is not easy to analyze, as the identification of false starts, reformulations, repetitions and replacements is open to interpretation.

Finally, due to the fact that the differences between the two test formats are minimal, one could assume that construct-irrelevant variance is not an issue when considering the computer-based test as a potential and valid alternative. However, as Kiddle and Kormos (2011: 355) point out, it may be possible that test administration conditions have an adverse effect on test-takers' oral performance at lower levels of language competence, and that the lack of interaction has an effect on test validity because it is more difficult to obtain extensive responses from lower level candidates in the computer-based test. In fact, I noticed that one of the informants who had a lower level of competence experienced more difficulties in the computer-based test. Not all researchers agree with this view, however. Kenyon and Malabonga (2001: 81), for example, state that one of the most obvious advantages of computer-based testing, and particularly that of computer-adaptive tests, is its "…ability to match task difficulty to examinee proficiency".

5.2 Differences in preferences

The second research question that this study aims to answer is whether the participants' attitudes differ when comparing the two test setups and what these attitudes are. The informants' preferences are, to a great extent, in line with previous research in that the majority indicated a preference for the face-to-face test. An even greater majority claimed to feel that they were better able to demonstrate their oral skills in the face-to-face test. The main reasons given for this preference are the more natural feeling of this particular testing format and the fact that it is easier to speak to a real person as opposed to a machine or an avatar. This may be due to the fact that, as Chapelle and Voss (2016: 120) point out, "[e]ven if people are accustomed to reading a computer screen, they may be less comfortable talking to a computer screen". Some of the test-takers also felt that they had fewer problems with their working memory and were more fluent in their performance in the face-to-face test. However, I feel that it is important to take into account that some informants did indeed prefer the format which lacked a human interlocutor because it made them feel less nervous. Anxiety related to speaking a foreign language as well as taking tests are important external factors to consider in an oral test, as they can adversely affect test performance. In fact, as Qian (2009: 123) points out, "…it would be reasonable to assume that, if a test-taker's state of mind or disposition is affected by

the testing mode in some negative way, the affective filter[32] may also be up to interfere with his or her test performance". In fact, a slight majority (~53%) of informants indicated having felt more nervous on one of the two tests, with the majority claiming to have felt more uneasy in the computer-based test.

An aspect which has significant implications for test validity and reliability is the question of whether the test-takers felt that both exams exclusively tested oral skills as opposed to other (irrelevant) skills. 80% of the informants were of the opinion that both tests were equally valid, but the other 20% claimed to feel disadvantaged on one or the other, or both of the tests. This is probably the most important aspect which needs to be considered when it comes to test design so as to assure fairness. Interpersonal skills were brought up in one of the informants' comments. This is of course a tricky matter, as some test-takers are likely to have weaker interpersonal skills than others. However, given that a (foreign) language test ideally consists of a number of different tasks testing different skills (e.g. written production), it is likely that such disadvantages even out. Factors such as IT skills or issues arising from a particular task type requiring a specific set of skills (e.g. one requiring a map and thus orientation skills), on the other hand, are more problematic because they inherently lead to construct-irrelevant variance, thus skewing test scores.

Tossavainen (2016: 40) elaborates on this point by saying that although it is not possible to always please everyone, it is enough, ethically, to attempt to be as just as possible using all available means. This requires research and awareness of the issues surrounding these ethical questions. This is, in fact, one of the reasons that I felt it was important to ask test-takers about their preferences. It would not have been enough to simply compare the CAF measures in terms of the two tests because, as one could see, although the differences in terms of the measures were small, the differences in preferences were far greater.

However, Qian (2009: 123) correctly claims that it is necessary to avoid making hasty conclusions regarding test-taker preferences. This is because these may be defined by the test or context. In other words, test-takers' attitudes are likely to be influenced by factors such as test quality, stakes or, as I pointed out earlier, the

[32] The affective filter relates to Krashen's *Affective Filter Hypothesis*, which acknowledges the view that emotions play an important role in foreign language learning (Mitchell *et al*. 2013: 45).

test-taker's interpersonal characteristics and/or cultural background. In fact, one aspect which most likely affects the results of studies such as the present one is test quality as well as the low stakes of the test. It was not possible to design a test which would replicate the high-stakes of the matriculation examination or even a course-specific summative assessment. I would argue that were test-takers' performances evaluated for something which had consequences for their future, their performance on the test(s) and their attitudes would most likely have been different.

5.3 Implications of the findings

Despite the fact that this study is not able to replicate a high-stakes test situation, it does have some important implications, especially for the Finnish matriculation examination context. There are a number of things which need to be taken into account when designing a high-stakes computer-based test.

The first is to reflect on the question of whether it really is the best way, all things considered, to test speech with the help of a computer. In terms of test validity, it is, after all, crucial to select the appropriate test depending on the purpose of the test and the intended use of its results (Malone and Montee 2010: 983). Moreover, considering that the communicative language teaching and testing paradigms accentuate interaction, it is of utmost importance to consider whether using a computer is the best way to capture this aspect or oral proficiency. As Kenyon and Malabonga (2001: 82) state, "…in instances where an evaluation of interactional competence is critical, it may be quite a while before it can be replicated with technology". Of course, progress has been made since then, but I would argue that it is, still today, incredibly difficult to evaluate interaction as well as pronunciation and suprasegmental features of speech with existing technology. As discussed earlier, it is incredibly difficult to determine which variety to consider the standard against which to assess.

I also think that it is important to remember that the computerization of a test does not mean that it is a mere conversion of an existing (oral) test. A computer screen with instructions is not the same as an assessor giving those same instructions and making sure the test-taker correctly completes the task. In my study, for example, a factor which may have adversely affected the informants' performance in the

computer-based test is the fact that some of the test-takers did not read the instructions carefully enough. Task 3, in particular, appeared to cause problems. These may in part have been due to the students having to scroll up and down the page. Fulcher (2003a: 390–392), for example, states that in the design of a computer-based test it is important to make sure that the navigational structure of the test is clear and does not require undue attention. The amount of information presented on a single screen should be limited because excessive detail makes it difficult to distinguish between what is relevant for the completion of a test task and what is not. In the interviews, the test-takers mentioned that scrolling caused difficulties in the computer-based test because it is easier to visualize the task and its instructions all at once. The effects of such factors should therefore be minimized so as not to affect test performance. However, considering these issues in light of the Finnish matriculation examination and particularly the spoken test administered via a computer in the future, it is likely that such problems will be minimized due to large-scale piloting and the significant washback effect that the test will likely have. In other words, students are likely to practice for the test and will therefore be familiar with the task types as well as the test structure.

5.4 Limitations of the study

Despite the aforementioned findings and their important implications, the present study is subject to a number of limitations. The first of these involves the fact that the data collection process was limited to just one school and 15 informants. The findings cannot therefore be generalized. Furthermore, the academic achievement of the students in the school the data was collected from is above national average[33]. Moreover, self-selection of informants may have resulted in only students with certain qualities (e.g. extroversion) to take part in the study, thus not giving an accurate picture of the average student. Some of the informants had also previously attended a bilingual school, where one of the teaching languages is English. These factors may affect the results of the current study. I originally aimed to conduct this study in two different schools, but due to certain restrictions, I had to limit this study to one school only. More research is definitely needed. It would be worthwhile to

[33] This observation is based both on the high GPA required to get into this particular upper secondary school as well as the mean results on the matriculation examination of the student population.

conduct a study with a greater number of participating teachers and students from a geographically wider area.

