universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„The relationship between item formulation and item facility in multiple-choice reading tasks “

verfasst von / submitted by

Mag. Gabriel Kummenecker, MA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Philosophie (Mag.phil.)

Wien, 2020 / Vienna, 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 190 344 350
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Lehramtsstudium UniStG UF Englisch UF Italienisch
Betreut von / Supervisor:	Univ.-Prof.in Mag.a Dr.in Christiane Dalton-Puffer
Mitbetreut von / Co-Supervisor:	Mag. Dr. Armin Berger, MA

Im Danke meiner Familie gewidmet.

Danksagung

Zunächst möchte ich mich bei Frau Univ.-Prof.in Mag.a Dr.in Christiane Dalton-Puffer und Herrn Mag. Dr. Armin Berger, MA für die stets motivierende Betreuung der Diplomarbeitsentwicklung und ihre vielen Ratschläge und sehr detaillierten Rückmeldungen bedanken. Im Zuge dieser Studienabschlussarbeit konnte ich sowohl meine quantitativen Forschungsfähigkeiten als auch meine fachwissenschaftlichen Kompetenzen im Bereich der Sprachentestung erheblich erweitern, was für mich auch beruflich von Vorteil ist.

Beim *Finnish Matriculation Board* möchte ich mich dafür bedanken, dass mir die Resultate von fünf finnischen Maturaterminen inklusive Übersetzungen finnischer Ausdrücke zugeschickt wurden. Somit hatte ich einen Datensatz, anhand dessen ich meine Forschungsfrage bearbeiten konnte.

Last but not least, möchte ich meiner Familie für ihre kontinuierliche Unterstützung danken, welche mir einen erfolgreichen Bildungsweg ermöglichte.

Thank you!

Table of contents

1 Introduction

It occurs that the items of reading tests are formulated in a complex way. This raises the question of whether the comprehension of the items might become an obstacle for examinees in testing situations. When the item is not understood, it cannot be solved. It can be assumed that the reading of the questions should not cause any comprehension difficulties but rather enable the examinees to show their reading competence with regard to the texts. As items are, however, sometimes formulated in a complex way, the research question of this project is to find out about the relationship between item formulations and the facility values of items. This might offer significant insights for item writers.

A colleague working for the *Finnish Matriculation Examination*, which is the Finnish school-leaving exam, offered the facility values of six exam dates of the multiple-choice items of their reading part of the exam for English as a foreign language. This data could be used for correlating the facility value with different features, such as lexical difficulty, of the item formulations of the *Finnish Matriculation Examination*.

At first, as many factors as possible leading to different facility values of the items had to be excluded in order to be able to focus on the relationship between the item formulation and the facility value. Finally, 76 MISD items testing text passages with a similar readability index could be identified.

With this corpus of comparable items, it could then be analysed to which extent the wording of the items correlates with the facility values. For this, as many features as possible concerning the item formulations were defined. These 22 features refer to lexical aspects as well as to the length of the items and clauses.

Then these different features were correlated in *SPSS* with the facility values of the items. The results showed the extent of the relationship between the formulations of the items and their facility values. Only three features showed significant correlation coefficients, two of which dealt with lexical difficulty in the items.

This introduction is followed by the theoretical background concerning language tests and in particular testing reading and multiple-choice items. This theoretical section also deals with a similar study by Freedle and Kostin and the *Common European Framework of Reference for Languages (CEFR).* This is followed by a chapter dealing with the context of the study, which presents the *Finnish Matriculation*

Examination, the research question as well as the hypothesises concerning the outcomes of this research project. This is followed by the methodology of this research project. At first, indicators of the quality of quantitative research are dealt with. Then it is explained how the 76 reading multiple-choice items were selected for this study. After this, the calculation of correlations is presented. This is followed by the definition of the chosen features of the item formulations. Then it is explained how the data was transferred from *Excel* into *SPSS* in order to calculate the correlations. This is followed by a chapter about the results of the correlations, which are also presented in a table. After this, the results of this study are discussed by making links back to the similar study as well as the theoretical background. Finally, the most important points of this research project are summarized in the conclusion. The items as well as all the tables consisting of the calculations of the different features of the chosen 76 items can be found in the appendices.

