UNIVERSITY OF HELSINKI

Assessing EFL Writing

in the Finnish Matriculation Examination

Jenni Tukiainen
Master's Thesis
English Philology
Department of Modern Languages
University of Helsinki
February 2016

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta: Humanistinen tiedekunta	Laitos: Nykykielten laitos	
Tekijä: Jenni Tukiainen		
Työn nimi: Assessing EFL Writing in the Finnish Matriculation Examination		
Oppiaine: Englantilainen filologia		
Työn laji: Pro gradu	Aika: helmikuu 2016	Sivumäärä: 71 sivua + liitteet

Tiivistelmä

Tutkielma käsittelee pitkän englannin ylioppilaskokeen ainekirjoitusten arvostelua. Tarkastelun kohteena ovat arviointikriteerit, joiden perusteella kirjoitelmat pisteytetään sekä virallisen ohjeistuksen että käytännön tasolla. Lisäksi tarkoituksena on selvittää, millaisia tekstejä kokelailta vaaditaan pitkän englannin kokeessa sekä millaisia kirjoitelman arviointiin liittyviä kysymyksiä ja huolia kokelaat tuovat ilmi Ylen Abitreenit-ohjelmassa, jossa kokeiden arvioijat kommentoivat ja vastaavat kokeeseen liittyviin kysymyksiin suorissa lähetyksissä heti kokeiden jälkeen.

Aineisto kattaa koekerrat syksystä 2003 syksyyn 2013 eli 21 koetta, joista kussakin on neljä vaihtoehtoista tehtävänantoa. Abitreenit-ohjelman taltiointien lisäksi aineistoon kuuluu viisi arvioitua kirjoitelmaa jokaisesta kokeesta eli yhteensä 105 kirjoitelmaa, ylioppilastutkintolautakunnan kokouspöytäkirjoja sekä koekohtaiset määräykset ja ohjeet vuosilta 2002, 2004, 2005, 2007 ja 2011.

Analyysissä käytettävät metodit ovat pääosin kvalitatiivisia eli deskriptiivisiä ja komparatiivisia. Jonkin verran mukana on myös kvantitatiivista analyysiä.

Analyysin perusteella kirjoitelmien arviointi nojaa määräysten ja ohjeiden asettamiin kriteereihin, jotka sisältävät paljon epämääräisiä kuvauksia sekä tulkinnanvaraisuutta. Esimerkiksi kaikille mahdollisille pisteille ei ole määritelty erillisiä kuvauksia. Esseisiin tehtyjen merkintöjen sekä arviointiohjeiden välillä ei ole paljoakaan yhteyttä, sillä merkinnät koskevat lähes yksinomaan kielivirheitä, kun taas ohjeissa kuvataan taitotasoja myös mm. kommunikatiivisuuden ja sisällön osalta. Arviointi näyttää täten perustuvan pitkälti arvioijan tulkintaan ja sisäistettyyn käsitykseen tietyn tasoisesta kirjoitelmasta. Tätä päätelmää tukevat myös kokousten pöytäkirjat, joiden perusteella kokouksissa käydään läpi mm. esimerkkitapauksia tietyn pistemäärän tasoisista kirjoitelmista. Aineiston perusteella arvioinnissa on kuitenkin nähtävissä jonkin verran epäjohdonmukaisuutta.

Tehtävänannoissa pyydetään kirjoittamaan tekstejä, joista useimmat vaativat akateemisia kirjoitustaitoja eli argumentaatiota ja ilmiöiden kuvaamista. Myös retoriset taidot painottuvat erityisesti puheissa. Loput tehtävistä keskittyvät faktatiedon esittämiseen tai menneiden tapahtumien kerrontaan. Esimerkiksi luovaa kirjoittamista ei edellytetä yhdessäkään tehtävänannossa.

Kokelaiden kysymykset ja huolenaiheet liittyvät hyvin konkreettisiin ja usein yksityiskohtaisiin kieli- ja muotoseikkoihin, jotka mahdollisesti johtavat pistevähennyksiin. Kokeiden arvioijat eivät pysty antamaan yksiselitteisiä vastauksia suurimpaan osaan kysymyksistä, ellei vastaus löydy suoraan ohjeista ja määräyksistä (esim. sanamäärärajoitukset).

Tutkielman tulosten valossa näyttää siltä, että kirjoitelmien arviointi pitkän englannin ylioppilaskokeessa perustuu suurelta osin arvioijan intuitioon, joka on kehittynyt niin tarkaksi, että sen voitaneen olettaa takaavan arvioinnin riittävän luotettavuuden. Monet tehtävänantoihin ja arviointikriteereihin liittyvät seikat näyttävät kuitenkin vuodesta toiseen aiheuttavan epävarmuutta sekä kokelaissa että arvioijissa.

Avainsanat: vieraan kielen arviointi, englanti vieraana kielenä, kirjallisten taitojen arviointi, pitkän vieraan kielen ylioppilaskoe, vieraan kielen arviointi lukiossa, arviointikriteerit, high stakes –kokeet, arviointiasteikot, arvioinnin luotettavuus, kielitaidon arviointi, kirjoitelmien arviointi

Säilytyspaikka: Helsingin yliopiston Keskustakampuksen kirjasto

Muita tietoja

Table of Contents

1 Introduction

Assessment is one of the most essential parts of foreign language teaching and learning in our society. It should be safe to say that everyone who has learned or studied a language at an institute such as school has come across assessment in one form or another - most likely in a variety of forms. The topic of this thesis is the assessment of the EFL (English as a foreign language) writing skills in the national matriculation examination in the Finnish upper secondary school. The English exam consists of four separate parts: listening comprehension, reading comprehension, structure and vocabulary, and written production. This study focuses on the assessment of the written production part of the advanced level English matriculation exam.

While, as Norris (2008) states, assessment has established its position in all formal education as a way of monitoring the learning outcomes, and teachers and administrators are increasingly expected to understand and adopt the principles of assessment, it is not self-evident whether one form of assessment is better than the other or whether any kind of assessment is necessary at all. The emphases on evaluation have varied through time (history), space (institutes), and mindset (ideologies), as is also explained by Weir (2005). According to Norris (2008) assessment can be considered "good" or "appropriate" depending on, for example, its purpose and context. Evaluation indeed varies in form, function and significance, and it may be concentrated on a very narrow or a very broad part of language ability. It can be considered anything between informal feedback to formal numeral grading. The purpose, means and the criteria of the assessment need to be clear to both the person being evaluated and the person conducting the assessment. Furthermore, it is necessary to ensure the validity of the assessment, and thus assessment itself needs to be evaluated (Norris, 2008).

The aim of this study is to analyse how written production is evaluated and what are the assessment criteria. In the light of the data at hand, I will also look at questions and concerns raised by the candidates who take the English matriculation exam. Although the purpose is not to evaluate but rather to describe the assessment, the

results of this study could potentially be used for evaluating the assessment. The research questions are:

1. What are the raters' criteria for evaluating the written compositions in the advanced English matriculation exams in the Finnish upper secondary school?
2. What kinds of texts are the candidates expected to write in the advanced English matriculation exams in the Finnish upper secondary school?
3. What concerns do the candidates have with regard to the assessment of the written compositions in the advanced English matriculation exams in the Finnish upper secondary school?

In the future I may be teaching English as a foreign language in the upper secondary school in Finland. It is therefore important for me personally, as for any upper secondary school teacher, to have a comprehensive understanding of the nature of the matriculation examination. The results of the matriculation examination are an important part of the three-year studies to an upper secondary school student in respect of her/his future study options, because many institutes of higher education give emphasis to the success in the matriculation examination.

It is important for a teacher to be able to guide the students so that they have a realistic idea of the nature of the matriculation exam, given the significance of the evaluation and results. During one of my teaching practices I came across many questions when assessing written compositions of a group of upper secondary school students. The written compositions, or essays, were part of a final course exam, but their ultimate purpose was to train the candidates for the upcoming matriculation exam, hence I was using the matriculation examination rating scale and criteria provided in the exam guidelines. Even with the drawn-up scale it was quite challenging to evaluate the essays and give a numeric grade to each of them, and although the grades ended up being quite well in line with the judgement of my mentor teacher, it felt as if most of the assessment was based on intuition. In this study I start approaching the questions related to the assessment of the essays from the points of view emerging from the guidelines and rating scales, but the other data add crucial information regarding the assessment process in practice.

A lot of research on language assessment is based on quantitative methods and statistics, for example when studying the correlation between time spent on learning and learning outcomes (Leung, 2012). Qualitative methods offer another means of studying language assessment. Leung (2012) gives an example of research on evaluating written assignments: how to analyse a given grade or score and the factors behind the result of the evaluation, such as interpretation of rubrics and content. In this study I mainly apply qualitative methods, but also quantitative perspectives are included.

Although the assessment of foreign language proficiency, including writing skills, has been studied widely, research on the matriculation examination of foreign languages seems to remain scant, and I have not come across a single study on the writing task in the English matriculation exam. In addition to other data such as official guidelines and a commentary programme, this study includes material of very limited access yet of high significance, i.e. authentic evaluated essays from the matriculation examination. Therefore I hope this study will shed some new light not only on the broad concept of assessment and grading of writing skills but also on the practical and concrete issues that teachers and other educational experts deal with as a considerable part of their work.

This thesis is structured as follows: chapter 2 first provides an overview of the theoretical concepts relevant to the topic of assessing EFL writing skills in the matriculation examination and then briefly introduces general upper secondary education and the matriculation examination in Finland. Also, some previous studies on topics related to this thesis are introduced at the end of chapter 2. The data and methods used in this study are described in chapter 3. Chapter 4 comprises the analyses of the data from several different perspectives. In chapter 5 the results of the analysis are summarized and the study as a whole discussed and evaluated. Conclusions based on the results and suggestions for further research topics are presented in chapter 6.

2 Background

This chapter focuses on some key definitions and practices related to foreign language assessment and evaluation of writing skills in particular. A lot of attention is given to rating scales, which hold a salient position in the assessment of the written production task in the English matriculation examination. Also the perspective of objectivity and reliability is given emphasis with respect to the concept of high-stakes tests. After the overview of the theoretical concepts, general upper secondary education in Finland is introduced and the matriculation examination explained in detail. At the end of this chapter, some previous research on EFL/ESL writing assessment is introduced and reviewed.

2.1 Foreign language assessment

In this study, I use the term "assessment" the same way as Taras (2005), i.e. as a judgement of students' performance with regard to the set standards, criteria and learning goals, which results in a comparative or numerical rating. Moreover, I use the term "evaluation" to refer to the process of assessment (Taras, 2005). I use the term "rating" synonymously with the term "assessment" when referring to the assessment done by raters.

Assessment can either be formative or summative. According to Harlen and James' (1997) definition, formative assessment means that the students' performance is evaluated based on what stage they are with respect to a specific skill or content. Formative assessment is typically based on continuous observation in the classroom. Summative assessment, on the other hand, is criterion- or norm-referenced, and the learning outcomes are judged based on the content and goals of, for example, a specific course (Harlen and James, 1997). As suggested by Taras (2005), all assessment will always have summative features. When thinking about summative assessment, testing is one of the first concepts that come to mind. Meyer (2009) argues that the importance and benefits of formative assessment are often neglected and that formative assessment has a crucial role in learners' development. However, I will here focus on testing as a means of assessing language skills and not discuss

other means, such as portfolios or peer assessment, because the focus of this study is on assessing writing skills through a language test.

Tests are often related to summative assessment because they are usually based on the content of a specific syllabus or a topic area of a language. McNamara (2000) distinguishes between two types of language tests regarding the testing method: "paper-and-pencil tests", which often have a fixed response format such as multiple choice, and performance tests, which demand a sample of communicative performance. The performance tests usually measure oral and written language skills (McNamara, 2000; Turner, 2012) and they have become increasingly common in language assessment (Turner, 2012). The test types vary also depending on the purpose of the test. Achievement tests measure the learning outcome in the frames of the set goals of a syllabus, whereas proficiency tests measure the readiness to face real-life needs for language use (McNamara, 2000; Knoch, 2009). Knoch (2009) adds two other test types: a placement test determines the level of the test taker for the purpose of e.g. selecting a suitable language course, and a diagnostic test focuses on identifying the test taker's strengths and weaknesses in some area of language proficiency.

2.2 Assessment of writing skills

When designing a test for assessing written production skills, it is necessary to first define certain properties of the task in order to ensure the validity of the test. As McNamara (2000) states, matters such as the complexity of the task, restrictions concerning the topic, the number and length of texts to be produced, how much support to provide in the rubric and in what time the task should be completed need to be considered.

The themes and topics in the writing task prompts play a significant role in the assessment of foreign language writing skills. Personal themes are easy for the examinee to identify with and they enable the examinee to draw material from their own experiences, as stated by Kranert (2013). When tasks require content knowledge

it will affect the way the task is performed (Shaw and Weir, 2007). There is a risk that some candidates lack the sufficient background information, for example, when candidates from different cultural backgrounds take the same writing task (Kranert, 2013). Kranert (2013) points out that the writing tasks which are part of an exam and based on a curriculum do not measure only writing skills but also knowledge on themes handled in the language courses. According to empirical research by Papajohn (1999 cited in Shaw and Weir, 2007), the topic of the writing task has a significant effect on the score.

The problem of inequality with regard to the required content knowledge could, perhaps, be eased by providing the test takers background reading or other input. A study by Lewkowicz (1997 cited in Shaw and Weir, 2007) indicates, however, that background reading provided in the task or the amount of support given in the task prompt may have a negative influence on the examinee's performance. According to the study, examinees who are given background reading produce less creative texts which contain a lot of ideas and linguistic material drawn directly from the input text (Shaw and Weir, 2007). From the assessment point of view, it is difficult to judge a performance which leans on someone else's production, at least unless the line for borrowing has been drawn in advance and made clear to the examinees (Shaw and Weir, 2007).

In addition to the task properties the test developers need to, of course, define the criteria for assessment and scoring. The criteria are based on the traits which make language proficiency. However, there is no consensus among researchers on the best definition for language proficiency; it may be seen as general competence, field-specific competence or task-based performance (Shin, 2013). Furthermore, the criteria depend on how the test developers understand the essence of the language use, and assessment of language proficiency is always subject to interpretation (McNamara, 2000). McNamara (2000) notes that assessment usually covers different aspects such as fluency, accuracy, organization and sociocultural appropriateness. According to Leung (2012), research on assessment of writing has specified the following features and factors teachers consider while evaluating essays: skills, creativity, process, genre, effort and achievement, EALness (English as a

second/additional language), criteria, content and register. The different perspectives on language competence can either be evaluated as a whole (holistic assessment) or as separate categories (analytic assessment), which may also be combined into an overall score (McNamara, 2000; Shaw and Weir, 2007; Turner, 2012). From the well-known standardized English language proficiency tests TOEFL represents holistic assessment whereas IELTS represents analytic assessment, as stated by Turner (2012). While a combination of different aspects of language proficiency tend to be taken into account in either case, as McNamara (2000) concludes the communicative perspective has become more and more emphasized in both language teaching and assessment.

The assessment of writing skills is most often based on rating scales, in which the criteria for judging the performance is determined (McNamara, 2000; Knoch, 2009; Turner, 2012). Turner (2012) emphasizes that the scales are designed to improve the reliability and simplify the rating. The official guidelines for the assessment of the written production part of the foreign language matriculation exam also provide rating scales. With standardized tests such as the matriculation examination, the assessment is done by raters who are familiar with the rating scales, but it could also be done as self-evaluation. For example, the Common European Framework of Reference for Languages (CEFR) by the European Council (2001), offers rating scales, or proficiency scales, for self-evaluation, although the scales are sometimes applied to rater-mediated assessment as well.

Rating scales in language assessment typically comprise different verbally described levels of language competence, which are connected to a certain numeric or descriptive grade or score. These levels are organized hierarchically in the way that each level needs to be acquired before moving to the next one (Turner, 2012). According to McNamara (2000), a typical rating scale has between three and nine levels which are organized from the highest skills to the lowest. Knoch (2009) states that rating scales of large-scale tests often contain between six and nine levels. For example, the rating scales for the production task in the English matriculation examination have eight level descriptions. McNamara (2000) argues that the descriptions of the levels should ideally be independent from each other, but in

reality the descriptions are often designed in a comparative manner. Designing the number and descriptions of the levels requires careful and thorough planning, e.g. where to set the boundaries and how close the descriptions are to each other, so that the scale would provide useful and relevant information (McNamara, 2000; Knoch, 2009; Turner, 2012). Knoch (2009) notes that the number of possible levels in a rating scale needs to be limited to the raters' capability of making distinctions between them. Other points which need to be considered are the purpose, the type (e.g. holistic or analytic) and the validity of the scale, the emphases in the criteria and the reporting of the scores (Knoch, 2009).

Another question regarding rating scales is whether they are designed for general use or a specific writing task. In addition to holistic and analytic assessment Shaw and Weir (2007), Knoch (2009) and Turner (2012) introduce primary trait scoring as the third type of assessment. This assessment type adds one principal feature of interest to the holistic assessment and is designed for a particular writing task (Shaw and Weir, 2007; Knoch, 2009; Turner, 2012). Knoch (2009) and Turner (2012) also talk about multitrait scales which focus on more than one trait. However, primary trait assessment is considered less convenient than holistic and analytic approaches due to its restrictions and complexity (Shaw and Weir, 2007). In the case of the matriculation examination guidelines, the rating scales are designed to be applied not only to the production tasks of a certain exam but to all the exams for which the guidelines are valid, which is made evident in the guidelines.