In addition to the amount of data posing a major limitation, the methods adopted to answer the present study's research questions should also be discussed in critical light. In fact, a limitation may arise from the fact that the tests were completed in the same order, i.e. the face-to-face test was always completed before the computer-based test. Such test practice effects are of course accentuated by the fact that the test tasks in the two tests are so similar. For the purposes of this study it was necessary to create two highly similar tests so as to answer the question of whether *testing mode alone* has an effect on oral performance. Also, as Skehan (2001: 182) points out,

> …there may be significant consequences when one task is chosen rather than another. Or to spell this out even more directly, if candidate performances are compared after having been elicited through the use of different tasks, the performances themselves may be very difficult to relate to one another.

For this reason, it was crucial that the tasks be comparable in terms of content. Wigglesworth (2001: 205) also adds that attention must be paid to task parameters because relatively minor changes in task characteristics and/or conditions can have a significant impact on the scores obtained.

It should be noted, however, that the two testing modes potentially lend themselves to testing the same construct (i.e. oral proficiency), but in different ways. This is because in designing a computer-based test, it is possible to take advantage of certain functionalities which are only available on computers. Perhaps the most obvious way of doing this is the integration of different modalities (e.g. video content) which may add to the authenticity of the test. However, computers also enable innovative response formats (Sawaki 2012: 428). In theory, the two testing modes could test different underlying skills of the same construct or even different constructs, considering that in a computer-based test, some basic computer skills are a prerequisite for successfully completing the test. More research is needed to determine the impact of multimedia on test performance because, as seen in the piloting phase of the computer-based test, videos can be cognitively more demanding to process in comparison to pictures.

Another question which requires some thought is that of whether the tasks chosen for the two tests were adequate. The aim was to achieve a balance in terms of having an array of tasks which vary in their cognitive load and the type of communicative features needed for completing them. The choice of tasks was given a lot of thought, as factors such as personal characteristics, topical knowledge and affective schemata are not allowed to interfere with task performance. Despite the attempt of reducing these effects, the instruction (map) task appeared to cause some of the informants undue stress and a feeling of being disadvantaged. In addition to the question of task quality is that of quantity, i.e. whether the number of tasks in the two tests is sufficient to get an accurate picture of the test-takers' oral skills and to answer the first research question. Due to time constraints it was not possible to exceed four tasks per test. However, as Mousavi (2009: 42) points out, "[i]t is an accepted fact that the longer the test, the more reliable the results will be".

Moreover, Moodle, the authoring system used, was not necessarily ideal. In fact, one of the biggest factors to have affected the results of the present study is probably the fact that it was impossible to replicate a high-stakes testing situation. The situation could have been made more realistic by having all of the test-takers complete the computer-based test at the same time in a language studio setting. Unfortunately, this was not possible due to the limited time and resources available. Also, the computer-based test would have been more realistic had it been possible to have the informants directly record their speech on the computer and had it been possible to incorporate a timer on the screen. More elaborate studies which have these resources available to them are needed.

A final limitation involves a feature which is inherent to the type of qualitative analysis used in this study. Although the analysis was done systematically and with great care, it did involve some interpretation and is likely to have been affected, to a certain degree, by human error. Spoken discourse is oftentimes characterized by ellipsis. In the CAF analysis, this resulted in the need to occasionally 'fill in the gaps' to reconstruct a grammatically complete utterance. Finally, the CAF measures need to be complemented by more qualitative and deeper-level analysis because, as Lennon (2000: 25) points out with regard to fluency, surface analysis findings

> are not unambiguous, since these features also have other perfectly legitimate, rhetorical, and even communicative functions in discourse. There is thus the problem of multifunctionality, which mere mechanical counting of surface features glosses over, ignoring the psycholinguistics of production.

The same argument can, in my opinion, be made with reference to accuracy and complexity. Therefore, in order to allow for comparability of studies, more research with well-established, standardized CAF measures is needed.

6 Conclusion

This thesis answers the following research questions:

1. *What differences are there in upper secondary school students' oral performance when comparing a face-to-face test to a computer-based equivalent in terms of:*
    a. *complexity*
    b. *accuracy, and*
    c. *fluency?*
2. *What types of attitudes do students have regarding these two tests?*

The key findings of this thesis are the following:

- The differences in terms of complexity, accuracy and fluency in the two tests are relatively small, suggesting that the testing mode does not have a major impact on oral performance.
- The measures used to determine complexity were (i) the amount of subordination and (ii) the type-token ratio. Test-takers indicated a greater amount of subordination and a higher type-token ratio in the computer-based test.
- The measures used to determine accuracy were (i) the percentage of error-free clauses and (ii) the number of errors per 100 words. The first measure indicated that approximately 73% of test-takers had a higher percentage of error-free clauses in the face-to-face test, but the second measure challenged this view, as it suggested that approximately 53% of test-takers made more phonological, morphological, syntactic and lexical errors in the face-to-face test. The results are therefore inconclusive.
- The measures used to determine fluency were (i) the percentage of silent pauses and (ii) the number of hesitation markers. The results show that test-takers had a higher percentage of silent pauses in the computer-based test and approximately 53% of test-takers hesitated more in the face-to-face test.

- The majority of test-takers showed a preference for the face-to-face test due to its interactive nature and felt that they were also better able to demonstrate the extent of their oral proficiency in the face-to-face test.

The results also indicate that Finnish upper secondary school students are open to the possibility of their oral skills being tested and that they generally would not mind a computer-based oral test, given that it is fair and valid.

References

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (1997). *Language testing in practice*. Oxford: Oxford University Press.

Biber, D., Conrad, S. and Leech, G. (2002). *Student grammar of spoken and written English*. Essex: Pearson Education Limited.

Brown, G. and Yule, G. (1983). *Teaching the spoken language: An approach based on the analysis of conversational English*. Cambridge: Cambridge University Press.

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

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

Bygate, M., Skehan, P. and Swain, M. (2001). *Researching pedagogic tasks: Second language learning, teaching and testing*. Essex: Pearson Education Limited.

Bygate, M. (2011). Teaching and testing speaking. In: Long, M. H. and Doughty, C. J. (eds.) *The Handbook of Language Teaching*. Malden MA: Wiley Blackwell.

Bygate, M. (2012). Speaking. In: Kaplan, R. B. (ed.) *The Oxford handbook of applied linguistics*. Oxford: Oxford University Press.

Cameron, D. (2001). *Working with spoken discourse*. London: SAGE Publications Ltd.

Canale, M. and Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics*, 1, 1–47.

Cassady, J. C. (2010). Test anxiety: Contemporary theories and implications for learning. In: Cassady, J. C. (ed.) *Anxiety in schools: The causes, consequences, and solutions for academic anxieties*. New York: Peter Lang Publishing.

Chafe, W. L. (1985). Linguistic differences produced by differences between

speaking and writing. In: Olson, D. R, Torrance, N. and Hildyard, A. (eds.) *Literacy, language, and learning: The nature and consequences of reading and writing*. Cambridge: Cambridge University Press.

Chaloub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In: Bygate, M., Skehan, P. and Swain, M. (eds.) *Researching pedagogic tasks: Second language learning, teaching and testing*. Essex: Pearson Education Limited.

Chapelle, C. A. and Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chapelle, C. A. and Voss, E. (2016). 20 years of technology and language assessment in language learning & technology. *Language Learning & Technology*, 20(2), 116–128.