2 Theoretical background

2.1. Language tests

2.1.1. Different types of language tests

There exist different categories for classifying tests, which are presented in this section. It will also be addressed into which of these categories the reading multiple-choice tasks of the present study fall.

Generally, there exist two different types of tests. Firstly, traditional paper and pencil tests are typically used for testing isolated language components (e.g. vocabulary) or the receptive skills of reading and listening. In professionally designed standardized exams, the test items often come in fixed response formats, where examinees are asked to choose from several possible responses. The most frequent of these formats is multiple-choice. Hence, the reading component with the multiple-choice items of the present study can be classified as a paper and pencil test. Secondly, there are performance tests which measure the examinees' skills in a communicative act, usually in writing or speaking (McNamara 2000: 5-64).

Concerning the test purpose, a distinction can be made between proficiency and achievement tests. Proficiency tests usually measure the acquired competence of a language, for example. Several proficiency tests are standardised such as the American TOEFL test, the British Cambridge English Examinations or the British-Australian IELTS test, which non-native English speakers have to take in order to study in the

USA, the UK or Australia (Brown, Davies, Elder, Hill, Lumley, McNamara 1999; 154). Furthermore, school-leaving examinations are also sometimes standardised. For instance, in Austria the standardised *Matura* became obligatory for all examinees in 2016. In Austria this examination is a prerequisite in order to be allowed to move on to tertiary education. In contrast, achievement tests look to the past, for instance after a course, in order to measure what students have learnt and to see to what extent the learning goals have been reached (McNamara 2000: 6-7).

The *Finnish Matriculation Examination* is a hybrid between a proficiency test and an achievement test. Concerning the latter, the aim of this exam is to measure whether the goals of the curriculum were reached (Ylioppilastutkintolautakunta – studentexamensnamnden a: n.d.). Consequently, it can be argued that the *Finnish Matriculation Examination* is an achievement test. However, at the same time, this exam provides information concerning the level which was reached in the target language, which is why it can also be seen as a proficiency test.

Nowadays it is state-of-the-art to have standardized tests for important exams, which have far-reaching consequences for the learners. This is why such tests are classified as *high-stakes tests* in contrast to low-stakes tests, which are less important for the examinees' lives (Kecker et al.: 2019: 393). Examples of such *high-stakes tests* are the *Cambridge English Language Assessment* or also school-leaving examinations such as the *Finnish Matriculation Examination*. These *high-stakes examinations* are often a prerequisite for the entry to certain educational institutions.

2.1.2. Quality factors in testing

Essential quality factors for tests are validity, reliability and objectivity (Hinger & Stadler 2018: 40). These aspects should be taken into account when producing standardized exams and consequently also when formulating multiple-choice items.

Validity is the most important concept with regard to testing (Fulcher & Davidson 2007: xix) and analyses to what extent an exam actually measures what it says it does. In other words, validity stands for the relationship between the performance of examinees in a test and the conclusions which can be drawn concerning the examinees' competence. Since the 1970s, a broader concept of validity has been discussed. Previously, social consequences of tests had been seen as a separate concept and were classified under policy issues (Davis & Xi 2016: 62). Later, researchers such as Messick (1989, quoted in Davis & Xi 2016: 62-63) have argued

that the social impact of tests is an important aspect within the concept of validity. The debate goes on whether the consequences of tests are part of validity (Davis & Xi 2016: 64).

It has been argued that it is import to think about the context of an examination when dealing with validity. Concerning important standardized exams the focus needs to be more on the interpretation of the results and the use of the assessment while the local context is less relevant. It needs to be taken into consideration that assessment in the context of a classroom differs considerably from a *high-stakes examination* (Moss 2003, Moss 2013, quoted in Davis & Xi 2016: 65), which the *Finnish Matriculation Examination* is.