2.3 Objectivity in assessing high-stakes language tests

A language test may have a massive impact on an individual's life concerning e.g. education, employment or the place of residence, as McNamara (2000) states. The purpose of foreign language assessment varies from solely checking whether the test taker has acquired a certain area of language knowledge (achievement test) or determining their level for selecting a suitable language course (placement test), to using the test results in decision-making when choosing applicants, for example, for a job or school, or granting citizenship. Those language tests which have a major

impact on an individual's life are called high-stakes tests (Kunnan, 2012). As the number of candidates taking a high-stakes test at once can be large, Kunnan (2012) notes that high-stakes tests are often standardized to simplify administration, scoring and reporting processes. It is hardly surprising that high-stakes test takers are under high pressure due to the significance of the results. Therefore, as Weir (2005) argues, it is of particular importance for the test developers to plan the content, conditions and assessment of the test accordingly.

As McNamara (2000) states, the rating process is always subjective, which makes it crucial that the raters be trained and qualified for the assignment. However, McNamara (2000) argues that even if the rating criteria and classifications are made unambiguous and clear to the trained raters who fulfil their task accurately, the rating will nevertheless remain subjective, as is proven in the comparison of ratings done by different raters. Furthermore, McNamara (2000) suggests that even ratings done by the same rater may sometimes prove inconsistent. According to Shaw and Weir (2007), one problem with the interpretation of rating scales is that they include ambiguous wordings such as "simple" when, for example, referring to linguistic features such as vocabulary or structures. Despite the endeavor to improve the validity of the rating scales (Shaw and Weir, 2007; Knoch, 2009; Turner, 2012), Shaw and Weir (2007) note that it is arguable whether written criteria can be all-encompassing in defining a certain level of language competence, and they talk about internalized representation, which raters develop as a result of experience. This adds to the challenges of the reliability and objectivity of the assessment. But, especially when communicative skills are in the centre of the assessment, is it even possible to remove this human inaccuracy that seems inevitable?

Despite the acknowledgment of the limitations of rater objectivity, which will always occur due to e.g. different tendencies in severity and emphases of the raters, McNamara (2000), Shaw and Weir (2007) and Knoch (2009) talk about rater training, which is important in terms of fairness and consistency of the rating. In practice, rater training often happens in moderating meetings, in which comparison is done and inconsistencies in rating discussed with the aim to achieve general agreement on the interpretation of the rating scales (McNamara, 2000). For example,

the matriculation examination board holds moderating meetings to discuss issues related to the assessment of the matriculation exams. McNamara (2000) also argues that psychological pressure encourages raters to aim for rating results which are in line with other raters.

The matriculation examination can be considered high-stakes testing, because the results have a considerable influence on higher education possibilities. For example, many universities give only part of the application score based on a separate entrance exam and the rest of the score comes directly from the matriculation examination results. Some university departments also accept applicants directly based solely on successful performance in specific subjects in the matriculation examination. The significance of the matriculation examination results is one of the main motivations for this study.

2.4 EFL and upper secondary school education in Finland

In the Finnish matriculation examination a candidate has to take four obligatory exams, one of which is a foreign language exam. If s/he wishes, s/he may also take optional exams (Matriculation Examination, Board 2011). In addition to the national languages, Finnish and Swedish, there are several foreign languages taught in the upper secondary school in Finland and tested in the matriculation examination. Some of these languages, such as Italian and Portuguese, are only taught in very few schools; English, on the other hand, is the most commonly learned foreign language and it is offered, if not in all, at least in most schools. In autumn 2015, for example, 16 247 of the 36 802 candidates took the advanced level and 407 candidates took the basic level English matriculation exam (Matriculation Examination Board, 2015). The number would, of course, be much higher if all the exams had to be taken at once - but the exams are given every semester and the candidate is allowed to take exams in three consecutive semesters. In comparison, in Swedish the number of the examinees in autumn 2015 was 588 (advanced level) and 6 829 (intermediate level), in Finnish as a foreign language 1261 (advanced level) and 306 (intermediate level),

and in German 124 (advanced level) and 396 (basic level), according to the statistics from the Matriculation Examination Board (2015).

In the advanced level English syllabus, the upper secondary school offers six compulsory courses and a minimum of two specialization courses. In addition, schools may offer applied courses (Finnish National Board of Education, 2003). Prior to taking the English matriculation exam the candidates need to have taken all the compulsory English courses (Matriculation Examination Board, 2011). Students are, however, advised to take at least the two specialization courses in addition to the obligatory courses before taking the exam.

2.5 The advanced English matriculation examination

The purpose of the language exams is to find out whether the candidates have acquired the knowledge, skills and maturity presented in the national curriculum for the general upper secondary education, as is stated by the Finnish National Board of Education (2011). The language exams are based on the curriculum and the content of both obligatory and optional courses offered in the upper secondary schools (The Finnish National Board of Education, 2011). The exams are also designed to reflect the CEFR, which emphasizes a communicative perspective in language competence. According to the CEFR, language skills include knowledge of the language and its usage in situations defined in the curriculum. According to the Matriculation Examination Board (2011), the language tests measure both reception skills and communicative production skills, and the different parts of the exam are designed to cover different types of communicative situations as comprehensively as possible.

In the matriculation examination, there are two levels in the English exam. The advanced level exam is based on the syllabus of the so-called “A language”, which means that the learning of the particular foreign language started in elementary school (The Matriculation Examination Board, 2011). For the sake of clarity, I use the term “curriculum” when I refer to the document which defines all the content, objectives and requirements for general upper secondary education and the term

"syllabus" when I refer to the extent, content and objectives of a certain subject included in the curriculum.

The exams are planned and implemented by the Matriculation Examination Board, which is chosen by the Ministry of Education every three years (Finlex, 1998) and which consists of a chairperson, two deputy chairpersons and a necessary number of other members (Finlex, 2005). The exam papers are first examined, assessed and scored by the subject teachers in schools before they are sent to the Matriculation Examination Board, where a smaller group of specialized raters, called censors, go through each exam and determine the final grades. According to Finlex (2005), the censors are persons who know the subject and the educational field well. In the analysis I use the term "rater" to refer to the whole group of teachers and censors who assess the matriculation exams. In the background section the term "rater" is not specified and generally refers to a person responsible for the rating of an exam or a performance.

There are seven possible grades, which from the highest to the lowest are: laudatur, eximia cum laude approbatur, magna cum laude approbatur, cum laude approbatur, lubenter approbatur, approbatur and improbatur (Finlex, 2005). The grades for each exam are given based on normal distribution, which means that the score requirements for each grade are defined separately for each exam. For example, five percent of the examinees always get the highest grade (laudatur) and five percent of the examinees do not pass the exam (improbatur), whereas, as is characteristic to normal distribution (McNamara, 2000), most grades will be close to the average. This kind of assessment is called norm-referenced measurement, in which the examinee's performance is evaluated in comparison with the other examinees' performance instead of fixed criteria (McNamara, 2000).

The structure of the advanced English exam has remained the same for a long time and it consists of four clearly separated parts: listening comprehension, reading comprehension, structure and vocabulary, and written production. The exam is taken in two parts on two separate days: the listening comprehension test is taken on one day and the written part, which includes the three other parts, is taken on another day. The examinees have up to six hours to finish the written exam. They can proceed as they choose, i.e. they need not follow the order of the tasks. The

production task is placed at the end of the test, but the examinees are advised to reserve enough time for this part because it encompasses almost half of the maximum score of the written exam (99/209) and one third of the maximum score of the whole exam (99/299). Thus, in the evaluation, the production task weighs more than any other single task in the exam.

2.6 Educational goals

The matriculation examination, as well as the general upper secondary education as a whole, is based on the national curriculum written by the Finnish National Board of Education. The exams which are included in the data of this study are based on the 1994 curriculum until spring 2008. From autumn 2008 onwards the exams are based on the 2003 curriculum. This was confirmed to me by a member of the National Examination Board.

The 1994 curriculum gives both general and level-specific learning objectives for foreign language learning but it does not set goals for individual languages. The general objectives of foreign languages include enhancing common knowledge and worldview, the ability to communicate in working life, understanding cultures and one's own cultural identity, readiness to work internationally, motivation to use the language and self-assessment skills. The objectives of written skills for the advanced level syllabus are variation in the use of vocabulary, idioms and structures, writing fluent text with the help of aiding tools if necessary, the ability to write a summary and applying a way of communication that is typical of the target language and culture (Finnish National Board of Education, 1994).

The 2003 curriculum for foreign languages generally emphasizes intercultural communication skills and cultural awareness, understanding and appreciation, especially in the European community. It also lists other objectives, such as ability to communicate in a way that is typical of the target language and culture and self-assessment and awareness of one's own learning strategies. The 2003 curriculum defines proficiency level objectives for English and other foreign languages. The skill level scale in the curriculum is a Finnish application of the skill levels presented in CEFR (The Matriculation Examination Board, 2011), which introduces

proficiency levels from A1 (basic skills) to C2 (native-like skills). In 1994 the CEFR proficiency scales did not yet exist and thus were not applied in the curriculum. The skills are divided into four categories: listening, reading, writing and speaking. According to the scale presented in the 2003 curriculum, the target level in advanced English is B2.1 in all of the four categories. The definition for the level B2.1 in writing is quoted in Appendix 3. In short, the candidate is expected to be able to write detailed text on personal and abstract topics, process, express and organize ideas, use broad vocabulary and have a fairly good command of orthography, grammar and punctuation. However, according to the description, complex structures and variation in expression and style cause problems at this level of proficiency.

2.7 Research on assessing EFL/ESL writing skills

Although I have not found earlier research on the particular questions raised in this thesis or an identical research design, the assessment of foreign language writing skills has been studied to a large extent. Here I briefly introduce five international studies which aim to answer question regarding assessment of EFL or ESL (English as a second language) writing skills.

Neumann's (2014) case study focuses on teachers' assessment of grammar in academic ESL writing. Neumann (2014) aims to specify the indicators of grammatical proficiency teachers take into consideration when assessing students' academic writing skills. Her purpose is also to find out students' perceptions about the criteria for the assessment of grammar in writing and what they perceive to be the influence of these criteria to their written production and learning. Neumann (2014) approaches the research questions with a mixed methods design, which includes both quantitative and qualitative methods. Quantitative approach was applied in the form of statistical techniques in analysing the students' essay exams to find out the teachers' assessment criteria regarding grammatical skills in writing. Qualitative approach was used in the form of student questionnaires and interviews with students and teachers to illuminate the context of the assessment and to answer questions related to the assessment criteria as well as students' ideas of the expectations concerning grammatical ability. In Neumann's (2014) study factors on the sentence

level accuracy emerge as the primary criterion for the assessment. Regarding the other research questions, the study shows that the students were aware of their teachers' focus in assessing the essays and that they adapted their writing according to the teacher's expectations.

Lee (2011) focuses her study concerning EFL writing assessment on another perspective, namely formative assessment as opposed to summative assessment, the latter of which, as she argues, does not much promote foreign language learning. The aim of the study is to provide evidence for the benefits of formative assessment of written proficiency. The study investigates an EFL teacher, the formative practices she uses in assessing her students' writing skills and their impact on the students' beliefs and attitudes towards EFL writing.

A study which concentrates on the assessment of EFL writing skills in upper secondary school was conducted in Göteborg by Oscarson (2009). However, the focus of this study is not on teachers' or other raters' assessment principles but students' self-assessment. The aim of Oscarson's (2009) study is to find ways to enhance students' lifelong foreign language learning skills. This was done by looking at students' perceptions of their writing skills with regard to the syllabus and by finding out whether self-assessment influences these perceptions. The data consists of students' self-assessment material on two graded writing tasks. The results of the study suggest that students have no trouble with assessing their general writing skills at a group level, but the ability to assess specific writing skills depends on the amount of practice they have had on self-assessment.

The expectations regarding foreign language writing skills were studied in the dissertation by Mo (2015). In this study, genre requirements in writing tasks are analysed by investigating the syntax of the task prompts and by applying content analysis in examining the rubrics. The study compares the differences between writing tasks on the state and the national levels and finds that task instructions on the national level are more specific than those on the state level. Mo (2015) also draws attention to problems regarding writing assessment at the state level, such as task prompt ambiguity, implicit genre expectations and incongruity between learning expectations and writing assessment.

Students' perceptions with respect to assessment of writing skills have also been studied, as was done by Montgomery and Baker (2007). The study explores the correlation between teachers' self-reflection on the amount of feedback they give on ESL written compositions from different aspects such as organization, vocabulary or ideas and content. Furthermore, the aim is to compare students' perceptions to teachers' perceptions on the amount of feedback given. The data consists of teacher and student questionnaires and the feedback on the written compositions. The results show that overall the teachers' and the students' perceptions on the feedback corresponded well to each other. However, according to the results, the students thought they received more feedback than they were given, and the teachers' perceptions on how much they give feedback on each aspect were not fully in line with the reality.

The studies introduced above were selected based on their relevance to the topic and research questions of this thesis. As it is evident, the earlier research differs from this thesis in perspective and research design. In chapter 3.2 below I compare the research methods described in this chapter to the ones I use in this thesis.

3 Data and methods

This chapter describes the data and methods used in this study. The different pieces of data will first be introduced separately, after which the methods are briefly explained and reviewed.

3.1 Data

The data consists of material obtained from the Matriculation Examination Board (MEB) and public online material created by the Finnish Broadcasting Company. The data from the MEB is in written form and the data from the Finnish Broadcasting company is in the form of video recordings. The data includes:

- production tasks in the written parts of the advanced English matriculation exams from autumn 2003 to autumn 2013, a total of 21 exams
- 5 evaluated and graded authentic essays from each exam, i.e. a total of 105 essays
- the official level-specific matriculation exam guidelines and regulations for foreign languages from the years 2002, 2004, 2005, 2007 and 2011
- meeting minutes from the MEB English censor meetings from 13 meetings between autumn 2009 and spring 2013
- the Finnish Broadcasting Company's (Yleisradio) commentary programme "Abitreenit" from spring 2010, autumn 2011, spring 2012, autumn 2012, spring 2013 and autumn 2013

I was granted a research permission from the Matriculation Examination Board to analyse data that is not publicly available. The non-public data includes the production tasks from autumn 2003 to autumn 2007, all the essays, the guidelines from 2002-2007 and the MEB meeting minutes. As I have agreed with the MEB, I will handle all the confidential data, i.e. student essays anonymously and will not reveal any identities or name schools.

The exams i.e. the production tasks are available online from spring 2008 onwards, as well as the commentary programme episodes from the exams mentioned above (Yleisradio, 2010; 2011; 2012a; 2012b; 2013a; 2013b). The 2011 guidelines are still in use and available online at the time of this study. The 2003 curriculum is also available online (Finnish National Board of Education 2003). The 1994 curriculum is publicly available in print.

The selection of the data was based on accessibility and relevance. I use all the material that was provided by the MEB. There are gaps within certain pieces of the data: the meeting minutes have not been archived systematically and video recordings of the Abitreenit episodes are not available from all exams. As the data descriptions below show, this study focuses on the censors' perspective, as all the comments and remarks apart from essay markings come from censors. Additional data, such as questionnaires, interviews with raters and candidates or data from Suomen englanninopettajat ry (the association of English teachers in Finland) could have been included, but due to the limitations of this thesis and the quality of the already included data, I have decided that there is no major need to look for any additional material. For future research, however, the above mentioned data could prove interesting.

3.1.1 The production task prompts

The data in this study includes exams from ten years: from autumn 2003 to autumn 2013. Because the matriculation examination takes place twice a year, in spring and in autumn, tasks from 21 exams are under examination. Each production task has four separate task prompts/rubrics, of which the examinee chooses one. This means that the 21 production tasks include a total of 84 task prompts. All the task prompts are listed in Table A in Appendix 2. The exams are listed under "TIME" with abbreviations such as "A12" (autumn 2012) or "S08" (spring 2008). The titles of the task prompts are listed under "TITLE".

The reason for selecting this time period is that it includes the last ten exams from which, at the time of the data collection of this study, it was possible to receive essays as data for research. The amount of exams also provides enough data for a thesis of this scale.

3.1.2 Essays

A sample of 105 authentic compositions was chosen as data for this study. The selection was done by a person working at the MEB, and the selection was presumably done at random. There are five essays from each exam from autumn 2003 to autumn 2013. The five essays are all written about the same rubric, i.e. they were selected based on which topic the examinee chose in the exam, so that more exact and reliable comparison between different performances can be done. This is the only basis for the selection, and no other factor, such as the score, was taken into account.

The essays are marked by at least two raters: one teacher and one censor. The markings made by different raters are distinguishable by the colour of the pen. Teachers use a red pen, the first censor uses a green pen and possible other censors use other colours. These instructions are written in the guidelines and stated in the censor meetings. In this study I look at all the markings and do not consider the differences between markings made by different raters within an essay.

The essays are confidential material, and I had access to them at the National Archives of Finland for one month and one week in September-October 2015. For research purposes, I have a permission to give examples from the essays in this thesis based on the notes I made in the archives. However, I was not allowed to make copies of the complete essays. Due to these limitations in ethics and accessibility, the extracts in the analysis section are relatively short.

3.1.3 The guidelines and regulations

The guidelines and regulations (*määräykset ja ohjeet*) for the matriculation examination are meant for teachers as well as the censors assessing the exams. They are updated more frequently than the curriculum. However, the contents of the publications have not changed radically in ten years and they have largely remained the same. The publications relevant for this study are from the years 2002, 2004, 2005, 2007 and 2011. Earlier years are irrelevant for the data, there are no other publications in between, and the 2011 version is currently the most recent one. The 2011 guidelines are publicly available online, giving also the students and parents the opportunity to get acquainted with the content. The guidelines cover all foreign

language exams in general, but they provide specific instructions for the advanced level exam. The guidelines provide a lot of detailed practical instructions for administration and implementation. There are also some instructions for the assessment and scoring of essays.