Clark, J. L. D. (1979). Direct versus semi-direct tests of speaking proficiency. In: Brière, E. J. and Hinofotis, F. B. (eds.) *Concepts in language testing: some recent studies*. Washington DC: Teachers of English to speakers of other languages, 35–49.

Council of Europe (2012). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

de Jong, N. H. (2016). Fluency in second language assessment. In: Tsagari, D. and Banerjee, J. (eds.) *Handbook of second language assessment*. Handbook of Second Language Assessment. Berlin: De Gruyter.

DigiTala, Puhumisen sähköisen arvioinnin tutkimushanke, https://blogs.helsinki.fi/digitala-projekti/about-digitala/ [Accessed 8 May 2019].

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Douglas, D. (2010). *Understanding language testing*. Abingdon: Routledge.

Drackert, A. (2015). *Validating language proficiency assessments in second language acquisition research: Applying an argument-based approach*. Frankfurt am Main: Peter Lang.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.

Ellis, R. and Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford

University Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Finnish National Board of Education (2010). Tiedote: *Vieraiden kielten ja toisen kotimaisen kielen suullisen kielitaidon arviointi lukiossa*, http://www.oph.fi/download/124430_Tiedote_49_2010.pdf [Accessed 8 May 2019].

Finnish National Board of Education (2016). *National core curriculum for general upper secondary schools 2015*. Helsinki: Finnish National Board of Education.

Foster, P. and Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research*, 3(3), 215–247.

Foster, P., Tonkyn, A. and Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354–375.

Fulcher, G. (2003a). Interface design in computer-based language testing. *Language Testing*, 20(4), 384–408.

Fulcher, G. (2003b). *Testing second language speaking*. London: Pearson Longman.

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly*, 11(2), 186–197.

Hildén, R. (2000). *Att tala bra, bättre och bäst: Suomenkielisten abiturienttien ruotsin kielen suullinen taito testisuoritusten valossa*. Helsinki: Hakapaino.

Hildén, R. and Vähähyyppä, K. (2016) Suullisten taitojen arviointi mukaan kielikokeisiin ehkä vuonna 2020. *Helsingin Sanomat*, 14.5.2016.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Huhta, A. (1993). Teorioita kielitaidosta – Onko niistä hyötyä testaukselle? In: Takala, S. (ed.) *Suullinen kielitaito ja sen arviointi*. Jyväskylän yliopisto.

Huhta, A. and Hildén, R. (2016). Kielitutkinnot ja muu laajamittainen kielitaidon arviointi Suomessa. In: Huhta, A. and Hildén, R. (eds.) Kielitaidon arviointitutkimus 2000-luvun Suomessa. *AFinLA-e. Soveltavan kielitieteen tutkimuksia* 9, 3–26.

Horwitz, E. K., Tallon, M. and Luo, H. (2010). Foreign language anxiety. In: Cassady, J. C. (ed.) *Anxiety in schools: The causes, consequences, and solutions for academic anxieties*. New York: Peter Lang Publishing.

Housen, A., Vedder, I. and Kuiken, F. (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins Publishing Company.

Isaacs, T. (2016). Assessing speaking. In: Tsagari, D. and Banerjee, J. (eds.) *Handbook of second language assessment*. Berlin: De Gruyter.

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

Kenyon, D. M. and Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessment. *Language Learning and Technology*, 5(2), 60–83.

Kenyon, D. M. and Malone, M. (2010). Investigating examinee autonomy in a computerized test of oral proficiency. *European Commission Report*. Ispra: European Union.

Kenyon, D. M. and Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal*, 84(i), 85–101.

Kiddle, T. and Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly*, 8, 342–360.

Koponen, M. and Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In: Riggenbach, H. (ed.) *Perspectives on fluency*. Michigan: The University of Michigan Press.

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

Lennon, P. (2000). The lexical element in spoken second language fluency. In: Riggenbach, H. (ed.) *Perspectives on fluency*. Michigan: The University of Michigan Press.

Luoma, S. (1997). Comparability of a tape-mediated and a face-to-face test of speaking: A triangulation study. University of Jyväskylä.

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

Malabonga, V., Kenyon, D. and Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, 22(1), 59 – 92.

Malone, M. E. and Montee, M. J. (2010). Oral proficiency assessment: Current approaches and applications for post-secondary foreign language programs. *Language and Linguistics Compass*, 4(1), 972–986.

Matriculation Examination Board, History. https://www.ylioppilastutkinto.fi/en/matriculation-examination/history [Accessed 8 May 2019].

Ministry of Education (2006). *Lukiokoulutuksen suullisen kielitaidon arviointiryhmän muistio*. Helsinki: Yliopistopaino.

Mitchell, R., Myles, F. and Marsden, E. (2013). *Second language learning theories*. Abingdon: Routledge.

Mousavi, S. A. (2009). Multimedia as a test method facet in oral proficiency tests. *International Journal of Pedagogies and Learning*, 5, 37–48.

Morrow, K. (1986). The evaluation of tests of communicative performance. In: Portal, M. (ed.) *Innovations in language testing*. Windsor: NFER-Nelson: 1–13.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*, 93, 836–847.

O'Loughlin, K. (1997). The comparability of direct and semi-direct speaking tests: A case study. Ph.D. dissertation. University of Melbourne.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge: Cambridge University Press.

O'Sullivan, B. (2008). *Notes on assessing speaking*. Cornell University: Language Resource Center.

Pižorn, K. and Huhta, A. (2016). Assessment in educational settings. In: Tsagari, D. and Jayanti, B. (eds.) *Handbook of Second Language Assessment*. Berlin: De Gruyter.

Pollari, P. (2016). Daunting, reliable, important or "trivial nitpicking?" Upper secondary students' expectations and experiences of the English test in the Matriculation Examination In: Huhta, A. and Hildén, R. (eds.) Kielitaidon arviointitutkimus 2000-luvun Suomessa. *AFinLA-e. Soveltavan kielitieteen tutkimuksia* 9, 184–211.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking

assessment: Affective effects on test takers. *Language Assessment Quarterly*, 6(2), 113–125.

Richards, J. C. and Rodgers, T. S. (2014). *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22(1), 27–57.

Sajavaara, K. and Dufva, H. (2001). Finnish-English Phonetics and Phonology. *International Journal of English Studies*, 1(1), 241–256.

Saleva, M. (1994). Kokeestako se kiikastaa: Lukion suullisen englannin kielen päättökokeen kehittelyä. In: Laurinen, L. and Luukka, M.-R. (eds.) *Puhekulttuurit ja kielten oppiminen*. Jyväskylä: Suomen soveltavan kielitieteen yhdistys (AFinLa), 52, 277–291.

Saleva, M. (1997). *Now they're talking: Testing oral proficiency in a language laboratory*. University of Jyväskylä.

Sawaki, Y. (2012). Technology in language testing. In: Fulcher, G. and Davidson, F. (eds.) *The Routledge Handbook of Language Testing*. Abingdon: Routledge, 426–437.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11(2), 99–123.

Skehan, P. and Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1(3), 185–211.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics*, 30(4), 510–532.

Skehan, P. (2001). Tasks and language performance assessment. In: Bygate, M., Skehan, P. and Swain, M. (eds.) *Researching pedagogic tasks: Second language learning, teaching and testing*. Essex: Pearson Education Limited.

Skehan, P. (2003). Focus on form, tasks and technology. *Computer Assisted Language Learning*, 16(5), 391–411.

Surface, E., Poncheri, R. and Bhavsar, K. (2008). Two studies investigating the reliability and validity of the English ACTFL OPIc with Korean test takers: The ACTFL OPIc validation project technical report.

Suvorov, R. and Hegelheimer, V. (2014). Computer-assisted language testing. In: A. J. Kunnan (ed.) *The Companion to Language Assessment*. Malden MA: Wiley Blackwell.

Thompson, G. L., Troy, L. C. and Nieves, K. (2016). Comparing the OPI and the OPIc: The effects of test method on oral proficiency scores and student preference. *Foreign Language Annuals*, 49(1), 75–92.

Tossavainen, H. (2016). Kielitestien eettisyydestä ja oikeudenmukaisuudesta. In: Huhta, A. and Hildén, R. (eds.) (2016). Kielitaidon arviointitutkimus 2000-luvun Suomessa. *AFinLA-e. Soveltavan kielitieteen tutkimuksia* 9, 27–43.

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

Ussama, R. and Sinwongsuwat, K. (2014). Conversation proficiency assessment: A comparative study of two-party peer interaction and interview interaction implemented with Thai EFL. *International Journal of Language Studies*, 8(4), 95–106.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. TESOL Quarterly, 23(3), 489–508.

Weir, C. J. (2005). *Language testing and validation – an evidence-based approach*. Houndmills: Palgrave.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In: Bygate, M., Skehan, P. and Swain, M. (eds.) *Researching pedagogic tasks: Second language learning, teaching and testing*. Essex: Pearson Education Limited.

Young, R. and Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14, 403–424.

Appendices

Appendix 1 – *Visual representation of direct and semi-direct test formats*

Figure 7 – *Face-to-face test setup (Underhill 1987: 28)*

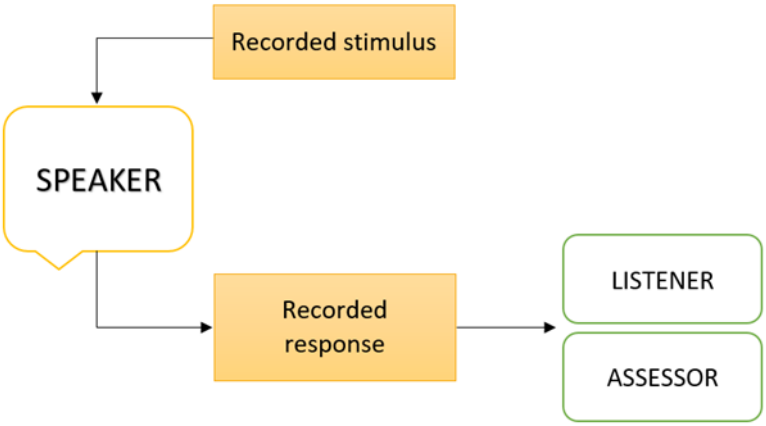


Figure 8 – *Tape-mediated test setup (Underhill 1987: 34)*

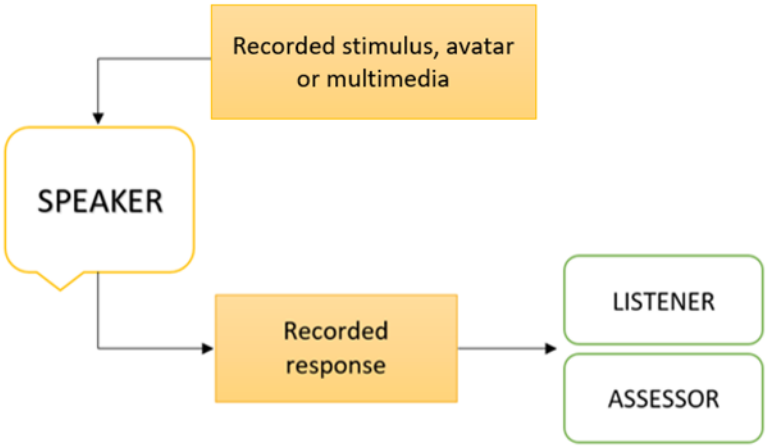


Figure 9 – *Computer-based test setup*

Appendix 2 – *Permission form*

Tutkimuslupapyyntölomake

Hyvä opiskelija,

Tutkin pro gradu -työssäni lukio-opiskelijoiden suullista kielitaitoa kahdessa eri koetilanteessa: perinteisessä kasvokkaisessa ja sähköisessä muodossa. Tutkimukseni aineisto koostuu siten äänityksistä, joita kerään sanelimen ja/tai tietokoneen avulla opiskelijoiden suorittaessa kokeet.

Olen lisäksi kiinnostunut siitä, kumpaa koemuotoa opiskelijat pitävät parempana, mistä kerään lopuksi lisätietoa kyselylomakkein sekä ryhmähaastatteluin.

Opiskelijoiden vastaukset/keskustelut transkriboidaan, joten saatan tutkimuksessani siteerata niitä. Vastauksista/keskusteluista ei kuitenkaan ilmene henkilökohtaisia tietoja tai tunnisteita. Materiaali on luottamuksellista ja se tullaan säilyttämään salassa opiskelijoiden henkilöllisyyttä suojellen, eikä sitä luovuteta ulkopuolisille. Äänitykset tuhotaan, kun niitä ei enää tarvita. Osallistuminen on vapaaehtoista, ja opiskelija voi halutessaan keskeyttää osallistumisensa koska tahansa.

Jos päätät osallistua ja pystyt tulemaan molempiin kokeisiin (##PÄIVÄMÄÄRÄ## & ##PÄIVÄMÄÄRÄ##) sekä haastatteluun (##PÄIVÄMÄÄRÄ##), allekirjoita lomake. Mikäli olet alaikäinen, pyydä vanhemmaltasi allekirjoitus.

Otathan yhteyttä, mikäli kaipaat lisätietoja. Yhteystietoni löytyvät sivun alareunasta.

Annan luvan käyttää suoritustani tutkimuksessa.

Opiskelijan allekirjoitus ja nimenselvennys: _____

Aika ja paikka: ____/____/2017_____

Annan luvan käyttää lapseni suoritusta tutkimuksessa.

Vanhemman allekirjoitus ja nimenselvennys: _____

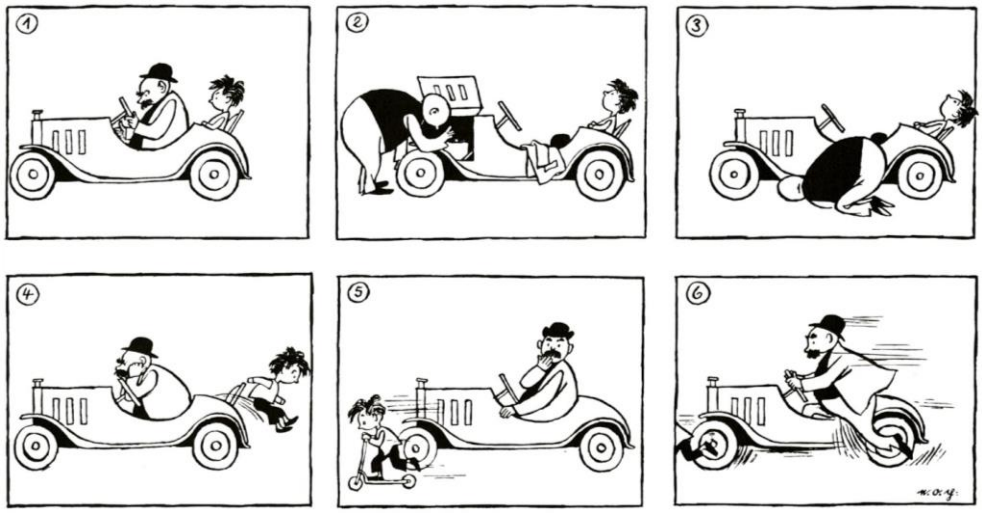
Aika ja paikka: ____/____/2017_____

Ystävällisin terveisin,
Essi Pohto

Yhteystiedot:
##SÄHKÖPOSTI##
##PUHELINNUMERO##

Appendix 3 – *Face-to-face test tasks*

Osa 1 – 1 minuutti
Tutki alla olevaa sarjakuvaa ja kerro sitten **englanniksi**, mitä siinä tapahtuu. Kuvaile **henkilöhahmoja** ja **tapahtumia**, ja pohdi, mikä **vika** autoon tulee.