3.1.4 Meeting minutes from MEB English censor meetings

The data includes minutes from the Matriculation Examination Board English censor meetings. The minutes are from the meetings held on 15.9.2009, 23.9.2009, 19.2.2010, 24.3.2010, 17.9.2010, 24.9.2010, 17.2.2012, 23.3.2012, 14.9.2012, 21.9.2012, 19.2.2013, 22.3.2013 and 5.4.2013. All meetings have not been systematically recorded by the MEB. There are at least two meetings for each exam; the first meeting is dedicated to the listening comprehension part and the second meeting is dedicated to the written part of the exam. Therefore, the examples given in the analysis section are from the second meetings. The structure of the meetings is quite fixed: they start with general announcements, after which the exam in question is reviewed section by section.

The meetings are attended by the chairperson and the censors who read and assess the English exams. In each meeting the participants are reminded that it is important for all the censors to be present in the meetings. According to the meeting minutes, the number of English censors varies between 21 and 42. The meeting minutes give the following information about the number and presence of the censors:

15.9.2009: 20 present, 1 absent

23.9.2009: 19 present, 2 absent

19.2.2010: 36 present, 1 absent

24.3.2010: 38 present (no mention about absences)

17.9.2010: 21 present, 3 absent

24.9.2010: 20 present, 4 absent

17.2.2012: 35 present, 4 absent

23.3.2012: 40 present, 1 absent

14.9.2012: 29 present, 1 absent

21.9.2012: 30 present, 0 absent

19.2.2013: 34 present, 5 absent

22.3.2013: 41 present, 1 absent

The number of censors seems to vary not only for each exam but sometimes also for the different parts (listening and written part) of the exams.

3.1.5 The commentary programme

In recent years the Finnish Broadcasting Company, or Yleisradio, has broadcast a live commentary programme called "Abitreenit" shortly after matriculation exams of different subjects, including the advanced English exams. The episodes are subject- and exam-specific, i.e. there are separate episodes for different subjects and exams. The episodes are published online on Yleisradio's website. The episodes from the English exams included in the data that are still available on the website are from spring 2010, autumn 2011, spring 2012, autumn 2012, spring 2013 and autumn 2013, which makes six episodes altogether.

In the programme there are two people representing the MEB and answering the candidates' questions about the exam and explaining what makes a good answer or a good essay. These persons belong to the group of censors. There are also a host and one candidate who took the exam present in the conversation. The host and the guest candidate ask the censors questions of their own or sent by the viewers via a live chat room. Most of the questions come from the candidates who took the exam. The different parts of the exam are gone through in order, so the comments related to the written production task are at the end of the programme. Usually the last 10-15 minutes are reserved for commenting on and answering questions about the written production tasks.

I have transcribed all the conversations which deal with the written production task from all six episodes. The excerpts included in the analysis are my translations from these transcriptions. I have replaced the censors' real names by numbering them. In the six episodes there are three different censors altogether, i.e. Censor 1, Censor 2 and Censor 3. The number always refers to the same person.

3.2 Methods

In this thesis I use qualitative and quantitative methods, i.e. mixed methods, when analyzing the data.

For the written production tasks, I have drawn up a table (see Table A in Appendix 2) where all the 84 task prompts/rubrics from the 21 exams are listed. The instructions, themes and text types vary. For example, in autumn 2008 the following rubrics were given: *1. A speech; 2. Opinion and advice; 3. Hard values, soft values; 4. An interesting event or period in history.* I have listed six different features of the production tasks: text type, context, argumentation, data, theme and support, which will be explained in more detail in the analysis chapter. To analyse the text types I use a categorization of genre analysis presented in Hyland (2004).

In addition to categorization I use comparative methods. I compare the contents of the guidelines from different years with each other, and from a comparative perspective I also look at what information the different pieces of data give with regard to the research questions. Furthermore, a comparative approach is applied to the analysis of the raters' markings in the essays and when analysing the differences between three essays of different grades.

I use qualitative content analysis (Dörnyei, 2007) when analysing the student essays, the MEB censor meeting minutes and the commentary programme. I apply quantitative methods to the analysis of the task prompts and the essays. I have counted the amount of certain text types and context regarding the task prompts. Descriptive information on all the essays is provided in Table B (Appendix 2). This includes information on the score, word count, number of paragraphs, and types of markings. The essay numbering in Table B, which is referred to in the analysis, is done for the purposes of this thesis and is not in any way connected to the examinee numbers.

Comparing the research designs in the studies described in chapter 2.7 and this thesis, many differences yet a lot of resemblance can be seen. For example, both Neumann's (2014) study and mine use mixed methods. However, Neumann has more emphasis on quantitative methods, whereas my methods are primarily qualitative. Neumann bases her quantitative analysis on statistical methods, while my

quantitative methods consist of constructing tables which contain statistical information and simple calculations. Neumann's approach to the analysis of the essays is researcher-centered in that the researchers label the features of interest in the essays. In my analysis I focus chiefly on the raters' markings. However, I also point out parts in the essays which are not marked when they are in conflict with the other markings. The data in Neumann's (2014) and Montgomery and Baker's (2007) studies include a questionnaire, and Neumann's study also includes interviews. The data gathered via questionnaires and interviews can be considered direct and explicit yet subjective and biased. For similar purposes I use authentic documented data (a commentary programme and meeting minutes), which was not created for the needs of this thesis but for the sake of validity and reliability of the exam. Thus it can be considered more objective albeit indirect and not targeted. Similar to the methods used by Mo (2015), I examine the task prompts and rubrics in the matriculation examination written production tasks on the syntactic and content levels.

4 Analysis

In this chapter I analyse the assessment criteria and emphases based on the task prompts, essays, guidelines, censor meetings and the commentary programme. I begin by looking at the descriptions of the scores on a more general level, after which I move on to analysing the assessment from different aspects. First, the aim is to find out how the communicative aspects are taken into account in the rating and designing the tasks. Secondly, I focus on the influence of content (length, structure and topic handling) on the rating. Third, I investigate linguistic features such as the correctness of grammar and spelling and what their role is in the rating. Throughout the analysis I include the candidates' comments and questions emerging from the commentary programme with respect to the three above mentioned aspects. Finally, I compare three essays of different scores with each other to see how the essays reflect the assessment principles in practice.

4.1 Scoring

In all the guidelines from 2002 to 2011 the possible scores to be given are 99, 97, 95, 92, 90, 88, 85, 82, 80, 78, 75, 72, 70, 68, 65, 62, 60, 58, 55, 52, 50, 48, 45, 42, 40, 35, 30, 25, 20, 15, 10, 5 and 0. According to the guidelines, other scores cannot be given. However, two exceptions occur in the data: essay number 18, which got 63 points, and essay number 6, which got 57 points (see Table B in Appendix 2). The reason behind these exceptional scores is point deductions, which are applied at the end of the grading process.

Table B in Appendix 2 shows all the scores marked in each essay (in the "Score" column) - first the one given by the teacher and then the MEB censor(s). Sometimes the essays have scores crossed out when the rater has either written a wrong score or changed her/his mind. The final score is marked in bold letters on the right. Instructions regarding the scoring are given in the guidelines and the censor meetings. Teachers are instructed to write their score on the final page of the exam paper so that it will not influence the censors' assessment (The Matriculation Examination Board, 2011). In the censor meetings the censors are reminded that

essays which are given 99 points by a censor will always be evaluated by a second censor. A second censor may evaluate 20-25 essays from each censor when there is a considerable point gap between the scores given by the teacher and the first censor. To lower the chance of biased or incongruous rating, the essays are always assessed by at least two raters. The teachers may often know the candidates and how they normally perform in English writing tasks, which could consciously or unconsciously influence the rating. This problem is much less likely to occur with the MEB censors, and because the final score is always given by a censor, such bias should not have a major impact on the exam results.

Table 1 shows where in the rating scale provided by the guidelines (according to the scales from years 2007 and 2011) the 105 essays are placed.

Score (out of 99)	Essays
99-90	12
88-80	30
78-70	24
68-60	21
58-50	12
48-40	6
35-20	0
15-0	0
TOTAL	105

Table 1: Distribution of the essay scores (according to the 2007 and 2011 scales).

The levels in the scales in the guidelines from 2002, 2004 and 2005 are called "very good"/*erittäin hyvä* (99-90), "good"/*hyvä* (88-80), "rather good"/*melko hyvä* (78-70), "satisfactory"/*tyydyttävä* (68-60), "sufficient"/*välttävä* (58-50), "weak"/*heikko* (48-35) and "very weak"/*erittäin heikko* (30-20). The scores from 0 to 15 do not have a

verbal equivalent. It is noteworthy that the scales differ slightly between the guidelines of different years: the lower levels in the 2007 and 2011 scales are 48-40 and 35-20. The 2007 and 2011 scales do not have any verbal equivalents for the scores. It should be noted that each of the eight levels in the rating scale is connected to four or five scores, yet the level descriptions offer no clues as to how to distinguish between the 33 different scores.

The data suggests that the assessment of the writing skills in the advanced English matriculation exam is not, in fact, based on the curriculum but rather on the level of students' English skills in Finland. In the spring 2010 Abitreenit episode a viewer comments that the exam was too easy, to which one of the censors responds that the exam was made to correspond the level B2.1 described in the curriculum and that the difficulty level of two of the texts in the spring 2010 exam is even higher than the candidates' expected level of proficiency. Due to the normal distribution of the grades, it is practically impossible to reach to the highest scores with level B2.1 skills. Also the definitions for the highest scores in the guidelines demand more than the objectives in the curriculum.

The rating scales represent analytic rating in the way that they distinguish separate factors which all have their own descriptions for each level, although it is stated clearly that communicativity should be considered as the main factor. However, as only one score can be given for the whole task, the rating can and perhaps should be considered holistic as well. The scales in the exam guidelines from 2007 and 2011 distinguish three factors to be evaluated: 1) communicativity, 2) context and structure and 3) linguistic richness and accuracy. The older guidelines distinguish four factors: 1) readability and language use, 2) expressive competence, 3) handling the topic and 4) linguistic errors. Based on these factors I have divided the next three parts of the analysis into looking at the evaluation from the perspectives of communicativity, content and linguistic features.

4.2 Communicativity

This section is divided into four subsections: text type, contextualization, emphasis on communicative skills and style. The first three subsections focus on what

communicative skills are expected from the candidates, whereas the last subsection provides concrete examples of how communicative features affect the assessment of the essays in practice.

4.2.1 Text types of the essays

When analysing the communicative factors in assessing EFL writing tasks, it is relevant to look at what text types the task prompts demand. I have applied the genre categorization based on Butt, Fahey, Feez, Spinks, and Yallop (2000) and Martin (1989) (reviewed in Hyland, 2004) when analyzing the text types of the writing task prompts. Here I use the term "text type" synonymously with Hyland's term "genre", because later in this thesis I will use the term "genre" for different purposes. I decided to use this categorization model because the categories are applicable for the kinds of texts produced at school as part of practicing writing skills in a foreign language. Table 2 shows the distribution of the different text types in the data based on the information in the column "T.TYPES" in Table A in Appendix 2.

Recount	6	7 %
Procedure	0	0 %
Narrative	0	0 %
Description	30	36 %
Report	6	7 %
Explanation	0	0 %
Exposition	39	46 %
Not specified	3	4 %
Total	84	100 %

Table 2: Text types of the task prompts

"Recount" means a description and reconstruction of past events (Hyland 2004: 29). Examples of recount tasks are, for example, the prompts *A speech* from spring 2009 and *My crazy festival* from spring 2008. The first one instructs the examinee to write a speech to be given at her/his grandparent's 80th birthday. The latter asks the examinee to write about a past or imaginary festival.

The main function of a "description" is to describe an event or a phenomenon, real or imaginary (Hyland 2004: 29). An example of a description is from spring 2013*: When angry, count [to] four; when very angry, swear.* Here the examinee is supposed to write about what makes her/him angry and what s/he does in such a situation. What distinguishes description from recount in this case is that the main idea is not to write about a past event in a story-like form, although it may contain such components, but rather to generalize and describe a phenomenon.

Hyland (2004: 29) defines "report" as a presentation of factual information, including classification and description. The definition as such is perhaps somewhat equivocal, but it is elaborated in the description of its purpose: a report first identifies a problem, then it provides analytical commentary, and finally proposes a solution (Hyland 2004: 33). An example of a report task is from autumn 2011: *Improving safety in traffic*, in which the examinee is given a diagram of car and moped casualties by age and is asked to give suggestions for concrete measures to increase traffic safety.

"Exposition" is an argumentative text which justifies a claim and may weigh different points of view (Hyland 2004: 29). An example of an exposition is *Shopping on Sundays* from spring 2011. The examinee is asked to state an opinion about the opening hours of shops and write about both advantages and disadvantages of having shops open seven days a week.

There are three genres introduced by Hyland (2004) which do not occur in the exam prompt data: procedure, narrative and explanation. "Procedure" is a set of instructions, such as manuals and recipes (Hyland 2004: 29). "Narrative" is a text written in the form of a story, its purpose is to entertain or instruct through reflection, and it includes orientation, complication, evaluation and resolution (Hyland 2004: 29, 33). "Explanation" reports an incident and specifies reasons behind it (Hyland

2004: 29). It differs from reports in that it has a more objective perspective, it does not necessarily problematize a phenomenon, and it does not offer a solution.

Some of the task prompts do not offer or allude to any specific text type. I have marked these cases as "not specified" in Table A (Appendix 2). These task prompts only provide the title and nothing else, leaving it up to the examinee to decide which text type would be the most suitable either for the title or themselves. One of these task prompts is *A lucky escape* from autumn 2012.

As seen in Table 2 above, the most common text types in the task prompts are exposition and description: 82 % of the 84 prompts represent these text types. Recount and report are equally frequent, 7 % of the total number of the task prompts each, and the remaining 4 % of the task prompts are unspecified. These numbers show that there is little variation in the text types, and almost half of the text types introduced are unrepresented. One purpose of the non-specified task prompts may be to leave room for text type variation. The text type analysis suggests that the writing skill that is considered the most important is exposition, i.e. argumentative writing, after which comes description, which emphasizes classification and description skills.

Another observation is that none of the tasks represents creative writing such as a narrative. Other types of creative writing are not included in the list of text types in Hyland (2004). The unspecified task prompts could and probably have, however, inspired compositions of a creative sort. According to the censors commenting on the requirements of the autumn 2012 exam production task in the Abitreenit programme, under the title *A lucky escape* the examinee could write any kind of text from any point of view. One of the censors says that "the sky is the limit" and that one could write almost about anything, be it fantasy or real life, as long as it fits the title. According to him, a scientific text about water purification, for example, would not fit.

4.2.2 Contextualization

Another interesting feature of the task prompts is the contextualization of the text to be produced (see Table A in Appendix 2). Table 3 below reveals that 40 percent of all the task prompts provide a specific forum for the composition; the sense of

authenticity in the writing task is thus highlighted. The rest of the task prompts do not specify a context. I have divided the forums into media and communities. "Media" comprises any kinds of published media, such as in the task prompt *Opinion and advice* from autumn 2008, in which the examinee is advised to write a letter to The New York Times, or in *Calling All Inventors!* from autumn 2013, in which the examinee is asked to write a letter addressed to the website Everyday Edisons. The media provided in the task prompts are authentic although they may not be familiar to all of the examinees.

The other forum in the task prompts I call "community". The texts meant for a community forum are typically speeches written for a group of strangers or friends/relatives. An example of a speech meant for a community is the task prompt *Speech* from autumn 2007, in which the audience is "an international group". The community forum also includes letters addressed to individuals, organizations or companies, as is the case in *Dear X* from spring 2013 and *Could I change this gift?* from autumn 2003.

Forum	Number of task prompts
Media	20
Community	14
Not specified	50
Total	84

Table 3: Context of the task prompts.

There has been an increase in the amount of task prompts which include a context in the form of a media outlet or a community. In Table 4 I have divided the 10-year-period into three groups, each of which includes seven exams. In the earliest period, from autumn 2003 to autumn 2006, only three out of 28 task prompts specify a forum. In the middle period, from spring 2007 to spring 2010, the number has increased to 12 out of 28. In the most recent period, from autumn 2010 to autumn 2013, two thirds of the task prompts include a context.

Task prompt	Context included
Autumn 2010 - autumn 2013	19 / 28
Spring 2007 - spring 2010	12 / 28
Autumn 2003 - autumn 2006	3 / 28

Table 4: The amount of the task prompts with a context included.

4.2.3 Emphasis on communicativity

The assessment guidelines for the advanced level foreign language exam emphasize communicative skills over other features. The more recent guidelines, 2007 and 2011, state that the assessment is to be done primarily based on the criteria described in the column "communicativity" in the rating scale, and the other criteria, i.e. content and structure, and linguistic richness and accuracy, support and complete the assessment. The criteria for communicative proficiency in the written production task are listed in Table 8 in Appendix 1, in which I have translated the content related to communicative features in the rating scale.

As we can see in Table 8, the criteria have become more exact and descriptive. However, they are still very vague and leave much room for interpretation. It is up to the rater's personal opinion whether a text is very clear, clear or fairly clear, what it means to write relatively naturally, or what is the difference between insufficient (puutteellinen) and poor (heikko) language proficiency. The vagueness of the criteria also shows in the difficulty of translating the criteria into English. Hence it is important to find out what other sources say about assessing the communicativity of the compositions.

4.2.4 Style

The guidelines state that one of the aspects to be taken into consideration when assessing the written compositions is the idiomaticity of the language use. The levels for scores 60-99 in the 2007 and 2011 rating scales describe how stylistically

appropriately the language fits in the context. In the earlier guidelines stylistic features are also included.

In the Abitreenit programme, the censors talk about the importance of proper style, especially when writing a speech or when addressing a letter to a person outside the familiar circle. Twelve out of the 84 task prompts ask the examinee to write a speech; thus rhetorical skills are part of the language proficiency expected from the candidates. The censors give examples of good ways to begin and end a speech, as, for example, in the autumn 2011 Abitreenit episode:

> [1]HOST: do you need to introduce yourself in the beginning and do you need to have a clear beginning or ending
> CENSOR 1: there has to be some kind of an opening line of course, otherwise it's not a speech, some kind of an opening line like "hello how are you today" or something "dear friends good to see you here", there needs to be something
>
> …
>
> HOST: should you use the beginnings and endings which are given, for example, in textbooks, always like five alternatives, do you have to use them or
> CENSOR 2: that is one option, of course, but like I said a moment ago there are many options, many ways to end, to thank, of course, "thank you for attention", or, and many different possibilities, but some kind of an ending
>
> …
>
> CENSOR 1: a good ending is also rhetorical such as "we should never under any conditions have any sort of nuclear power now or ever be safe period" and then the end, everyone knows this is a rhetorical form

In the spring 2012 one of the censors also comments as follows:

> CENSOR 1: and excuse me but now I should also note that a person should understand some psycholinguistics, how to write to the BBC or to the managers of Yleisradio, to anyone who should be respected, and one should write by using a proper opening and so on "dear Sirs dear Madam slash Madam" etc

[1] I have transcribed the conversations which deal with the written production task from all six episodes of the Abitreenit programme. The extracts I include in the analysis are rough translations from the Finnish transcripts. The passages in quotes are original English utterances. The words in brackets are my additions for clarification. This applies to all the extracts taken from the Abitreenit programme in this analysis.

In the spring 2012 episode the censors comment on the proper register, i.e. the proper style in a certain social context, for a task:

> HOST: in number two, is it ok to begin briefly with “hi” or “hello”
> CENSOR 1: “hi” excuse me but “hi” is quite impolite, of course we definitely won’t take points because it has some kind of an opening but of course it affects the grade as a whole
> CENSOR 2: exactly
> CENSOR 1: because in our assessment there is this appropriateness and all that is related to the register etc

The markings on the essays also indicate that stylistically appropriate openings and endings in certain text types are expected and that inadequate versions are penalized in scoring. For example, essay number 086 on *My kind of TV-programmes* has a stylistic error marked in the opening: “Hey! Letter to THE BBC directors!” Systematic point deductions are, however, only applied when a salutation or a signature is omitted and not when it is only stylistically inappropriate. Details about which essays include elements such as salutation, date or signature and which essays were penalized are presented in the columns “Paragr.” and “Other” in Table B in Appendix 2.

Some of the features in the student essays that the raters pay attention to are idiomaticity and stylistic features. Seventy-three percent of the essays have at least one marking related to idiomaticity or style, as is shown in the column “Sty” in Table B in Appendix 2. In this section I discuss the stylistic features because they are closely related to communicative factors in writing. Other linguistic features such as orthography and punctuation are discussed in chapter 4.4 below. The types of errors marked which are related to style or idiomaticity can be divided into five categories: unconventional ways of addressing the recipient, incorrectly used fixed phrases, made-up phrases, colloquial language and other unidiomatic word choices. Examples of each category are given in Table 9 in Appendix 1. The examples are selected from different essays, and they are not linked to each other. The number of the essay is marked in each example.

In the examples presented in Table 9 the meaning is usually conveyed successfully despite the stylistic incorrectness. However, sometimes the raters add a question mark when part of the text is too ambiguous, as in the following cases:

- 008: “behind of everything”
- 030: “to every single breath we have”
- 031: “you can and you take it”
- 032: “then we are dit it”
- 039: Now days the society of an other culture divine and concer the happits of the other culture”
- 067: “I don’t think the new motorway would rise down the rush amount”
- 069: “If that is possible I hope so that to happening!”
- 104: “and only orginal soldier training system in -40 celcius”

I would argue that most of these examples would be intelligible in context, but the question mark written by the raters suggests that these cases have a more serious impact on the grading compared to the stylistic error types in Table 9 because they disturb the communicativity of the text. One of the lower-score essays (no: 097) has a comment from a censor saying the message is successfully conveyed despite numerous linguistic and stylistic errors. The score was raised from 40 to 48, apparently due to the communicative qualities.

Furthermore, the following comments, which I have translated into English from Finnish and Swedish, are related to communicative dimensions and written on the essays by the raters:

- 053: “a lot of repetition”
- 055: “a factual mistake”
- 067: “The meaning is unclear in several passages.”
- 067: “Only one fully correct sentence!”

The meeting minutes from 2010 report comments on the communicative merits and flaws of some of the essays discussed in the meetings. Examples of these comments are gathered from the meeting minutes in Table 10 in Appendix 1. The comments

imply that the communicative element indeed plays a crucial role in the assessment not only in the guidelines but also in practice.

4.3 Content

This section focuses on the aspects related to the content of the essays. I have divided this section into the following subsections: length, structure and topic handling. The structural features, especially those which have to do with genre and coherence, are closely related to communicativity as well, but in the rating scales in the guidelines the structural features are described together with the content features, and I decided to follow this division.

4.3.1 Length

In the Abitreenit commentary programme one of the most frequently asked questions about the written production task is about the word limit: whether and how many points will be deducted if the essay length exceeds or goes under the limit given in the task instructions. Each time the same question regarding the length limits is asked by several candidates, although the assessment guidelines provide clear penalty instructions for compositions which are too long or too short. In the older exams, before spring 2005, the task instructions stated that the length of the essay should be 150-200 words. Since spring 2005 the instructions have asked for 150-250 words. It is, however, possible to exceed the limit and write up to 312 words without sanctions, but after this systematic and gradual point deductions are to be applied. Essays with less than 90 words are penalized with a 30-point deduction, which is also mentioned in Abitreenit in spring 2012.

The guidelines provide exact instructions for point deductions when the length exceeds or goes under the word limit. In the 2002 and 2004 versions the deduction for overly long essays was 5-10 points, but since the 2005 guidelines came into effect there has been no maximum limit for the deduction. The number of points to be deducted increases by five points for every 25-percent excess of the word limit. For

essays which are too short, deductions are also applied in five-point intervals. In the guidelines from all years the minimum deduction for short essays is five points. The guidelines from 2002 and 2004 state that for essays which only comprise "a few sentences" the final score will be below 30. The 2005 guidelines elaborate the definition of a short essay by stating that a composition which goes under the instructed length by 40 percent or more would get a final score of less than 30 points. In the 2007 guidelines it is stated that the score of an essay with less than 90 words would automatically be below 30 points, whereas according to the 2011 guidelines an essay with less than 90 words would get less than 15 points. This suggests that the length of the essay has become an even more important factor in the evaluation.

The length of the essay does not only have a negative influence on the grading but the censors in Abitreenit in 2010 and 2011 say that good writers write almost 300 words in the production task. According to the censors, this is because it is easier for the examinee to show her/his language skills by writing a longer essay, and on the other hand, it is difficult to assess the level of the examinee if the text is too short. As seen in Table 5, there is indeed a positive correlation between the length of the essay and the score, which supports the censors' perceptions of the influence of the length on the quality of the essay. None of the essays included in the data were short enough for point deductions; the shortest essays have 149 words (essays number 012 and 022). However, essay number 033 from autumn 2006 exceeds the word limit with its 356 words, yet the markings on the essay refer to no point deductions.

Score (out of 99)	Essays	Average number of words
99-90	12	239
88-80	30	227
78-70	24	210
68-60	21	192
58-50	12	192
48-40	6	176

Table 5: The average number of words in the essays with certain score.

When considering the essay length as an indicator of good writing skills in an exam, it is relevant to acknowledge that the time constraints in the exam may greatly influence the examinee's ability to produce a text of a certain length. As the production task is placed at the end of the exam, many if not most examinees may leave it until last despite the possibility to proceed in the exam in an order of their choice. If the other parts of the exam have consumed a lot of time and energy, it may be difficult for the examinee to fulfil the essay length requirements.

4.3.2 Structure

Many candidates who comment and ask questions in the Abitreenit programme are unsure about how they are expected to formulate their essay when they are asked to write, for example, a speech, a letter or an article. One of the censors in Abitreenit (autumn 2012) comments that the different genres and text types that the examinees should write in the production tasks are presented in the textbooks used in upper secondary schools in Finland but it is up to the teacher whether the candidates have had practice in writing these genres. Here I use the term "genre" when I refer to the differences between texts which represent, for example, a speech, an article or a letter to the editor, to make a distinction between the genre categorization in Hyland (2004) - or "text type" categorization, as I have called it - and the genres I discuss here.

A speech is one of the most distinct genres in the task prompts and it is included regularly throughout the exams in the data. Twelve of the 84 task prompts instruct the examinee to write a speech. In several Abitreenit episodes the censors emphasize that a speech requires a beginning in the form of a salutation and an ending. They even give several stylistically appropriate examples, as mentioned above. Here I focus on the content point of view, i.e. what elements a genre requires, whereas earlier I looked at the formal requirements of different genres in the light of stylistic correctness, i.e. what those elements should look like.

In the case of other genres than a speech, the censors seem more uncertain about the conventions and requirements. One question in Abitreenit in 2010 was about whether the form of the text should be more like a letter or merely stating an opinion, and

whether a name or a nickname and date should be included in a letter to the editor. The censors' answers to this question were not completely in line with each other and appeared to be based on their intuition, because the final criteria would be decided in the censors' meeting and were not set at the time of the broadcast.

Some problems also occur when trying to define what formal properties a certain context requires. In the spring 2012 Abitreenit episode a candidate asked whether an email letter instead of a traditional letter would be approved, and the censors answered that it would be approved but that it would still require a proper opening and ending. Furthermore, in autumn 2012, a question arose about whether an article to be published in a newspaper should be signed or not. The censor's response was:

> CENSOR 1: yeah it's not required, I mean usually they are signed but the editors put the signature there, so in that sense there doesn't need to be any signature

Similar questions were discussed regarding the letter to the editor in the same exam:

> CENSOR 1: yeah and usually these online magazines, they have those text boxes for the text, but of course they always require a name or initials like in this example "JC by email" there has to be a name

In spring 2013 the question was about adding the date to the letter:

> CENSOR 3: and a letter requires an opening and ending and whether the date is required, which usually is the case with letters, it will probably be discussed later [in the censor meeting]

Thus one of the most problematic issues regarding the structure when an authentic-like response is expected is that the traditional print media genres have different formal conventions than online media genres, and even if the task prompt states for which type of magazine the response is meant, the difference between these two channels, from the assessment point of view, seems to remain unclear for the candidates and censors.

The censors' meeting minutes reveal more about the final decisions regarding the requirements for these formal genre conventions. In spring 2010 some example

essays were discussed at a meeting, and it was decided that the speech requires an opening and an ending and that the task which asks to write an article requires a signature, i.e. a name or a nickname. If the name/nickname was missing it led to a deduction of two points. Similar decisions were recorded also in later meetings: in autumn 2010 a five-point deduction for missing the elements of a speech: an opening and a closing line, in spring 2012 a two-point deduction for missing a "natural" (*luonteva*) opening in a speech or a signature in a letter, and also in autumn 2012 a two-point deduction resulted from missing a signature. It was also decided that the speech in task prompt number 2 in autumn 2012 required a "natural" opening and that the letter in spring 2013 required a signature but not a separate closing. Deductions for the missing elements were two points as well. Furthermore, some essay markings also have to do with genre requirements. If a letter to the editor, for example, is missing a signature, it is marked with an empty line at the end of the essay and penalized with a two-point deduction, as in essay number 086.

Coherence of the text is one of the important features that are under examination when evaluating the essays. It is closely related to the structure of the essay, although it undoubtedly has much to do with communicative aspects as well. According to the censors in the Abitreenit episodes in 2010, 2012 and 2013, a good structure for an essay includes a beginning with an introduction, an end with a conclusion and a middle part with a couple of separate ideas and subject matters in their own paragraphs.

Table 6 indicates that this norm of a good structure for an essay is passed on to the candidates, and most of them divide their essay into 3-5 paragraphs. According to the calculation, six paragraphs seem to give the highest score whereas less than three paragraphs are connected to lower scores. However, the essays which scored the highest, 99 and 97 points, only have three and four paragraphs respectively. The essay with the lowest score, 40 points, also has three paragraphs. Therefore it seems that writing at least three paragraphs is appreciated in the evaluation, but whether there are three, five or eight paragraphs does not seem to make much difference. The content of these paragraphs is more likely to have a greater importance.

Paragraphs	Number of essays	Average score
1	4	61
2	1	60
3	35	75
4	39	71
5	22	74
6	3	82
7	0	-
8*	1	72

*Table 6: Number of paragraphs and the average scores. *An estimate (the paragraphs were not clearly separated).*

The assessment guidelines from 2007 and 2011 specify as criteria for the highest scores, 90-99 points, that the text should be logical and that varying cohesive devices should be used competently. If the structure is illogical, no more than 48 points should be given.

4.3.3 Topic handling

All the rating scales in the guidelines from 2002-2011 provide criteria for assessing how well the topic has been handled. The topic-related criteria for a certain level in the scale are listed in Table 11. The data shows that variability is appreciated over one-sidedness, and that creative texts are considered better than conventional texts.

In the commentary programme the censors are asked multiple times whether it would decrease the score if the text does not stand out from the crowd. The censors do not directly say that the essay should differ much from the others but rather that it is tedious to read essays which are much alike.

A lot of other questions arise regarding the essay topic in the Abitreenit programme. For example, in spring 2010 one of the task prompts instructed to write a speech “on

a topic you find important to all young Europeans" (*Dear Fellow Europeans*). In the episode in spring 2010 candidates asked whether particular topics would be accepted, including nature conservation, the mental health of the youth and the need for the state to support young people's language learning. The censors considered these examples good topics, as well as any topic that the examinee finds important to the European youth.

Although the topics are often left open and different kinds of perspectives are accepted, the task prompts also have restrictions that are not always clear to the candidates or even the censors. For example, in spring 2010, the task prompt *Can one person change the world for the better?* instructed the examinee to "write about one person living today who you believe will do much good for the world". In the Abitreenit programme right after the exam one candidate asked if writing about Santa Claus was accepted, and both the censors thought it would definitely be accepted. The following dialogue, like many parts of the dialogue in the programme, has a humorous tone.

> HOST: and one of the viewers [examinees] thought Santa Claus was this kind of a person
> CENSOR 1: well of course Santa Claus
> CENSOR 2: yes that is fine
> CENSOR 1: as long as we remember that he lives in Korvatunturi
> CENSOR 2: yes
> CENSOR 1: of course it is good
> HOST: and you will probably get extra points if you remember to mention he lives in Finland
> CENSOR 1: yes
> CENSOR 2: mmh
> CENSOR 1: at least it gets the censor who reads it into a good mood, it will definitely bring good points

In the censor meeting on March 24th 2010, however, it was decided that if the examinee wrote about a fictional person or a person who has died, it would lead to a deduction of five points. It is probable that Santa Claus would be considered a fictional person because in the meeting held on March 22nd 2013 a similar question arose, and the meeting minutes state that if the letter (*Dear X*) is addressed to a fictional person such as Santa Claus, a point deduction of five points will be applied.

It also requires thorough planning to decide what kind of topics and themes are appropriate for the production task. In the matriculation examination the themes are selected based on the curriculum, which defines the themes for each course (The Matriculation Examination Board, 2011). One candidate in the spring 2010 Abitreenit episode criticized one of the four given topics, *James Bond - a hero of the past?*, saying that such a topic is not fair towards people who are not familiar with James Bond movies. The censors replied that it would be possible to write about James Bond even without having seen any of the films and presumed that there would be very few candidates who would have never heard of James Bond. They also noted that there were three other topics to choose from so there should be something for everyone. I have listed the main themes of the task prompts in the column "THEME" in Table A in Appendix 2. The themes vary from very specific to broad, e.g. the theme "James Bond" versus the themes "values" or "ethics". Based on the comments of the censor in the Abitreenit spring 2013 episode, the level of expertise on a certain theme or topic does not seem to be very important:

> HOST: in number one you had to write about climate change so was it supposed to be your own experience or should there also be some facts?
> CENSOR 1: well according to the instructions it should be "what is your stand on this question", and it could be based on your own experience, or if someone knows has read more about it, someone else less, but it's such a topical issue…

In the Abitreenit autumn 2011 episode the censors commented on the content requirements as shown in the following dialogue:

> CENSOR 2: here we see whether one can write in English, whether one can assert an opinion and so on, that is the point here
> CENSOR 1: yeah
> CENSOR 2: that we don't [look at] the content in that way but of course it is interesting to read what points of view the young people have on this issue
> CENSOR 1: yes the content needs to make sense but it is not like the mother tongue test

In the commentary programme the censors often emphasize the importance of following the task instructions strictly:

> HOST: in number two, if you only wrote about one TV-programme will you lose points

CENSOR 2: there is “what kinds of programmes you would like to watch”
CENSOR 1: yes
CENSOR 2: it asks what kinds, many
CENSOR 1: yes and how could sugges*tions* be only one

In most of the task prompts the examinees are expected to express their opinions and views on a particular matter, yet only some of these task prompts specifically ask the examinee to justify their opinion. In 23 out of the 84 task prompts argumentation is demanded in the instructions, as is shown in the ”ARG” (argumentation) column in Table A in Appendix 2 (these task prompts are marked with “Y” in the column). In the Abitreenit episode in spring 2010 when discussing the task prompt *Can one person change the world for the better?* the importance of argumentation becomes evident:

CENSOR 2: but there is an important thing at the end, “defend your view”, justify, defend, defend your view, you need to be able to do that
CENSOR 1: yeah
HOST: is it wrong if you ended up with the answer that no single person can change the world for the better but it requires help from others too
CENSOR 1: well it reads, I mean in the instructions [read] “defend your view”
CENSOR 2: if you can “defend” I mean the question is also in the interrogative form, can one person do this and that, that is what is asked here, s/he has then come to the conclusion that no, they can’t, or can, it has both options but as long as you can “defend”, justify, defend your view, I would say that both are ok

A further question is whether the same argumentation skills are required also when the task instructions do not specifically ask the examinee to justify and defend their opinion. For example, in the autumn 2010 exam only the task prompt discussed in the excerpt above instructs the examinee to defend her/his view. However, also two of the other task prompts have to do with expressing one’s opinion and view:

Dear Fellow Europeans (autumn 2010)
“You will be participating in European Youth Week and will give a speech on a topic you find important to all young Europeans. Write this speech.”