Osa 2 – 2 minuuttia
Tutki hetki seuraavaa kahta kuvaa. Kerro sitten **englanniksi,** miten ne **muistuttavat** toisiaan ja miten ne **eroavat** toisistaan. Kumpi kuva on sinulle **omempi**?

Kuva A

Kuva B

Osa 3 – 5 minuuttia

Teeskennellään, että olen ystäväsi, joka tulee käymään Englannista. Keskustelemme siitä, mitä kaikkea minun tulisi nähdä Helsingissä. Valitse kartasta **3-5** kohdetta, jotka ovat **mielestäsi näkemisen arvoisia** ja kerro sitten **englanniksi**, **miksi** ja **miten** pääsemme paikasta toiseen.

Osa 4 – 5 minuuttia

Olet päättänyt pitää välivuoden ja etsit töitä. Voit joko olla oma itsesi tai teeskennellä olevasi fiktiivinen henkilö. Valitse **yksi** näistä kolmesta vaihtoehdosta ja **kerro itsestäsi** ja **miksi** haluaisit kyseisen työpaikan ja miksi juuri sinä olisit sopiva siihen. Minä toimin haastattelijana.

Support Worker

Are you able to care for others? Do you have good listening skills? Do you understand how older people feel?

If you have answered **YES** to the above, you might be just who we are looking for.

We provide excellent training, a free uniform and a great work atmosphere. If you are interested, please contact us at: goldencare@gc.co.uk.

GOLDEN CARE

Baristas wanted!

G'Day! Have you always wanted to work in a café in Sydney, Australia and make a quid? Well, this might be your chance!

We are recruiting young baristas from around the world who love tea and coffee!

Are you outgoing and a hard worker who doesn't forget to smile? Then hurry and e-mail us at: welovecoffee@coffeehouse.au.

Summer Camp Jobs

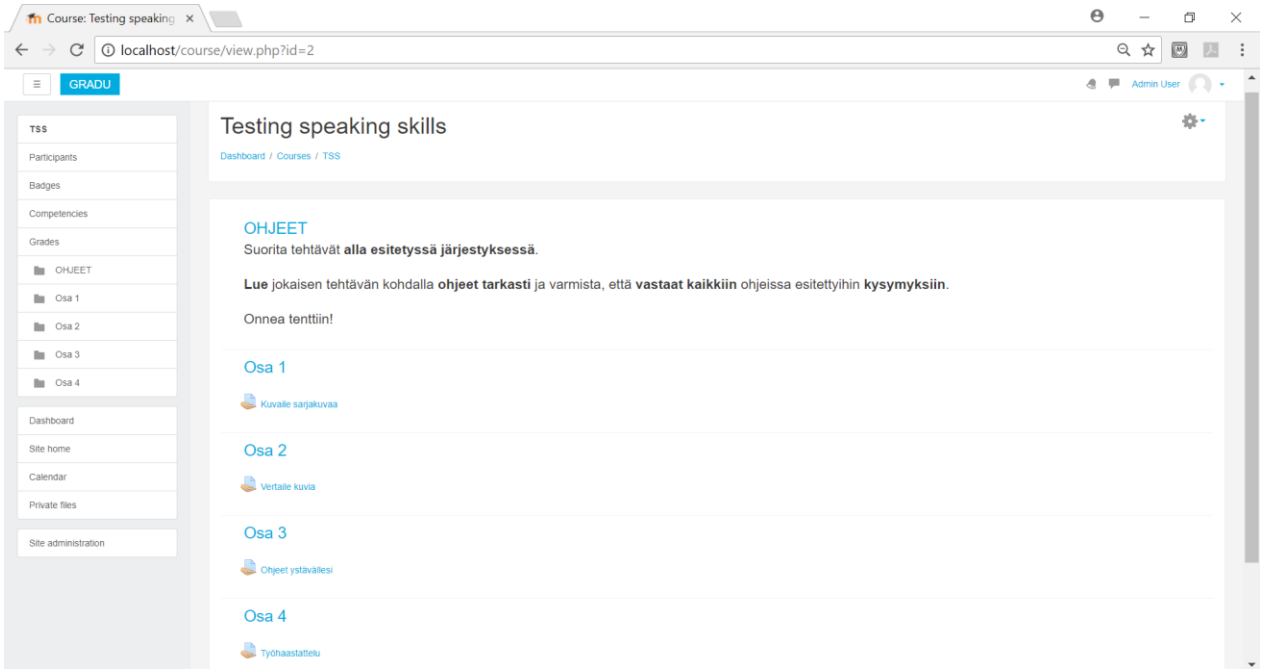
Hey you!

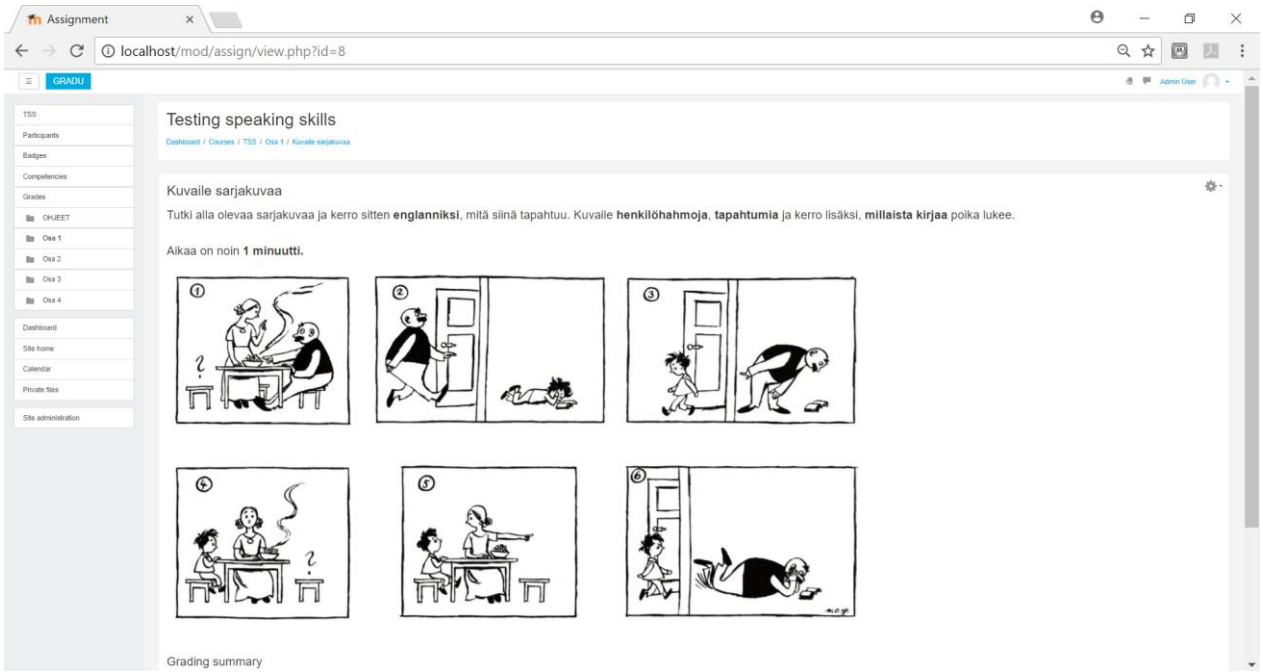
Would you be interested in working at our summer camp in Sanger, California this year?