No way! (autumn 2010)
“You live in England and there are plans to build a new motorway which will pass very close to your home. Write a letter to be published in a local paper giving your opinion about the plan.”

Furthermore, the task prompts vary significantly regarding the support and input provided in the instructions. In the "SUPPORT" column in Table A in Appendix 2, I have listed which task prompts include some kind of ideas or questions to support the examinee's writing process. Some task prompts even provide examples which the examinee may use in their composition. Most of the supporting guidance is either "content" or "questions". It may also be examples, clarification of the instructions, limitations or conventions. Here are examples of each type of supporting guidance (I have highlighted the relevant passages with bold font):

> *Chalk and Talk or More Modern Technology* (spring 2012)
> "You have been invited to give the students' point-of-view in a youth delegation **discussing what makes good teaching** (=content). **Are traditional methods, like blackboards and exercise books** (=examples)**, enough, or do we need the latest gadgets?** (=questions) Write your speech."

> *A speech* (autumn 2008)
> "**Choose either A or B**. (=clarification)
> A You want to start a Slow Food Society in your community. Give a welcoming speech and **explain to your audience why the slow food movement is important**. (=content)
> B You want to start a Fast Food Club in your community. Give a welcoming speech **providing some arguments in favour of fast food**. (=content)
> **Remember appropriate ways of beginning and ending a speech**." (=conventions)

> *The future - our responsibility* (spring 2004)
> "**What are our responsibilities towards future generations? How do our present activities affect the future?** (=questions) This time, **do not write about nature conservation**." (=limitation)

The supporting guidance inevitably affects the production of the essay one way or another. It gives the examinee more information about the expectations, and thus it may also have a crucial influence on the evaluation. It may also restrict the examinee's thinking process and creativity as noted by Shaw and Weir (2007). As discussed above, the censors emphasize the importance of following the instructions in the task prompts carefully. When the instructions are more detailed, the assessment may also be more straightforward - and may possibly focus more on the content - compared to task prompts which have none or very little information on the expectations.

Some task prompts include a small piece of data, e.g. statistics (as in *American Pet Ownership* from spring 2012), an initiative to be replied to (as in *Letter of the month* in autumn 2012) or a reference to part of the text in the reading comprehension part of the exam (as in *Could our society do without the police?* in autumn 2007). All these cases are listed under the heading "DATA" in Table A in Appendix 2. Task prompts which include or refer to such data have become more common in the 10-year time span, as can be seen in Table A, but they are still only 14 out of the 84 task prompts altogether. The data-prompts which require analytical skills are from spring 2012, autumn 2011 and autumn 2009. In the spring 2012 episode of Abitreenit the censors answered a question regarding the task prompt *American Pet Ownership*, saying that the data needs to be included by drawing conclusions from the statistics:

> HOST: about number four, somebody focused mainly on dogs, cats and fishes so should they have somehow mentioned all of these in the essay after all
> CENSOR 2: I don't think so, you just need to draw conclusions about these animals and then discuss [them], s/he has of course included some of these so I don't think all of them need to be written about
> CENSOR 1: yeah
> CENSOR 2: it is enough if you handle this topic
> CENSOR 1: you should of course follow the instructions here also, so don't repeat these numbers when it clearly reads "do not repeat these numbers in your composition but instead draw conclusions from them"

The meeting minutes have several remarks about topic handling and content. In autumn 2009, if the examinee handled both aspects in *Dancing - I just love it! / Dancing - not for me, thank you!* a point deduction of five points was applied. In the same exam in *The most important gadget of modern life* the statistics needed to be discussed and in addition "something personal" was also expected. In spring 2012 a deduction of five points was applied if the topic handling in *My kind of TV-programmes* was one-sided, which is in line with the censors' comments in Abitreenit in the excerpt discussed above. In the same exam in *American Pet Ownership* if the statistics were not handled and the topic handling was one-sided, five points were taken for each flaw. In autumn 2012 it was noted that five points should be deducted if the title was misinterpreted. However, the task prompt with no instructions at all, *A lucky escape*, was mentioned as an exception and it was stated that deductions should be made as seldom as possible as long as content fit the title.

The cases in which writing about a fictional or a deceased person was penalized are discussed above.

Table 7 shows how four individual essays in the spring 2010 meeting were commented on in terms of the content. Five points were deducted in two cases where the instructions were not satisfactorily followed or the content deviated from the topic. Two other essays were described to be typical or conventional with regard to the content. How the conventional topic handling affected the evaluation is not clear as point deductions are not noted. It seems, however, that the two essays with "typical topic handling/content" have other merits to increase the score, because in the 2007 guidelines the rating scale level description for 70-78 points reads that "the reader handles the topic conventionally". Considering that content-related features are not the only factor or even the first priority in the assessment, it is not surprising that these essays were evaluated higher than what the content-related criteria would suggest. Also, the level for 80-88 points in the scale does not mention anything about creativity or conventionality; only the level for 90-99 points requires creative topic handling.

Exam	Score of the essay	Comments
Spring 2010	85	typical in structure and topic handling
Spring 2010	80	typical in language and content
Spring 2010	80	-5p, task not fulfilled
Spring 2010	65	-5 p for task deviation

Table 7: Comments on content and topic handling in individual essays from a censor meeting.

The markings in the essays in general have little to say about how well the examinee managed to handle the topic. In a few cases there were written comments in the essays about the topic being insufficiently handled. In the first example, in essay number 037 on *What should society protect?* in spring 2007, each of the three raters wrote a comment about the content, which they thought did not fit the topic and even seemed prepared in advance. The final score deduction was 20 points, leaving the final score at 60. In the second example, in essay number 047 on *My crazy festival* in spring 2008, one of the raters wrote "crazy?" next to the score meaning that the examinee wrote about a festival which the rater did not consider "crazy" enough. The other rater added: "topic/content -10 p". The same 10-point deduction was done with essay number 046 for the same reason. The third example, essay number 085 on *What is a family?* in autumn 2011, has a comment from the teacher saying that the essay deviated slightly from the topic and a comment from the censor saying that no problems were discussed, which lead to a deduction of five points.

The guidelines provide instructions for situations in which the examinee has misunderstood the task prompt or plagiarized the text partly or completely. The minimum deduction for deviation from the task prompt is five points. The maximum deduction was 50 point until 2007, at which time this limit was removed. When caught of plagiarism, the score will automatically be less than 40, and complete plagiarism leads to zero points.

4.4 Linguistic features

One of the most common concerns the candidates seem to have about the assessment of the written production task is the role of linguistic errors. For example, in autumn 2011 these questions were discussed in the Abitreenit commentary programme, as is shown in the following dialogue.

> HOST: about the content, if you wrote the word "unrenewable, non-renewable" will you lose points
> CENSOR 1: "unrenewable, non-renewable" hmm
> CENSOR 2: well we need to look at the whole, of course, what kind of an essay it is

> CENSOR 1: yes
> CENSOR 2: but of course it is
> CENSOR 1: I mean there still seems to be some kind of an idea that we deduct points for something, but we look at the whole
> CENSOR 2: yeah
> CENSOR 1: we look at [what is] positive in it not [what is] negative
> HOST: so you read them through positive eyes
> CENSOR 1: yeah I mean how much it expresses, how much it communicates, how mature this expression is
> CENSOR 2: how the topic has been handled, for example, that is the most important, and what kind of vocabulary there is, when we get these essays with maximum score there can always be some minor error in English
> CENSOR 1: yeah
> CENSOR 2: so many things influence
> CENSOR 1: yeah even when you have the maximum score there can be some
> CENSOR 2: article errors or misspelling or something like that there can be

Similar concerns about linguistic errors arose in spring 2013:

> CENSOR 3: so many have asked about these criteria, what are the criteria for the evaluation, if I wrote some word wrong then how many points will I lose, but that is not how we assess them, we don't count that there's a wrong word, minus two points
>
> …
>
> CENSOR 1: so apparently this idea still lives in the memories of the grandmothers, in the minds of today's youth that there was a time when two or four points were taken and so on

The other censor continues by saying that if someone misspelled one word several times, it will definitely not be counted as several mistakes but only one mistake. This suggests that even though the mistakes may not be counted and systematically used for defining the score, they are still taken into account. In fact, the markings made on the graded essays chiefly refer to linguistic errors such as misspelling.

In the analysis of the 105 evaluated essays from the 21 exams included in the data, I have focused on the markings made by the raters (teachers and censors). The guidelines instruct the raters to mark all incorrect and inadequate parts of the essays with specific signs. Stylistic errors, which were discussed in the communicativity section, are underlined with a broken line. Errors in word order are marked with a horizontal arrow, and a missing word is marked with a vertical arrowhead. Other

errors are marked by underlining. Margins are asked to be left empty apart from adding question marks if some part of the text is unclear. The raters are allowed to add brief comments at the end next to the score.

Table B in Appendix 2 lists all 105 essays in chronological order. In addition to score, word count, the number of paragraphs and stylistic features, I have listed linguistic markings made by the raters (teachers and censors). These markings point out errors in structure, orthography, punctuation and lexicon. All the essays have such markings ranging from only a few to an abundance of lines and arrowheads. The sign "Y" ("yes") in Table B marks which essays have a certain type of a linguistic marking.

a) Structural (grammatical) errors consist of all kinds of morphosyntactic errors. Here are some examples of structural errors marked in the essays:

- double negation: "never couldn't"
- use of articles: "most of the boys" (instead of "most boys")
- verb tense: "people had already been destroyed so much nature"
- use of reflexive pronouns: "people want to feel themselves free"
- word order: "That's one reason why I like so much dancing."
- comparison of adjectives: "more faster"
- S-V concord: "there lives many children" / "everyone who don't"
- irregular noun plurals: "bacteries"
- non-standard verb forms: "choosed" (instead of "chosen")

Many of the types of grammatical errors do not hinder the communicativity of the text, although in some cases the meaning changes due to the error, e.g. in "most of the boys". It is possible that not only the amount but also the severity of the grammatical errors influences the assessment.

b) Orthographic errors are marked when words that are semantically correct are misspelled. For example:

“suprise”, “beuaty”, “puplicly”, “ladyes”, “exsample”, “concider”, “whit”, “cowerment”, “ofcourse”, “thuff guy”, “doen’t”, “extra-ordinary”, “lifes”, “where ever”, “ideé”, “legendarig” “rythm”, “cant”, “crouded”, “singel”, “watc”

Whether the severity of the misspelling has any influence on the evaluation is not made explicit in the data. It could be, however, that an error such as “extra-ordinary” affects the overall impression of the writer’s proficiency less negatively than, for example, an error with a very basic word such as in “whit” (“with”).

c) Punctuation errors include errors in punctuation and capitalization. Most of the markings in the essays have to do with commas, but other types of errors are marked as well, such as incorrectly used question marks. Other examples are:

- “artists, like”
- “of course _ it’s hard for a mom to…”
- “finnish.”
- “When I was a little girl _”
- “Fortunately_"

While going through the markings in the essays, I observed that the importance of comma rules in English seems to vary between different raters. Some raters mark all comma errors they can find, some mark only some of the comma errors in the essays whereas some do not mark any comma errors at all. Therefore it seems that whether correct punctuation has effect on the grading is up to the rater.

d) Lexical errors mean that the examinee has used a semantically incorrect word which may or may not be orthographically correct and may or may not exist in the lexicon. Examples of lexical errors marked in the essays are:

- “the clock” (instead of “time”)
- “borned” (instead of “created”)
- “forceless” (instead of “powerless”)
- “workexperiment” (instead of “work experience”)

- “mannered” (instead of “brought up”)
- “heart starts to bomp” (instead of “pound”)

As with the other error types, the data offers no information on whether all lexical errors have equal weight in the overall evaluation. On the basis of my other data concerning communicativity, it is probable that the severity of the lexical mistakes depends on whether the meaning is understandable and whether the message is conveyed successfully even if not idiomatically.

Sometimes it is problematic to define which type of error is in question. For example, in several cases it is not unambiguous whether an error is orthographic or lexical. In most of these cases I have categorized the errors based on an assumption of what is the most likely category. For example, in one essay the examinee wrote “manors” when s/he presumably meant “manners”. It is likely that s/he merely misspelled the semantically correct word. Another similar example is the use of the word “loose” instead of “lose”. Furthermore, it is not clear whether words which are obviously misspelled due to the interference of other languages, as in “jobb”, “ideé” and “legendarig”, should be counted as orthographic or lexical errors. Here I have counted them as orthographic errors. In the case in which the examinee wrote “lukio” instead of “upper secondary school” I counted it as a lexical error because it is clear that the examinee did not have access to the correct English equivalent. A structural error may also resemble an orthographical error, e.g. “it’s” instead of “its”. In a few cases I have left the category open, and these cases are marked with a question mark in Table B in Appendix 2.

All essays have at least a few error markings, even the essay which was given 99/99 points. The 99-point essay has error markings related to structure and orthography. Other essays which have error markings of the same types got 62 points, 82 points, 68 points, 82 points and 92 points. The 99-point essay had only three markings:

- “in _ late 15th century”
- “Pal<u>e</u>olithic”
- “each path is <u>as</u> justified”

In comparison, the above mentioned essay which got 62 points has several markings related to structure and orthography. There are three orthographical errors (“luckie”, “millionare” and “civils”) and structural errors in verb forms (tense, S-V concord), prepositions, articles, noun number, word order and cohesive devices in a sentence. It is safe to say that the errors in the 62-point essay are not only more numerous but also of a more serious kind.

There is some inconsistency in the raters’ error markings. One examinee wrote the following sentence using an incorrect word order: “How we can become a grown-up?” The only markings have to do with the number of the noun. One examinee has orthographic errors in “achivement” and “doesnt” and another examinee wrote “se eye to eye”, which likewise were not marked. A similar case happened with the word “allmost”, which was marked as an error in one composition and left unmarked in another. Punctuation is also marked with some inconsistency, as in the case of “Dear directors.”, which has a full stop instead of a comma, and when using the phrase “for example” within a sentence without separating it with two commas. Some of the other essays, on the contrary, are heavily marked for punctuation errors.

The singular linguistic errors may not weigh much in the assessment but nevertheless, even in the assessment guidelines they are not invisible. In fact, the rating scales in the guidelines have quite a lot to say about the correctness of vocabulary and grammar, as seen in Table 12 in Appendix 1. The essays worth 90-99 points may have “some mistakes” whereas the essays worth 0-48 points have “a lot of mistakes” according to the 2002-2005 guidelines. The amount of mistakes is also mentioned in the 2007-2011 guidelines, albeit only for scores lower than 68. Therefore it is important that the errors be marked accurately.

Interestingly, although none of the individual linguistic errors give minus points apart from possibly affecting the quality of the composition as a whole, even the smallest errors in the titles were systematically penalized according to the guidelines before 2011. Writing the number and the title of the task prompt perfectly correctly used to have a greater weight on the grading than any individual linguistic feature. In the guidelines from years 2002-2005 a title missing a character such as a question

mark led to the deduction of 2 points. If the title was changed, it meant a 5-point deduction also in the 2007 guidelines. After the 2011 guidelines came into effect, the title was no longer needed unless specifically asked for in the task prompt, as long as the number was written.

In the 105 essays in the data, the last point deduction for an incomplete, changed or omitted title was done in the spring 2012. In spring 2011 one essay omitted the task number but the title was otherwise correctly written so no deductions were made. The other essays with an incomplete, changed or omitted title before spring 2012 have point deductions according to the instructions in the guidelines.

4.7 Comparison of three rated essays

Finally, I compare three essays with each other in order to get a more comprehensive idea of how the evaluation functions. I have focused on the following aspects: What is the examinee writing about? What arguments does s/he use? What kind of vocabulary and grammatical structures does s/he use? How coherent is the essay? How does the writer express her/himself? How does s/he manage to convey the message?

The essays which I have chosen for this part of the analysis are essays number 036, 038 and 039 from spring 2007 on the topic *What should society protect?* The selection is based on the scores; from the data of 105 essays and 21 topics this topic has the widest range of scores. From the five essays on the topic in question I selected the highest, a middle and the lowest score. The scores of the three essays are 97, 78 and 55, so there is an example of - using the terms mentioned in chapter 4.1 - a “very good”, a “rather good” and a “sufficient” essay. The word count of the essays is 250, 234 and 166 respectively.

The task prompt *What should society protect?* has the following instructions:

> "Society currently protects among other things old buildings, endangered animals, landscapes, even people. In your opinion, what are the most important things that should be protected and why?"

The guidance of the task prompt includes examples and a question and asks for argumentation.

First I look at how the writer handles the topic and which themes s/he discusses. The writer of the 97-point essay writes about humanitarian issues, protecting people, mental health problems and social exclusion. S/he focuses on the idea "even people" given in the task guidance, and s/he specifies the problem by concentrating on issues such as mental health. The writer of the 78-point essay discusses animals, landscapes and nature, which were also given as examples in the task prompt. S/he handles these themes at a more general level. The writer of the 55-point essay concentrates on culture: old buildings (which were given as an example), minorities, customs etc. S/he handles these themes at an even more general level.