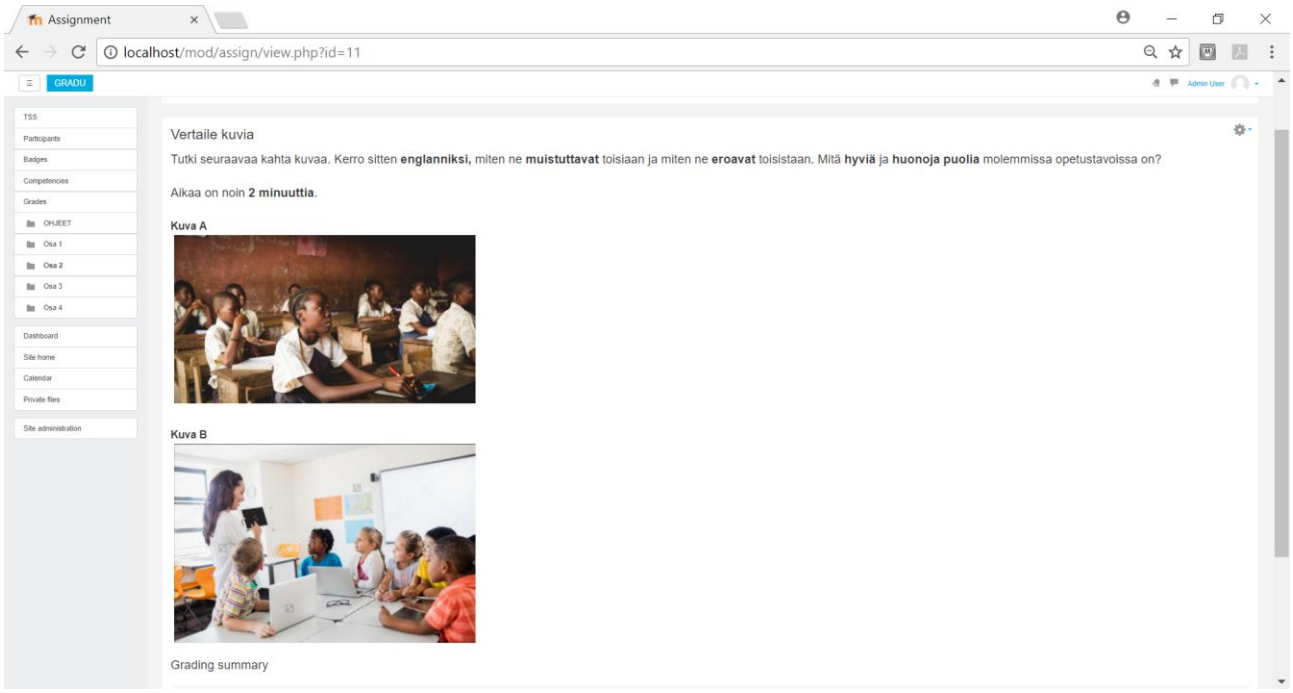
We are looking for young athletic and social individuals with enthusiasm to organize activities for 6-10-year-old kids. Good English skills are a must!

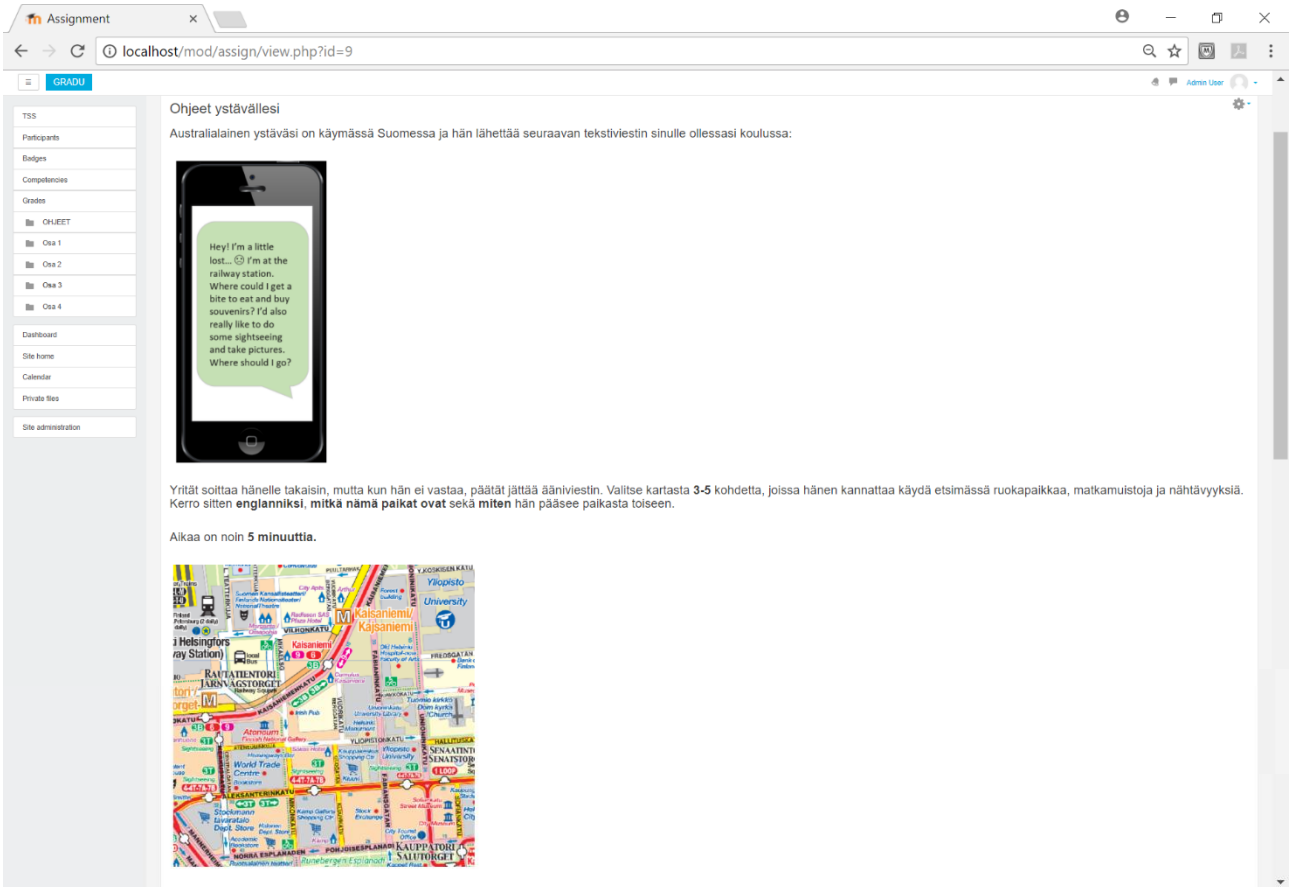
If you are interested, please contact us at: summer2017@camp.com. We will get back to you as soon as possible!