The 97-point essay gives several examples from different points of view, e.g. Western world vs. Third World countries. It discusses the causes and consequences and proposes education as a solution and justifies the writer's idea. It also leaves space for other opinions: "I don't mean to say those things aren't important". The 78-point essay states that while everyone has their own opinion about what is worth protecting, people can share these opinions and act together to promote their idea. Then it talks about nature and animals and defends the writer's opinions by saying, e.g. "it is great that we have these nature conservation areas, because there animals can live in peace and nature stays untouched." (the markings shown are the ones made by the rates). There is only one concrete example, which deals with trees. The 55-point essay also includes argumentation, e.g.: "old buildings are the main thing in our culture, it's like a landscape tells us who we are". It also provides some concrete examples of what culture comprises.

The 97-point essay uses varied and advanced vocabulary, such as "alienation", "ancient", "beloved", "corporation", "disorders", "exist", "extent", "folklore", "flora and fauna", "irregularities" and "recovery". The writer uses many synonyms (e.g.

irregularities/illnesses/disorders) instead of repeating the same words. S/he also uses more complex grammatical structures fluently: “we find worth protecting”, “it does exist, even if not to the extent that does...”, “a problem far more difficult”, “the spreading of irregularities of the mind” and many more. There are only two minor linguistic markings, on structure and orthography. The 78-point essay also has some vocabulary of a more advanced kind such as “nature conservation”, “opportunities”, “untouched” and “endangered” (given in the task prompt). However, there is a lot of repetition. For example, the words “society” and “protect” are used together in a clause six times in the essay, and the word “protect” alone is used nine times altogether. The grammatical structures are not quite as advanced as in the 97-point essay and they often contain some minor errors, e.g. “everyone has their own opinion _ what should be protected and what not”, “which are close to their heart_” and “in my opinion it is also good that society protects _ for example _ old trees, because as far as I’m concerned_ it is important to save something old while we are creating new things.” The essay has markings on structure, punctuation and lexicon. The 55-point essay also includes some more advanced vocabulary, although misspellings are frequent; “aband[on]”, “achivement” (no marking) and “majoroty”. The grammatical structures are not as complex as in the other two essays, and there are a lot of errors, e.g. “it’s like a landscape but only that it tells us who we are?, Were we are from?” and “every culture have it’s own happits”. There are markings on structure, orthography and punctuation.

The structure of the 97-point essay is clear and logical. The essay begins by arousing the reader’s interest and introduces the topic by giving examples: “the birthplace of a beloved author or a composer, an ancient poem or a folklore, the flora and the fauna”. The first paragraph ends with an ethical question directed to the reader. The second paragraph starts with an overview of the current situation and provides examples, presents the problem and offers a solution. The third paragraph includes a conclusion with arguments and finally brings the thoughts together by inviting everyone to participate in making the world a better place.

The 78-point essay also has the thoughts divided into separate paragraphs, but not in as coherent and compact way and it seems less thought out. The first paragraph is a general introduction and does not express the writer’s own opinion or give examples,

but it provides a fluent transition to the second paragraph: “Fortunately_ people can also, in groups, protect _ things which are close to their heart_.” The second paragraph begins with the writer’s opinion about animal rights and includes some argumentation. In the third paragraph the writer introduces another idea concerning the importance of protecting landscapes, which s/he justifies by describing them “beautiful and rare” and that both animal and humans need them. The fourth paragraph consists of only one sentence and it continues with the nature theme by explaining why nature conservation is important. The fifth paragraph concludes that nature is the most important thing and that it should be protected: “nature has given us everything we need_ and I think it’s time to give something back.”

The 55-point essay has no separate paragraphs, seems even less thought out and is missing a clear main idea. It begins by stating the writer’s opinion: “In my opinion sosiety must and should protect the culture.” Then the writer moves on to talk about old buildings and why they should be protected, after which s/he states that also customs of different cultures are important and that cultural minorities suffer from discrimination, which s/he argues is “because it [the customs of minorities] may seem savage to them [the majority]”. S/he concludes the essay by stating that “it [the savageness?] doesnt (no marking) give us the rights to rule over the smaller one”.

In respect of expressiveness and communicativity, the differences between the three essays are quite evident. The 97-point essay has a coherent, clear and well thought-out structure, which makes it very easy to read. The composition proves that the writer is really thinking and processing the themes s/he writes about, which results in a text that has something interesting to offer to the reader. There are several examples, although they lack a personal or a more concrete dimension, leaving the essay slightly vague. The expressions are rich and imaginative in many parts, i.e. the writer uses unconventional expressions such as “everyone’s right to view the Eurovision Song Contest”, which makes the essay stand out from the crowd. The 78-point essay also manages to convey the message quite clearly due to the logical structure (moving from animals to landscapes and then nature in general) and fluency of the language. However, it seems that the writer does not have a very clear idea and opinion about the topic, which shows in the way s/he moves from one idea to another and staying at a rather general level. The expressions are conventional but quite well

mastered. The 55-point essay has some problems in conveying the message, which is due to incoherence, unclear references and severe errors in grammatical structures and spelling, but also because the writer does not seem to have a clear idea in her/his mind and s/he jumps from one idea to another. An example of an unclear sentence is: "Now_days the society of an other culture divine and concer the happits of the other culture, because it may seem savage to them." The 55-point essay is clearly less communicative than the other two and struggles to express the ideas of the writer.

In conclusion, it seems that the features of the three essays are in line with the scores given. Based on the 2005 guidelines, which were used for the spring 2007 exam, an essay with 97 points should be pleasant to read, authentic-like and fluent with rich, idiomatic and varied expressions. The topic handling should be creative and varied as well, and the language may include some mistakes. The example essay fulfils all these criteria. An essay with 78 points, on the other hand, should be quite easy to read and relatively natural, and although the basic structures are in order, the structures and vocabulary are unvaried. The topic handling is conventional and not very diverse, and more errors are allowed. This more or less applies to the 78-point essay, which is at the highest end of the scale for these criteria. An essay with 55 points should, according to the guidelines, be relatively difficult to read, and some passages may be unclear. The language use is inadequate and the expressions are basic and one-sided, as is topic handling. It has quite many mistakes in basic structures and disturbing interference of other languages. These criteria seem to agree with the given score. Thus, although the evaluation guidelines and instructions are as vague as they are, the raters seem to have a shared understanding of how to apply the evaluation criteria in the grading.

5 Discussion

In this chapter I first discuss the findings based on the analysis, after which I reflect on the study as a whole.

5.1 Findings

As expected, this study implies that the assessment guidelines are the primary basis for the assessment of the essays. According to the analysis, the rating process involves intuition and, in Shaw and Weir's (2007) terms, internalized representation of the rating scales. This is not surprising, given the vagueness and generic nature of the scales and the fact that each level of the rating scale is connected to several different scores. Both of the rating scale types in the data describe eight levels of proficiency although the number of possible scores is 33. As such, the nuances between different scores within the eight levels are subject to the interpretation of the raters. Even with a rating scale with 33 level descriptions the same problem with interpretation would remain because the definitions would be difficult to differentiate. However, the MEB meetings function as a kind of rater training, in which the aim is to find agreement on the assessment principles. Whether this is enough to ensure consistency between different raters depends on the individuals and how well the issues discussed and decided in a meeting also reach the censors who are absent. Nevertheless, the minimum of two raters for each essay increases the reliability of the assessment.

A question regarding the scoring system and criteria is how much they are in line with the curriculum, since the highest scores require skills well above the set target level. One way to change this would be to adjust the description of the top end of the rating scale to correspond the level B2.1. If this was done, a much greater number of examinees would reach the highest scores. However, due to the normal distribution of the grades, the results would do very little justice to the examinees' performances. Another way would be to change the target level in the curriculum. In fact, a new curriculum for general upper secondary education was published in 2015 and it still presents B2.1 as the target level for the advanced English syllabus. However, the

criteria for the B2.1 level have changed to focus entirely on communicative aspects. The level B2.1 in the 2015 curriculum is generally described as, translated from Finnish: “the basic level of independent language proficiency” (Finnish National Board of Education, 2015), which still is not quite equal to the requirements of the highest scores and grades and does not describe the quality of the best essays in the data. One may argue that the set target level describes an average or “good” performance, which is justifiable in the sense that the expectations are not set unreasonably high for most candidates, but at the same time it raises the question: why do I not get the highest score when my performance achieves all the objectives set in the curriculum?

The analysis of the task prompts reveals that the text types represented in the task prompts vary very little, and that academic writing skills, which require argumentation, description of a phenomenon and sometimes even analytical skills, are the most emphasized types of text expected from upper secondary students in EFL writing. These skills are undoubtedly relevant for students who plan on continuing their studies in universities. The rest of the task prompts ask for a presentation of factual information or a reconstruction of past events. Furthermore, some task prompts, i.e. speeches demand rhetorical skills. The unspecified task prompts do, however, leave room for creative writing, but it is difficult to say how the raters would treat a composition of an examinee who decided to write, for example, a short story, a song or a poem. At least a short story i.e. narrative would provide the same linguistic possibilities as the text types which are included in the production tasks.

One feature of the production tasks which was not much covered in this study is the genre of the texts. For example, a speech stands out as one of the most common and distinguishable genres. Many candidates seem to be perplexed about the formal properties required in different genres such as a letter to the editor or an article. Also the differences between the conventions for texts to be published in online and print media forums confuse the candidates and do not seem to be very clear to the raters either, as the analysis suggests.

The role of communicative aspects is much emphasized at the theoretical level, but in the essay markings it shows surprisingly little. The influence of communicative

factors shows mostly in the markings related to stylistic errors and use of inappropriate register. Sometimes the essays contain markings in parts of the text which are fully understandable and acceptable, which creates a conflict between the ideas of communicative fluency and communicative practicality.

Including a context in the task prompt has become more and more common. The context presumably helps the candidate to imagine her/himself in the role in which s/he is supposed to write and increases the feel of authenticity and meaningfulness. The question is whether the contexts are relevant and whether they represent situations which are likely to occur in the lives of the candidates.

Similar to the issues with communicative features, the influence of content-related features on the assessment remains rather ambiguous. Based on the results of the analysis, the length of the essay has a relatively clear and important role in the assessment. Questions regarding length seem to trouble the candidates every semester, although detailed instructions are available and teachers should instruct the candidates before the exam. In addition to essay length, coherence and logical organization seem to be some of the best perceived factors in the assessment. Otherwise the censors' comments and the guidelines on content are more imprecise. For example, sufficient topic handling is mentioned in the rating scales but no details are given. The guidelines note that point deductions will be applied in cases where the examinee has deviated from or changed the title. The censors comment on the importance of logical topic handling but other than referring to the title or what the task prompts say they cannot give specific answers to the candidates' questions regarding topic handling requirements, e.g. from what point of view it is acceptable to write and whether creativity weighs in the assessment.

Some of the topics in the task prompts are based purely on personal experiences, but most topics require content knowledge. The censors' comments indicate that in tasks which require content knowledge e.g. about climate change, the level of the examinee's expertise on the topic does not have a major impact on the grading. As mentioned above, the degree of support provided in the task prompt influences the examinee's performance (Shaw and Weir, 2007). The comparison of the three essays in chapter 4.5 supports this idea, as the themes and vocabulary in the essays were clearly influenced by the examples given in the task prompt. Whether the amount of

support given in the task prompt significantly affects the perspective of the assessment as well remains unresolved. Another content-related problem pointed out is that the examinee may find it difficult to identify with the topics given. The data suggests, however, that the topics vary not only in themes but also in how specific they are.

Linguistic errors seem to be a common concern among the candidates. According to the censors, the concern over direct point deductions related to e.g. spelling and grammar is unnecessary. Yet, the markings on the essays are not in line with the idea of linguistic errors having only a minor role in the assessment. Undoubtedly the function of linguistic error markings is beyond merely pointing out orthographic or grammatical errors, for these errors cannot be entirely isolated from the other factors, i.e. communicativeness and content. As the linguistic errors are something concrete to lean onto and simple to mark, their significance should not be swept under the rug.

What is more striking than the significance of linguistic features is that the assessment seems to have a strongly negative point of view. Almost all the markings on the essays focus on errors, and only a few essays also have comments or markings of a positive kind. In the commentary programme the censors say, in contrast, that they focus on the positive attributes of the essays. Since the higher level descriptions in the rating scales are merit-oriented and given that the rating process is for the most part invisible (i.e. all the justification for a given score does not show in the essay markings), this is probably true to a large extent.

Overall, the connection between the criteria set in the guidelines and the markings in the essays remains loose. The guidelines describe proficiency from several aspects and emphasize communicativeness, whereas the markings focus almost entirely on linguistic errors. Thus the assessment seems to rely on the interpretation and internalized apprehension. This becomes evident also in the meeting minutes, which show that example essays for different scores are discussed in the meetings.

5.2 Reflection on the study

The data mainly included material from the Matriculation Examination Board. Some of the material is available publicly, which simplified the reporting process. The handling of the non-public material required more caution, and especially the analysis of the essays proved challenging due to the time and accessibility constraints and confidentiality. Furthermore, a large part of the data required translation into English, which considerably added to the challenges of the analysis and reporting.

The analysis provides a lot of information about the criteria for assessing the essays. The raters use the official guidelines as the basis of their evaluation and grading, but also individual factors influence the assessment process. Because the guidelines are very vague and general, they do not give detailed instructions for evaluating a certain task. For example, they do not provide separate criteria for speeches, letters to the editor, book reviews etc. This is why analysing other pieces of data as well was crucial in order to understand the assessment process more extensively.

The analysis of the essays gave surprisingly little information about the assessment process, because the essay markings are not much related to the rating scale criteria. Apart from some concrete examples the meeting minutes did not provide a lot of material on the rating process either. On the other hand, I was surprised how much information emerged from the Abitreenit commentary programme. This probably has to do with the fact that the programme is public and the censors are there to comment and answer questions about the assessment, whereas in the non-public material there is no immediate need to report and explain the decisions regarding the grading.

In this study the focus was on the censors and not the teachers, although no distinction between the essay markings made by teachers and censors was made. This is both due to the limitations of the data and the fact that the censors are responsible for the final score. The censors also represent the matriculation examination board who design the exams. Nevertheless, the teachers' role in the assessment should not be underestimated, and it may only be speculated whether and how much the markings made by the teacher affect the censor's assessment process – and whether or not, in reality, the censors never see the scores given by the teachers before giving their own score.

The comparison of the three essays brought the different aspects which emerged from the other data together and showed that the scores given did indeed agree with the level descriptions in the rating scale, at least on a broader level. Whether the exact scores represent the performance of the examinees is difficult to say because there are no descriptions for these nuances. It must also be noted that the comparison involved my personal interpretation, and therefore it is left debatable whether my interpretations are comparable with the interpretations of the raters.

The task prompts in themselves say a lot about the expectations regarding the EFL writing skills of an upper secondary school student. Because most of the task prompts include more instructions than the mere title, it is possible to draw conclusions about what kinds of texts are required from the candidates and what kinds of writing skills are considered useful and important for upper secondary school graduates.

With regard to the candidates' concerns, the data was limited to the questions selected and asked by the host and the guest candidates in the commentary programme. However, the selected questions concern many different areas and they presumably give a relatively good idea of what the most frequently raised issues are. Also, it is often mentioned in the programme that a specific type of question has been asked by many candidates or viewers. To get more data on the candidates' questions and concerns it would have been convenient to have the access to the messages sent to the Abitreenit chat room, but unfortunately the old chat conversations are not available. Furthermore, interviews or questionnaires could give more information on this issue.

6 Conclusion

The purpose of this study was, first of all, to find out more about the assessment of the written production task in the English matriculation exam by looking at the criteria in and outside the official guidelines and regulations. Secondly, the aim was to determine what kinds of texts the students in general upper secondary school are expected to produce. Third, the study aimed to identify concerns raised by candidates who take the matriculation examination.

Overall, this study supports the impression of some degree of intuition and vagueness involved in the assessment of the essays, which I expressed in the introduction. It does not, by any means, suggest that the assessment was arbitrary; on the contrary, many factors indicate that a lot of effort is put to ensure the examinees a reasonable level of fairness. If any suggestions for improvement should be made, I would propose more transparency in the essay markings for justification of the rating. Also, consistency in error marking is crucial and may require some more attention from the raters. With regard to the text types of the task prompts, some alternatives are offered (e.g. argumentative, descriptive and recounting). The task prompts also vary in how detailed they are, and some of them have no guidance at all. The guidance in the task prompts (or the lack of it) and the themes are things that raise questions among candidates. Also linguistic correctness and formal properties seem to be some of the major concerns for many of them. Although the candidates desire concrete and unambiguous assessment criteria, most of their questions can be answered only partially; the censors are able to give clear answers only when the information is explicitly stated in the guidelines.

Further research could reveal more about the holistic rating process and give additional information about factors in the assessment which do not show in the essay markings. For example, the censors themselves claim that in the assessment factors other than linguistic correctness weigh the most and that the focus of the assessment is on the merits rather than flaws. As this is contradictory to the essay markings, rater interviews and questionnaires would undoubtedly illuminate these issues. Furthermore, it would be interesting to compare the censors' assessment principles to those of the teachers' to see whether there are any differences.

However, comparing the censors' and the teachers' error markings in the essays for this purpose would be complicated, because the censors do not mark elements which have already been marked by a teacher or another rater.

The text type analysis provides information on what kinds of writing skills are expected from upper secondary school students, but analysis on genre would definitely add to this information. In fact, my original intention was to include an analysis on the genre of the tasks, but it proved challenging due to the difficulty to categorize the task prompts which do not specify the genre or even provide a context as a clue. Also research on context in the task prompts would bring another perspective on what kinds of texts are asked from the candidates. This could be studied by asking upper secondary school students and/or graduates in what kinds of contexts they need or think they will need English language writing skills.

To be able to find out more about the candidates' concerns regarding the assessment of the written production task, it would be convenient to gain access to the messages sent to the Abitreenit chat room during the commentary programme. This could be done during the future episodes of the programme. Other practical way of getting data on this question would be student questionnaires.

To conclude, the results of this study will hopefully provide new perspectives for an upper secondary school English teacher who is evaluating either course essays (which are usually evaluated based on the matriculation exam criteria) or the actual matriculation exam compositions. Furthermore, although the study is by no means all-encompassing it points out several aspects which have a potential to prove useful to test developers and raters. In a broader perspective the study specifies some of the expectations in foreign language learning in general upper secondary school education, and thus aids teachers in planning courses and especially the practice of written skills.

References

Primary sources:

Yleisradio. 2010. Abitreenit: Yo-kokeet. 2010 kevät: englanti pitkä oppimäärä. [online] Video available at: <http://oppiminen.yle.fi/abitreenit/englanti/yo-kokeet/2010-kevat-englanti-pitka-oppimaara> [Accessed 9 December 2015].

Yleisradio. 2011. Abitreenit: Yo-kokeet. 2011 syksy: englanti pitkä oppimäärä. [online] Video available at: <http://oppiminen.yle.fi/abitreenit/englanti/yo-kokeet/2011-syksy-englanti-pitka-oppimaara> [Accessed 9 December 2015].

Yleisradio. 2012a. Abitreenit: Yo-kokeet. 2012 kevät: englanti pitkä oppimäärä. [online] Video available at: <http://oppiminen.yle.fi/abitreenit/englanti/yo-kokeet/2012-kevat-englanti-pitka-oppimaara> [Accessed 9 December 2015].

Yleisradio. 2012b. Abitreenit: Yo-kokeet. 2012 syksy: englanti pitkä oppimäärä. [online] Video available at: <http://oppiminen.yle.fi/abitreenit/englanti/yo-kokeet/2012-syksy-englanti-pitka-oppimaara> [Accessed 9 December 2015].

Yleisradio. 2013a. Abitreenit: Yo-kokeet. 2013 kevät: englanti pitkä oppimäärä. [online] Video available at: <http://oppiminen.yle.fi/abitreenit/englanti/yo-kokeet/2013-kevat-englanti-pitka-oppimaara> [Accessed 9 December 2015].

Yleisradio. 2013b. Abitreenit: Yo-kokeet. 2013 syksy: englanti pitkä oppimäärä. [online] Video available at: <http://oppiminen.yle.fi/abitreenit/englanti/yo-kokeet/2013-syksy-englanti-pitka-oppimaara> [Accessed 9 December 2015].

Secondary sources:

Council of Europe, 2001. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR).* [online] Available at: <http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf> [Accessed 6 December 2015].

Dörneyi, Z., 2007. *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

Finlex, 1998. *Lukiolaki*. Edita Publishing Oy. [online] Available at: <http://www.finlex.fi/fi/laki/ajantasa/1998/19980629> [Accessed 6 December 2015].

Finlex, 2005. *Valtioneuvoston asetus ylioppilastutkinnosta*. Edita Publishing Oy. [online] Available at: <http://www.finlex.fi/fi/laki/ajantasa/2005/20050915> [Accessed 6 December 2015].

Finnish National Board of Education, 1994. *Lukion opetussuunnitelman perusteet 1994.* 2nd ed. Helsinki: Painatuskeskus.

Finnish National Board of Education, 2003. *National Core Curriculum for Upper Secondary Schools 2003*. [online] Available at: <http://www.oph.fi/download/47678_core_curricula_upper_secondary_education.pdf> [Accessed 9 February 2016].

Finnish National Board of Education, 2015. *Lukion opetussuunnitelman perusteet 2015*. [online] Available at: <http://www.oph.fi/download/172124_lukion_opetussuunnitelman_perusteet_2015.pdf> [Accessed 9 February 2016].

Harlen, W. and James, M., 1997. Assessment and Learning: Differences and Relationships Between Formative and Summative Assessment. *Assessment in Education*, [e-journal] 4(3), pp.365-379. Available at: <http://www.tandfonline.com.libproxy.helsinki.fi/doi/abs/10.1080/0969594970040304> [Accessed 21 February 2016].

Hyland, K., 2004. *Genre and Second Language Writing*. Michigan: University of Michigan Press.

Knoch, U., 2009. *Diagnostic Writing Assessment: The Development and Validation of a Rating Scale*. [e-book] Frankfurt am Main: Peter Lang AG. Available through: University of Helsinki Library website <https://helka.linneanet.fi/cgi-bin/Pwebrecon.cgi?LANGUAGE=English&DB=local&PAGE=First> [Accessed 11 February].

Kranert, M., 2013. *Korrigieren, Prüfen und Testen im Fach Deutsch als Fremdsprache*. [pdf] Available at: <http://www.geisteswissenschaften.fu-berlin.de/we04/germanistik/studium/studiengaenge/master/master_daf/download/Michael-Kranert---Korrigieren_-Pruefen-und-Testen-im-Fach-Deutsch-als-Fremdsprache-1.pdf> [Accessed 5 February 2016].

Kunnan, A.J., 2012. High-Stakes Language Testing. *The Encyclopedia of Applied Linguistics*, [e-journal] Available at: <http://onlinelibrary.wiley.com.libproxy.helsinki.fi/doi/10.1002/9781405198431.wbeal0504/abstract> [Accessed 4 May 2015].

Lee, I., 2011. Formative Assessment in EFL Writing: An Exploratory Case Study. *Changing English: Studies in Culture and Education*, [e-journal] 18(1). Abstract only. Available at: <http://www.tandfonline.com/doi/full/10.1080/1358684X.2011.543516> [Accessed 21 February 2016].

Leung, C., 2012. Qualitative Research in Language Assessment. *The Encyclopedia of Applied Linguistics*, [e-journal] Available at: <http://onlinelibrary.wiley.com.libproxy.helsinki.fi/doi/10.1002/9781405198431.wbeal0979/abstract> [Accessed 4 May 2015].

McNamara, T., 2000. *Language Testing*. Oxford: Oxford University Press.

Meyer, W.H., 2009. 'When you just get a mark and a nasty comment, what's that called?' 'Summative assessment'. Creating an Enabling Environment for Formative Assessment. *Southern African Linguistics & Applied Language Studies*, [e-journal] 27(2), pp.215-228. Available at: <http://www.tandfonline.com.libproxy.helsinki.fi/doi/abs/10.2989/SALALS.2009.27.2.8.871> [Accessed 9 February 2016].

Mo, Y., 2015. Exploring Task and Genre Demands in the Prompts and Rubrics of State Writing Assessments and the National Assessment of Educational Progress (NAEP). *Dissertation Abstracts International Section A: Humanities and Social Sciences*, [dissertation abstracts] 76(1-A(E)). Abstract only. Available at: <http://search.proquest.com.libproxy.helsinki.fi/docview/1692315950/abstract?source=fedsrch&accountid=11365> [Accessed 21 February 2016].

Montgomery, J.L and Baker, W., 2007. Teacher-Written Feedback: Student Perceptions, TeacherSelf-Assessment, and Actual Teacher Performance. *Journal of Second Language Writing*, [e-journal] 16(2), pp.82-99. Available at: <http://www.sciencedirect.com.libproxy.helsinki.fi/science/article/pii/S1060374307000318> [Accessed 21 February 2016].

Neumann, H., 2014. Teacher Assessment of Grammatical Ability in Second Language Academic Writing: A Case Study. *Journal of Second Language Writing*, [e-journal] 24, pp.83-107. Available at: <http://www.sciencedirect.com.libproxy.helsinki.fi/science/article/pii/S1060374314000204> [Accessed 11 February 2016].

Norris, J. M., 2008. *Validity Evaluation in Language Assessment*. [e-book] Frankfurt am Main: Peter Lang AG. Available through: University of Helsinki Library website <https://helka.linneanet.fi/cgi-bin/Pwebrecon.cgi?LANGUAGE=English&DB=local&PAGE=First> [Accessed 11 February].

Oscarson, A.D., 2009. Self-Assessment of Writing in Learning English as a Foreign Language: A study at the Upper Secondary School Level. *Göteborg Studies in Educational Sciences 277*. [e-book] Göteborg: Acta Universitatis Gothoburgensis. Available at: <http://eric.ed.gov/?id=ED505960> [Accessed 21 February 2016].

Shaw, S.D and Weir, C.J., 2007. *Examining Writing. Research and Practice in Assessing Second Language Writing*. Cambridge: Cambrige University Press.

Shin, S-Y., 2013. Proficiency Scales. *The Encyclopedia of Applied Linguistics*, [e-journal] Available at: <http://onlinelibrary.wiley.com.libproxy.helsinki.fi/doi/10.1002/9781405198431.wbeal1423/abstract> [Accessed 4 May 2015].

Taras, M., 2005. Assessment – Summative and Formative – Some Theoretical Reflections. *British Journal of Educational Studies*, [e-journal] 53(4), pp.466-478. Available at: <http://www.tandfonline.com.libproxy.helsinki.fi/doi/abs/10.1111/j.1467-8527.2005.00307.x> [Accessed 11 February 2016].

The Matriculation Examination Board, 2011. *Toisen kotimaisen kielen ja vieraiden kielten kokeiden määräykset*. [pdf] Available at: <https://www.ylioppilastutkinto.fi/images/sivuston_tiedostot/Ohjeet/Koekohtaiset/fi_maaraykset_kielikokeet.pdf> [Accessed 4 May 2015].

The Matriculation Examination Board, 2015. *Tilastot: Ylioppilastutkintoon osallistujat kokeittain*. [pdf] Available at:

<https://www.ylioppilastutkinto.fi/images/sivuston_tiedostot/stat/FB2015ST2001.pdf> [Accessed 6 December 2015].

Turner, C.E., 2012. Rating Scales for Language Tests. *The Encyclopedia of Applied Linguistics*, [e-journal] Available at: <http://onlinelibrary.wiley.com.libproxy.helsinki.fi/doi/10.1002/9781405198431.wbeal1045/abstract> [Accessed 10 February 2016].

Weir, C. J., 2005. *Language Testing and Validation: An Evidence-Based Approach*. [e-book] Basingstoke: Palgrave Macmillan. Available through: University of Helsinki Library website <https://helka.linneanet.fi/cgi-bin/Pwebrecon.cgi?LANGUAGE=English&DB=local&PAGE=First> [Accessed 11 February 2016].

Appendix 1

Table 8: My translation from the original Finnish rating scale descriptions in the guidelines from years 2002-2011 (communicativity/readability and language use).

Score (out of 99)	**Criteria: communicativity (guidelines 2007-2011)**	**Criteria: readability and language use (guidelines 2002-2005)**
99-90	The writer is able to convey the message very clearly, realistically, fluently and vividly. The text is very easy to read.	Pleasant to read. Authentic and fluent.
88-80	The writer is able to convey the message clearly and relatively fluently. The text is easy to read.	Easy to read. Fluent.
78-70	The writer is able to convey the message fairly clearly and relatively naturally. The text is fairly easy to read.	Fairly easy to read. Relatively natural.
68-60	The writer is able to convey the message satisfactorily. The text is partly difficult to read.	Some elements make the reading difficult. Some faltering in language proficiency.
58-50	The writer is able to convey the message only unclearly. The text is difficult to read in many passages. Singular passages may be left unclear.	Relatively difficult to read. Some passage may be left unclear. Insufficient control of the language.
48-40	The writer is able to convey the message poorly and in some passages not at all. Many passages are left unclear.	Difficult to read. The meaning is left unclear in some passages. Poor control of the language.
35-20	The writer is hardly able to convey the message. The text is very difficult to read.	Very difficult to read. The meaning is left unclear in several passages. Very poor control of the language.
15-0	The writer is not able to convey the message.	Due to the nonexistent language proficiency the text is almost or completely incomprehensible. (*Olemattoman kielitaidon vuoksi lähes tai täysin käsittämätön tuotos.*)

Table 9: Types of stylistic errors marked in the essays.

1) unconventional ways to address the recipient	**2) incorrectly used fixed phrases**	**3) made-up phrases**	**4) colloquial language**	**5) other unidiomatic word choices**
076: "Dear mr/mrs editor"	020: "storm in tea cup" (used incorrectly)	051: "question marks left open" (instead of "questions left unanswered")	033: "grand-mum"	058: "to use my money" (instead of "spend")
089: "Dear BBC"	042: "hold my horses" (used incorrectly)	072: "the 'looking for worker' site"	025, 035, 037: "like" (used colloquially)	073: "jobless" (instead of "unemployed")
078: "Hey!" (in a formal letter)	098: "every and each one of us"	048: "under a permission"	038: "I'm" (instead of "I am")	099: "steady opinions"
087: "Hello, the BBC directors!"	059: "I and my brother"	062: "deep in me I might also envy the dancers"	040: "Protecting is saving something for a future. For someone who is coming after us." (incomplete sentence)	037: "trust for the already big and dominating enterprises is so big"
088: "Dear Mr BBC directors"	090: "March 16th of 2012"	078: "what comes to" (instead of "when it comes to")	100: "bad" (instead of "badly")	051: "army differences"
086: "Hey! Letter to THE BBC directors!"	070: "at first" (instead of "firstly/first of all")	096: "There Mark Twains words are giving his opinion"	083: "oldish countries"	057: a man (instead of the passive "one")

Table 10: Comments on communicative features in individual essays from censor meetings.

Exam	Score of the essay	Comments on the essay
Spring 2010	70	communicative although contains many errors
Spring 2010	52	a lot of linguistic problems but communicates
Spring 2010	45	poorly communicative
Autumn 2010	95	a lot of merits but some artificiality
Autumn 2010	66	repetitive
Autumn 2010	65	communicatively readable
Autumn 2010	58	has content and communicates
Autumn 2010	55	repetitive
Autumn 2010	40	poorly communicative

Table 11: My translation from the original Finnish rating scale descriptions in the guidelines from years 2002-2011 (content/ handling of topic).

Score (out of 99)	**Criteria: content (guidelines 2007-2011)**	**Criteria: handling of topic (guidelines 2002-2005)**
99-90	The writer handles the topic very diversely and creatively.	Creative and diverse.
88-80	The writer handles the topic diversely.	Clear but quite conventional.
78-70	The writer handles the topic conventionally.	Conventional and quite narrow.
68-60	The writer handles the topic quite one-sidedly.	Handling of the topic is narrow.
58-50	The writer handles the topic one-sidedly and/or repetitively.	One-sided.
48-40	The writer handles the topic insufficiently and/or very repetitively.	Inadequate due to insufficient language proficiency.
35-20	The writer handles the topic very insufficiently.	Rudimentary due to poor language proficiency.
15-0	The writer handles the topic completely insufficiently.	Unconnected sentences or task not performed.

Table 12: My translation from the original Finnish rating scale descriptions in the guidelines from years 2002-2011 (linguistic richness and accuracy/expressivity and linguistic flaws).

Score (out of 99)	**Criteria: linguistic richness and accuracy (guidelines 2007-2011)**	**Criteria: expressivity and linguistic flaws (guidelines 2002-2005)**
99-90	The writer uses very rich and diverse, idiomatic, appropriate expressions and manages it very well.	Rich, idiomatic and diverse. Some mistakes.
88-80	The writer uses rich, diverse, appropriate expressions and manages it well.	Vocabulary is appropriate but not very diverse, fairly diverse structures. Some mistakes and inauthentic expressions.
78-70	The writer uses sufficient, common, mostly appropriate expressions and manages it relatively well.	Unvaried structures and vocabulary. Manages basic structures. A greater number of mistakes and inauthentic expressions.
68-60	The writer uses quite narrow, highly frequent expressions which are (possibly) only partly appropriate. S/he makes several mistakes.	Unvaried structures and vocabulary. Clear difficulties to produce foreign language. Some mistakes even in the basic structures. Interference.
58-50	The writer uses narrow, unvaried expressions and makes a lot of mistakes.	Basic and one-sided. Quite a lot of mistakes even in the basic structures. Disturbing interference.
48-40	The writer uses very narrow, simple expressions and makes a lot of mistakes.	A lot of mistakes.
35-20	The writer uses rudimentary expressions and mostly incorrectly.	A lot of mistakes.
15-0	The writer uses very rudimentary expressions and almost completely incorrectly.	A lot of mistakes.

Appendix 2

Table A (1/6): Titles and features of task prompts in the exams from autumn 2003 to autumn 2013.

EXAM	**TASK**	**FEAURES**					
TIME	TITLE	T.TYPE	CONTEXT	ARG	DATA	THEME	SUPPORT
A13	1. Calling All Inventors!	Description	Medium		Initiative: website	Innovation	Content
	2. Cultural Awareness	Report	Community	Y	None	Cultural awareness	Content
	3. Share What You're Reading!	Description	Medium	Y	Initiative: website	Books	Content
	4. Being European	Description	Medium		None	European values	Questions
S13	1. Climate change – reality or myth?	Exposition	Medium	Y	None	Climate change	Question
	2. Let bygones be bygones?	Exposition	Medium		None	National history	Questions
	3. Dear X	Recount	Community		None	Personal life	Content, clarification
	4. When angry, count [to] four; when very angry, swear.	Description	Not specified		Quote: writer	Anger management	Questions, limitation
A12	1. Letter of the month	Exposition	Medium	Y	Initiative: online magazine	Health	Content
	2. A speech	Description	Community		None	Public relations	Examples
	3. Finland and the European Union	Exposition	Medium		None	Euroean Union	Questions
	4. A lucky escape	Not specified	Not specified		None	Not specified	None
S12	1. Chalk and Talk or More Modern Technology?	Exposition	Community		None	Education	Questions, examples, content
	2. My Kind of TV-programmes	Exposition	Medium	Y	None	TV-programmes	Content

Table A (2/6): Titles and features of task prompts in the exams from autumn 2003 to autumn 2013.