Appendix 4 – *Computer-based test tasks*



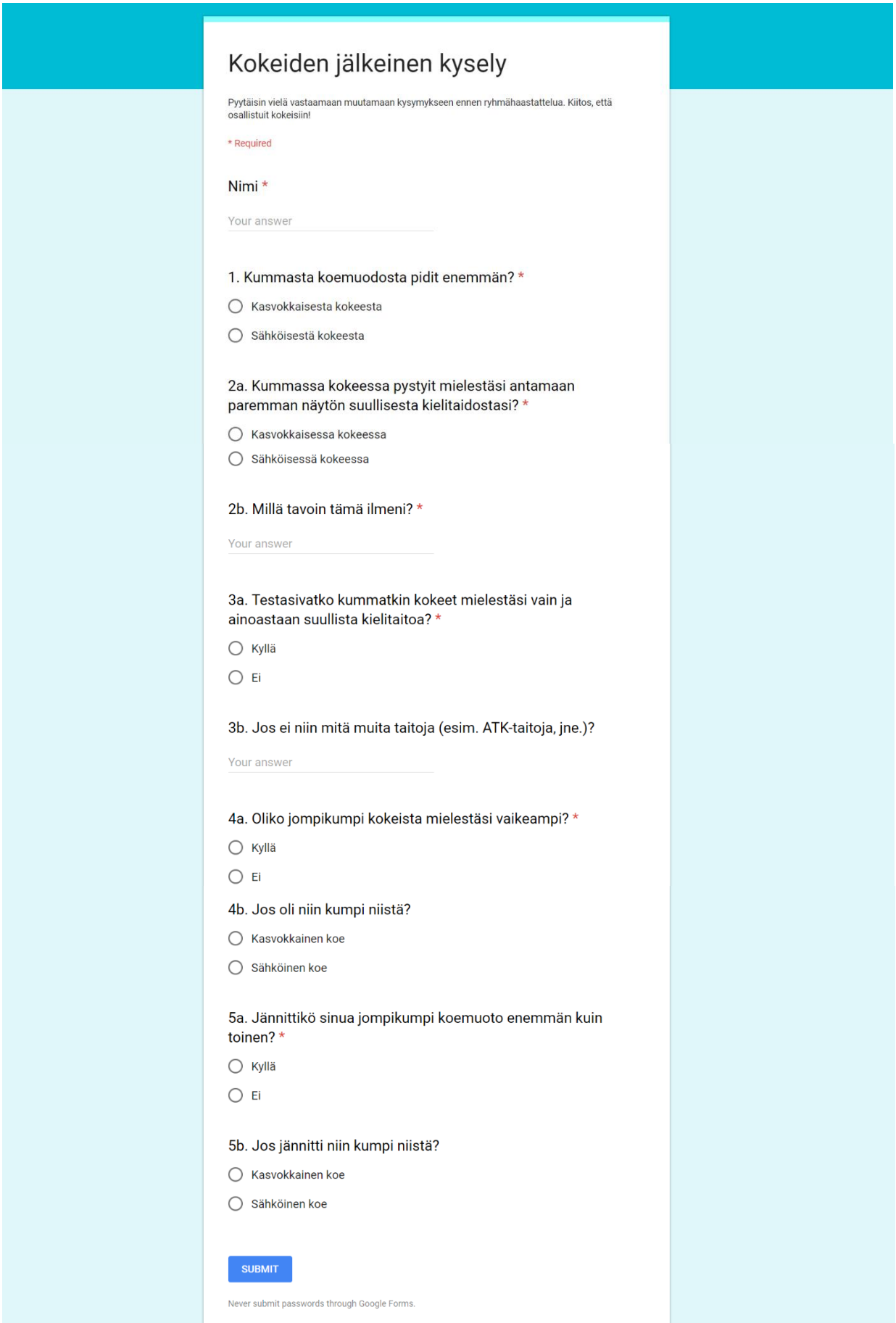



Assignment
localhost/mod/assign/view.php?id=11
GRADU
Admin User
TSS
Participants
Badges
Competencies
Grades
OHJEET
Osa 1
Osa 2
Osa 3
Osa 4
Dashboard
Site home
Calendar
Private files
Site administration
Vertaile kuvia
Tutki seuraavaa kahta kuvaa. Kerro sitten englanniksi, miten ne muistuttavat toisiaan ja miten ne eroavat toisistaan. Mitä hyviä ja huonoja puolia molemmissa opetustavoissa on?
Aikaa on noin 2 minuuttia.
Kuva A
Kuva B
Grading summary


Assignment
localhost/mod/assign/view.php?id=9
GRADU
Admin User
TSS
Participants
Badges
Competencies
Grades
OHJEET
Osa 1
Osa 2
Osa 3
Osa 4
Dashboard
Site home
Calendar
Private files
Site administration
Ohjeet ystävällesi
Australialainen ystäväsi on käymässä Suomessa ja hän lähettää seuraavan tekstiviestin sinulle ollessasi koulussa:
Hey! I'm a little lost... ☹ I'm at the railway station. Where could I get a bite to eat and buy souvenirs? I'd also really like to do some sightseeing and take pictures. Where should I go?
Yrität soittaa hänelle takaisin, mutta kun hän ei vastaa, päätät jättää ääniviestin. Valitse kartasta 3-5 kohdetta, joissa hänen kannattaa käydä etsimässä ruokapaikkaa, matkamuistoja ja nähtävyyksiä. Kerro sitten englanniksi, mitkä nämä paikat ovat sekä miten hän pääsee paikasta toiseen.
Aikaa on noin 5 minuuttia.


Assignment
localhost/mod/assign/view.php?id=14
GRADU
Admin User
TSS
Participants
Badges
Competencies
Grades
OHJEET
Osa 1
Osa 2
Osa 3
Osa 4
Dashboard
Site home
Calendar
Private files
Site administration
Työhaastattelu
Olet päättänyt pitää välivuoden ja etsit töitä. Voit joko olla oma itsesi tai teeskennellä olevasi fiktiivinen henkilö. Valitse yksi alla olevista kolmesta vaihtoehdosta. Kun olet päättänyt, minkä vaihtoehdoista valitset, siirry sivun alaosaan ja aloita diaesitys. Kate (avatar) toimii puhekumppaninasi ja haastattelee sinua.
Aikaa on 5 minuuttia.
Vaihtoehto 1
A once-in-a-lifetime chance to work at sea!
We are looking for adventurous young people to work on the Mariner of the Seas cruise ship.
We expect you to love traveling and to have experience in customer service, as we are known to provide an extraordinary level of service and care to our guests.
We provide excellent training, the chance to work with people from all over the word and best of all, free world travel.
If you are interested, please contact our human resources department at: cruise@mariner.com.
Vaihtoehto 2
Elephant Sanctuary in Plettenburg, South Africa
Do you feel compassion towards animals and have a strong need to protect a threatened species?
If you answered 'yes' to the above questions, we want you on our team!
The Elephant Sanctuary in Plettenburg is home to nearly 200 elephants. We are looking for animal-loving and responsible individuals to take care of our friendly giants.
Call us at +2756077321, if you are up for the challenge!
Vaihtoehto 3
Hotel receptionist needed
Aloha! We are looking for a part-time receptionist who could work 20 hours a week at our Honolulu hotel in Hawaii.
Applicants should be organized and dependable with good communication skills. They should also be able to perform general office assistant functions.
We provide:
- A basic salary of $10/hour
- Excellent benefits
- A friendly work atmosphere by the sea!
Contact us at hr@meridien.com.
voki
4/7
Työhaastattelu
Kate (avatar) toimii puhekumppaninasi ja hän haastattelee sinua. Vastaa kysymyksiin mahdollisimman täydentävästi.
Kun olet valmis, siirry eteenpäin painamalla nuolta.
1/7
Huom! Aloita diaesitys painamalla nuolta oikealla puolella (ei siis play-nappia)!
Huom! Aloita diaesitys painamalla nuolta oikealla puolella (ei siis play-nappia)!
Grading summary
Participants
7
View all submissions
Grade

Appendix 5 – *Questionnaire*



Kokeiden jälkeinen kysely

Pyytäisin vielä vastaamaan muutamaan kysymykseen ennen ryhmähaastattelua. Kiitos, että osallistuit kokeisiin!

* Required

Nimi *

Your answer

1. Kummasta koemuodosta pidit enemmän? *

- Kasvokkaisesta kokeesta
- Sähköisestä kokeesta

2a. Kummassa kokeessa pystyit mielestäsi antamaan paremman näytön suullisesta kielitaidostasi? *

- Kasvokkaisessa kokeessa
- Sähköisessä kokeessa

2b. Millä tavoin tämä ilmeni? *

Your answer

3a. Testasivatko kummatkin kokeet mielestäsi vain ja ainoastaan suullista kielitaitoa? *

- Kyllä
- Ei

3b. Jos ei niin mitä muita taitoja (esim. ATK-taitoja, jne.)?

Your answer

4a. Oliko jompikumpi kokeista mielestäsi vaikeampi? *

- Kyllä
- Ei

4b. Jos oli niin kumpi niistä?

- Kasvokkainen koe
- Sähköinen koe

5a. Jännittikö sinua jompikumpi koemuoto enemmän kuin toinen? *

- Kyllä
- Ei

5b. Jos jännitti niin kumpi niistä?

- Kasvokkainen koe
- Sähköinen koe

SUBMIT

Never submit passwords through Google Forms.

Appendix 6 – *Interview questions*

HAASTATTELUKYSYMYKSET

Teema 1: *graduun liittyvät kokeet*

1. Pystyittekö näyttämään vahvuutenne sekä heikkoutenne molemmissa kokeissa?
2. Tuntuiko teistä siltä, että teidän kielenkäyttönne olisi muuttunut riippuen koetyypistä? Miten?
3. Ovatko kokeet teidän mielestänne toisiinsa verrattavia? Koitteko että reiluudessa oli eroja kokeiden välillä? Miksi?
4. Olivatko tehtävät toisiinsa verrattavissa?
5. Tuntuiko teistä jompikumpi koe vaikeammalta? Miksi?
6. Jännittikö teitä jompikumpi koe enemmän? Miksi?
7. Mitä ovat suurimmat erot näiden kahden koetyypin välillä?
8. Mitä sähköisessä kokeessa voi tehdä, mitä perinteisessä haastattelumuotoisessa kokeessa ei voi tehdä ja päinvastoin?

Teema 2: *suullinen kielitaito ja sen testaaminen*

1. Kuinka tärkeänä pidätte vieraan kielen puhumisen taidon opettamista?
2. Entä sen testaamista?
3. Tulisiko sen olla pakollista vai vapaaehtoista?
4. Miten suullista kielitaitoa voisi paremmin harjoitella tunnilla ja/tai sen ulkopuolella? Minkälaisten harjoitusten/tehtävätyyppien kautta esim.?

Teema 3: *sähköiset kokeet*

1. Kaisa Vähähyyppä ylioppilastutkintolautakunnan pääsihteeri on sanonut, että ”[s]uunta on se, mitä nuoret toivovat”. Mitä mieltä olette digitalisaatiosta ja sähköisistä kokeista? Onko tämä toivomanne suunta?
2. Mitä mieltä olette siitä, että tietokone voisi automaattisesti arvioida opiskelijoiden suorituksia? Onko se teidän mielestänne reilua?

Haluatteko vielä lisätä jotakin?

Appendix 7 – *Transcription conventions used*

Transcription conventions used	
(.)	Pauses lasting less than one second
(2), (3), (4), (n)…	Pauses lasting over 2 seconds; the number of seconds is indicated in parentheses
#speech#	Code-switching
//	Phonemic transcription
(speech)	Indecipherable or unclear speech
°speech°	Noticeably quieter speech
SPEECH	Louder and/or stressed intonation
[speech]	Overlapping speech
{speech}	Transcriber's comments and remarks
@	Laughter
.t	Alveolar suction click
spee-	False starts
spee:ch	Prolonging of the prior sound or syllable
<speech>	Words or phrases spoken more slowly than surrounding discourse
>speech<	Words or phrases spoken more quickly than surrounding discourse
.h / .h:	Inhalations; longer inhalations are depicted by .h:
h / h:	Exhalations; longer exhalations are denoted by h:
speech	Non-existent lexeme
[…]	Omitted speech in the interviews

Appendix 8 – *Hesitation phenomena per marker in the two tests*

	Hesitation					
	Face-to-face				*Total*	*Total relative to total number of words uttered*
	*Fs.**	*Rep.*	*Ref.*	*Repl.*		
Informant 1	10	3	8	–	**21**	0.037702
Informant 2	41	8	21	–	**70**	0.067633
Informant 3	20	1	12	–	**33**	0.070362
Informant 4	16	1	9	1	**27**	0.035294
Informant 5	15	1	7	1	**24**	0.032967
Informant 6	53	10	30	1	**94**	0.108796
Informant 7	20	6	7	–	**33**	0.054726
Informant 8	42	5	14	1	**62**	0.057461
Informant 9	25	34	6	2	**67**	0.051223
Informant 10	49	28	18	–	**95**	0.045894
Informant 11	19	6	11	2	**38**	0.043728
Informant 12	43	18	22	2	**85**	0.124451
Informant 13	23	9	11	1	**44**	0.101617
Informant 14	21	1	12	1	**35**	0.062837
Informant 15	19	3	6	2	**30**	0.049587

Table 7 – *Hesitation phenomena in the face-to-face test*

* False starts: Fs., Repetitions: Rep., Reformulations: Ref. and Replacements: Repl.

	Hesitation					
	Computer-based				*Total*	*Total relative to total number of words uttered*
	*Fs.**	*Rep.*	*Ref.*	*Repl.*		
Informant 1	9	1	7	1	**18**	0.047619
Informant 2	21	2	14	–	**37**	0.040615
Informant 3	25	5	18	1	**49**	0.089253
Informant 4	18	6	11	–	**35**	0.049716
Informant 5	13	2	6	1	**22**	0.035541
Informant 6	24	5	20	1	**50**	0.077519
Informant 7	13	2	9	–	**24**	0.045283
Informant 8	73	12	44	2	**131**	0.087625
Informant 9	24	3	20	2	**49**	0.037809
Informant 10	35	25	18	–	**78**	0.042810
Informant 11	20	6	14	–	**40**	0.053192
Informant 12	53	19	33	–	**105**	0.104686
Informant 13	17	5	10	–	**32**	0.083117
Informant 14	24	1	11	–	**36**	0.057234
Informant 15	33	6	25	1	**65**	0.098336

Table 8 – *Hesitation phenomena in the computer-based test*

* False starts: Fs., Repetitions: Rep., Reformulations: Ref. and Replacements: Repl.