TIME	TITLE	T.TYPE	CONTEXT	ARG	DATA	THEME	SUPPORT
	3. "If I ran the country, I'd throw Halloween on the bonfire"	Exposition	Medium	Y	Quote: history magazine	Festivals	Questions, content
	4. American Pet Ownership	Description	Not specified		Figures, statistics	Pets	Content, question, limitation
A11	1. Does Finland need more nuclear power?	Exposition	Community	Y	Content	Nuclear power	Content
	2. Breaking the law?	Exposition	Not specified		None	Ethics	Questions
	3. What is a family?	Description	Medium		None	Family	Questions
	4. Improving safety in traffic	Report	Medium		Diagram, statistics	Traffic safety	Content
S11	1. Eat in, eat out, eat away	Description	Not specified	Y	None	Food	Questions
	2. Money down the drain	Report	Not specified		None	Consumption	Questions
	3. Shopping on Sundays	Exposition	Medium		None	Consumption	Questions
	4. Programming children 24/7	Exposition	Not specified		None	Parenting	Question
A10	1. Travelling – enjoying yourself or risking your life?	Exposition	Medium		None	Travel	Question
	2. Any job is better than no job at all	Exposition	Not specified		None	Work	None
	3. To compete or not to compete?	Exposition	Not specified		None	Competition	Questions
	4. Speech	Description	Community		None	Tourism	Content
S10	1. Dear Fellow Europeans	Description	Community		None	European values	None
	2. No way!	Exposition	Medium		None	Environment	None
	3. Can one person change the world for the better?	Exposition	Not specified	Y	None	Ethics	None

Table A (3/6): Titles and features of task prompts in the exams from autumn 2003 to autumn 2013.

TIME	TITLE	T.TYPE	CONTEXT	ARG	DATA	THEME	SUPPORT
	4. James Bond – a hero of the past?	Description	Medium		None	James Bond	None
A09	1. Seeking Notoriety on the Net	Exposition	Medium		Extract: news magazine	Online behavior	Clarification
	2. My problem and how to solve it	Description	Medium		None	Problem solving	Question
	3. Dancing – I just love it! / Dancing – not for me, thank you!	Description	Not specified		None	Dance	Question, clarification
	4. The most important gadget of modern life	Description	Not specified		Poll, statistics	Technology	None
S09	1. What am I?	Description	Not specified		None	Science	Questions
	2. Living in the past, living in the future	Description	Not specified		None	Role playing	Questions, examples
	3. A speech	Recount	Community		None	Anniversary	None
	4. Why extreme? How extreme?	Exposition	Not specified	Y	None	Extreme sports	Questions
A08	1. A speech (2 alternatives)	Exposition	Community	Y	None	Food	Content, clarification, conventions
	2. Opinion and advice	Report	Medium		Extract: newspaper	Charity	Content
	3. Hard values, soft values	Exposition	Not specified		None	Values	Questions
	4. An interesting event or period in history	Description	Not specified	Y	None	History	Questions
S08	1. My crazy festival	Recount	Not specified		None	Festival	Examples, clarification
	2. The right voting age	Exposition	Not specified	Y	None	Voting age	Question

Table A (4/6): Titles and features of task prompts in the exams from autumn 2003 to autumn 2013.

TIME	TITLE	T.TYPE	CONTEXT	ARG	DATA	THEME	SUPPORT
	3. Saving my town	Report	Medium		None	Environment	Question, clarification (bold)
	4. In praise of reading	Description	Medium		None	Reading	Content
A07	1. I wish I had the guts to interfere…	Recount	Not specified		None	Personal experience	Content
	2. The lost boys	Exposition	Not speficied		None	Gender differences in education	Questions, limitation
	3. Could our society do without the police?	Exposition	Not specified		Reading comprehension	Police forces	Questions
	4. Speech	Description	Community		None	Charity	Content, examples
S07	1. Why do tabloids sell?	Description	Not specified		None	Evening papers	Question
	2. What should society protect?	Exposition	Not specified	Y	None	Values	Question
	3. China - my new home country?	Description	Community		None	Working in China	Questions
	4. The art of gift-giving	Description	Not specified		None	Gift-giving	Question
A06	1. Advertising and children	Exposition	Not specified		None	Ethics	Questions
	2. Big brother is watching you	Exposition	Not specified	Y	None	Advertisement	Questions
	3. Being a grown-up	Exposition	Not specified		None	Developmental psychology	Questions
	4. [none - give your own title]	Exposition	Not specified		Reading comprehension	Innovation	Question, limitations

Table A (5/6): Titles and features of task prompts in the exams from autumn 2003 to autumn 2013.

TIME	TITLE	T.TYPE	CONTEXT	ARG	DATA	THEME	SUPPORT
S06	1. Safety on the roads	Report	Not specified		None	Traffic safety	Question
	2. Girls meet boys, boys meet girls	Description	Not specified		None	Relationships	Question
	3. The gadget of the future	Description	Not specified		None	Innovation	Question, examples
	4. What sort of army?	Exposition	Not specified	Y	None	Military service	Question
A05	1. Art and us	Exposition	Not specified	Y	None	Art	Examples, questions
	2. Spectator sports	Exposition	Not specified	Y	None	Spectator sports	Questions
	3. Mind your manners	Exposition	Not specified	Y	None	Manners	Questions
	4. Globalization	Exposition	Not specified	Y	None	Globalization	Content
S05	1. Useless PE lessons	Exposition	Not specified		Extract: newspaper	Physical education	Content
	2. Me - a humanitarian worker?	Description	Not specified	Y	None	Humanitarian work	Questions
	3. Euro elections	Exposition	Not specified		None	Euro elections	Question
	4. Finland will fall when Nokia falls	Exposition	Not specified		None	Finnish economy	Question
A04	1. Strange ways of making money	Description	Not specified		None	Work	Question
	2. Revenge	Recount	Not specified		None	Revenge	Question
	3. Monopoly and competition	Exposition	Not specified		None	Economy	Questions
	4. I witnessed history	Description	Not specified		None	History	Content, question
S04	1. What are taxes for?	Not specified	Not specified		None	Taxation	None

Table A (6/6): Titles and features of task prompts in the exams from autumn 2003 to autumn 2013.

TIME	TITLE	T.TYPE	CONTEXT	ARG	DATA	THEME	SUPPORT
	2. The future - our responsibility	Exposition	Not specified		None	Nature conservation	Questions, limitation
	3. In praise of a great Finnish invention	Description	Community		None	Innovation	Content
	4. Our values	Not specified	Not specified		None	Values	None
A03	1. Could I change this gift?	Recount	Community	Y	None	Product return	Content, conventions
	2. Curing criminals	Exposition	Not specified		None	Crime	Questions
	3. Today's pressures	Description	Not specified		None	Pressure management	Questions
	4. Speech	Description	Community	Y	None	Cultural competition	Content, examples

*Table B (1/5): Markings and other features in the essays.***

Exam	Task	No	Score	Word	Paragr.	sty	Str	ort	pun	lex	Title	Other
A 13	3.	105	78/**78**	162	3	Y	Y	Y	Y		changed	
		104	62/**68**	173	4		Y	Y	Y		OK	unclear meaning
		103	62/**60**	220	3+sal.	Y	Y	Y			missing	
		102	58/**60**	248	4	Y	Y	Y		Y	OK	
		101	55/**58**	204	3	Y	Y	Y	Y	Y	incomplete	
S 13	4.	100	58/**58**	158	3	Y	Y	Y	Y	Y	missing	’/, at the end
		99	82/**80**	258	3	Y	Y		Y	Y	incomplete	
		98	72/**72**	285	4	Y	Y	Y		Y	OK	unclear spelling
		97	40/48/**48**	177	4	Y	Y	Y	Y	Y	incomplete	message apprehensible
		96	70/**70**	157	4	Y	Y			Y	OK	
A 12	1.	95	82/**80**	290	4+sal.+sig.	Y	Y	Y		Y	OK	
		94	60/**62**	197	5+sig.	Y	Y	Y		Y	incomplete	
		93	88/**90**	216	5		Y				OK	no signature = -2p
		92	80/~~83~~, **80**	164	3+sal.	Y	Y	Y	Y	Y	missing	no signature = -2p
		91	75/**75**	198	3+sal.+sig.	Y	Y	Y		Y	OK	
S 12	2.	90	68/**68**	239	4+sal.+date+sig.	Y	Y	Y	Y	Y	OK	
		89	70/70-72/**70**	236	4+sal.+sig.	Y	Y	Y	Y	Y	missing = -5p	unclear meaning, no surname in sig.
		88	40/**43**	156	1+sal.	Y	Y	Y	Y	Y	OK	unclear spelling, no sig. = -2p
		87	57/**57**	160	4+sig.	Y	Y	Y			changed = -5p	

Table B (2/5): Markings and other features in the essays.

Exam	Task	No	Score	Word	Paragr.	sty	str	ort	pun	lex	Title	Other
		86	56/**56**	163	5+sal.	Y	Y	Y	Y	Y	OK	signature missing = -2p
A 11	3.	85	70/65/**65**	211	5	Y	Y	Y	Y	Y	incomplete = -2p	(comment illegible) = -5p
		84	92/**92**	248	4+sig.+date		Y	Y			OK	
		83	80/**80**	262	4	Y	Y	Y	Y	Y	OK	
		82	82/**82**	260	3+sig.	Y	Y	Y	Y	Y	OK	
		81	82/**80**	252	6	Y	Y	Y		Y	OK	
S 11	3.	80	90/**90**	243	5+sal.+date+sig.	Y	Y				OK	
		79	85/**82**	265	4+date+sal.(in paragr.)+sig.		Y	Y		Y	OK	
		78	63/**68**	210	5+add.+sal.+sig.	Y	Y		Y	Y	incomplete = -2p	
		77	88/**85**	276	4+sal.+sig.	Y	Y	Y	Y	Y	OK (no number)	
		76	75/**75**	203	3+date+sal.+sig.	Y	Y	Y		Y	OK	
A 10	2.	75	75/**75**	245	4		Y	Y		Y	OK	
		74	82/**82**	215	4		Y	Y			OK	
		73	75/**75**	180	1	Y	Y	Y	Y	Y	OK	unclear spelling
		72	70/**68**	156	3	Y	Y	Y	Y	Y	OK	weaknesses in vocabulary
		71	95/**92**	259	3	Y	Y				OK	
S 10	2.	70	75/**75**	174	5+sig.	Y	Y	Y	Y		OK	
		69	58/**55**	165	4	Y	Y	Y	Y	Y	OK	unclear meaning
		68	62/**60**	150	4+sig.	Y	Y	Y			incomplete = -2p	unclear meaning

Table B (3/5): Markings and other features in the essays.

Exam	Task	No	Score	Word	Paragr.	sty	Str	ort	pun	lex	Title	Other
		67	48/**45**	151	3+sig.	?	Y	Y	Y	Y	OK	unclear meaning, only 1 correct sentence (+)
		66	50/**52**	233	4+sig.		Y	Y	Y		OK	
A 09	3.	65	82/**85**	247	4	Y	Y	Y	Y		OK	
		64	62/**62**	204	4	Y	Y	Y	Y	Y	OK	unclear meaning
		63	88/**88**	210	4		Y	Y	Y		OK	
		62	92/**90**	192	3	Y	Y	Y		Y	OK	
		61	88/**88**	241	5		Y	Y	Y	Y	OK	
S 09	4.	60	82/**82**	220	4		Y	Y	Y		OK	
		59	55/**52**	265	5	Y	Y	Y	Y	Y	OK	
		58	72/**75**	229	4	Y	Y	Y	Y	Y	OK	
		57	85/**85**	231	5	Y	Y	Y	Y	Y	OK	
		56	70/**72**	232	3		Y	Y		Y	OK	
A 08	4.	55	80/**82**	238	3		Y	Y	Y	Y	OK	factual error, repetition, + +
		54	78/**75**	193	3	Y	Y	Y	Y	Y	OK	Repetitious
		53	65/**60**	170	5	Y	Y			Y	OK	repetitious
		52	75/**70**	192	3	Y	Y	Y	Y	Y	OK	
		51	85/**85**	214	5	Y	Y		Y	Y	OK	
S 08	1.	50	72/**70**	180	4		Y	Y	Y	Y	OK	
		49	85/**85**	221	3		Y	Y	Y	Y	OK	
		48	92/**92**	271	5	Y	Y		Y	Y	OK	syllabication
		47	76/**68**	233	4		Y	Y			OK	off topic = -10p
		46	65/**55**	202	4	Y	Y	Y	Y	Y	OK	off topic = -10p
A 07	1.	45	75/**75**	212	5		Y	Y	Y	Y	OK	
		44	92/**90**	250	5		Y		Y		OK	

Table B (4/5): Markings and other features in the essays.

Exam	Task	No	Score	Word	Paragr.	sty	Str	Ort	pun	lex	Title	Other
		43	80/**80**	241	3	Y	Y	Y	Y	Y	OK	
		42	75/**72**	217	8?	Y	Y	Y		Y	OK	
		41	92/**95**	233	6		Y	Y	Y	Y	OK	
S 07	2.	40	80/**82**	168	4	Y	Y	Y	Y		OK	
		39	55/**55**	166	1	Y	Y	Y	Y		OK	unclear meaning
		38	78/**78**	234	5	Y	Y		Y	Y	OK	
		37	86-80/70-60/**60**	250	2	Y	Y	Y	Y	Y	OK	off topic = -20p
		36	97/**97**	250	3	Y	Y	Y			OK	
A 06	3.	35	75/**75**	221	4	Y	Y	Y		Y	OK	
		34	90/**90**	262	5	Y	Y			Y	OK	
		33	72/~~68~~, **70**	356	6	Y	Y	Y	Y	Y	OK	unclear meaning
		32	50/**50**	225	4	Y	Y	Y		Y	OK	unclear meaning
		31	52/**52**	195	4	Y	Y	Y		Y	OK	unclear meaning
S 06	4.	30	70/**68**	178	4	Y	Y	Y	Y	Y	OK	unclear meaning
		29	92/**90**	219	3		Y	Y	Y		incomplete (-2p)	
		28	70/**70**	168	1	Y	Y	Y		Y	OK	
		27	82/**85**	251	5		Y	Y	Y	Y	OK	
		26	60/**60**	174	4	Y	Y	Y		Y	incomplete = -2p	
A 05	3.	25	78/82-85/**82**	234	3	Y	Y	Y			OK	
		24	85/**88**	245	4	Y	Y	Y			OK	
		23	82/**82**	263	3		Y	Y			OK	
		22	52/62/**62**	149	3		Y	Y	Y	Y	OK	
		21	88/**88**	211	3	Y	Y		Y		OK	

Table B (5/5): Markings and other features in the essays.

Exam	Task	No	Score	Word	Paragr.	sty	Str	ort	pun	lex	Title	Other
S 05	2.	20	45/42-45/**45**	214	5	Y	Y	Y	Y		OK	unclear spelling & meaning
		19	88/82/**82**	216	3	Y	Y	Y		Y	OK	
		18	68/**63**	182	3	Y	Y	Y		Y	incomplete = -2p	
		17	80/75/**75**	214	4	Y	Y	Y	Y		OK	
		16	45, ~~35~~/40-35/**40**	160	3		Y	Y		Y	OK	
A 04	1.	15	70/**70**	192	3		Y	Y		Y	OK	
		14	62/**68**	178	3	Y	Y	Y			OK	
		13	85/80-78/**78**	195	4	Y	Y	Y		Y	OK	incomplete main clause
		12	62/**62**	149	4		Y	Y			OK	
		11	72/**72**	174	3	Y	Y	Y	Y	Y	OK	unclear spelling
S 04	4.	10	85/**85**	194	5	Y	Y	Y	Y		OK	
		9	88/**88**	189	3	Y	Y	Y	Y		OK	
		8	80/**80**	164	4	Y	Y		Y		OK	unclear meaning
		7	95/97-99/**99**	227	4+footnote		Y	Y			OK	
		6	52/58/**57**	165	5		Y	Y	Y	Y	OK	
A 03	3.	5	80/**85**	170	3	Y	Y	Y	Y	Y	OK	
		4	45/**45**	197	5	Y	Y	Y	Y	Y	OK	
		3	90/**88**	210	3	Y	Y	Y	Y		OK	
		2	65/**68**	167	4	Y	Y	Y		Y	OK	
		1	68/**68**	193	3	Y	Y	Y	Y	Y	OK	

**Explanatios of abbreviations and headings:

Exam = time of the exam

Task = number of the task prompt

No = number of the essay

Score = score of the essay

Word = number of words in the essay

Paragr. = number of paragraphs in the essay

str = structural errors

ort = orthographic errors

pun = punctuation errors

sty = stylistic errors

lex = lexical errors

Title = remarks on the title in the essay

Other = other comments in the essay written by the raters

sig. = signature

sal. = salutation

add. = address

Appendix 3

The objectives of the target level B2.1 for advanced English writing in the National Core Curriculum for General Upper Secondary Education Intended for Young People (2003):

”• Can write clear and detailed texts about a variety of areas of personal interest and about familiar abstract topics, and routine factual messages and more formal social messages (reviews, business letters, instructions, applications, summaries).

• Can express information and views effectively in writing and comment on those of others. Can combine or summarize information from different sources in his/her own texts.

• Can use broad vocabulary and demanding sentence structures together with linguistic means to produce a clear, cohesive text. Flexibility of nuance and style is limited and there may be some jumps from one idea to another in a long contribution.

• Has a fairly good command of orthography, grammar and punctuation and errors do not lead to misunderstandings. Contributions may reveal mother tongue influences. Demanding structures and flexibility of expression and style cause problems.”