The most frequently cited types of validity are content, construct and criterion-related validity. While the last one is statistical, the first two are conceptual (Davies et al. 1999: 221).

- Content validity refers to the domain which should be tested (Davies et al. 1999: 222). This might, for instance, be an occupational domain or the language as a whole and its possible uses. Concerning general proficiency tests, where the whole language represents the target, the content becomes identical with the construct (Davies et al. 1999: 34). This is the case of the proficiency test this study has its focus on, the *Finnish Matriculation* Examination, which tests General English and does not focus on a specific domain. Concerning content validity, items need to be chosen which represent the domain. This is done by testing professionals and might require a needs analysis (Davies et al. 1999: 222) about the situations in which the examinees need which aspects of the target language. In order to represent General English, short text of a wide range of fields are being tested in the reading part of the *Finnish Matriculation Examination.*
- Construct validity deals with the quality of a test with regard to the theoretical model it is based on (Davies et al. 1999: 222). More recent views of construct validity take a broader variety of testing factors into consideration such as times and settings, the investigation of the behaviour of raters and candidates or the differences in the performance across different groups (Davies et al. 1999: 33).

- Criterion-related validity is calculated statistically with regard to how close a test is to its criterion (Davies et al. 1999: 222), which is an external variable such as another test, a syllabus, a performance in the real word or the judgement of a teacher (Davies et al. 1999: 37). When the criterion is another existing test or some other form of measurement in the same domain, this is referred to as concurrent validity (Davies et al. 1999: 222).
- Face validity is a further type of validity, which describes to what extent a test seems to measure what it said it does according to untrained observer such as an examinee (Davies et al. 1999: 37, 222). For instance, when just taking a glance at a task of Business English, the person should already get the impression, for instance by a picture and/or the title, that this task does, indeed, test Business English.

Reliability deals with the consistency concerning the results (McNamara 2000: 136). This refers to the extent of agreement among the results of a test with itself or another examination. Ideally, the agreement would be the same. But one has to take the measurement error into consideration, which can happen because of bias concerning the selection of items, the time of testing or the raters (Brown et al. 1999: 168). Concerning the raters this means whether different raters will come to the same results with regard to the performance of an examinee. As the subject matter of this study are multiple-choice items with a published key regarding the correct answer to each item, it can be assumed that different raters will usually reach the same results. In fact, multiple-choice items can also be scored digitally. Test developers of *high-stakes* objective tests such as the *Finnish Matriculation Examination* have to ensure, for instance, that the selected items measure the same ability and competence level as far as possible at each exam date. Otherwise it would be unfair to the examinees if one exam was easier than another one. In the case of the higher *Finnish Matriculation Examination* the reading items cover a range from B1.2. to C1 and the difficulty of the different items should be as consistent as possible over the different exam dates

On the one hand, it is possible that a test is reliable but not valid. A test can produce consistently the same results even though it does not measure what it is supposed to. On the other hand, a test is only valid when it is also reliable. If a test

does not measure its construct[1] consistently, the consequence is that the construct cannot always be measured accurately (Alderson, Clapham & Wall 1995: 187).

Another important concept concerning testing is objectivity. This means that during the live administration of tests as well as the rating afterwards, subjectivity should be reduced as far as possible (Hinger & Stadler 2018: 41). If a certain performance always leads to the same score, it is objective. Closed test methods such as multiple-choice do not cause any difficulty in this regard. In open test methods, for instance, when testing speaking or writing, objective assessment is much more difficult to achieve as people decide on the score with the help of the descriptions for the target levels (Glaboniat & Peresich 2018: 357-358). Multiple-choice items are referred to as objective, because they ask for a constructed response. The correct answer can be chosen out of several provided answers for each item.

Typically, an objective item consists of a stem, which addresses the problem, and of several choices. Multiple-choice items consist of at least three options (Brown et al. 1999: 132). In the analyzed reading multiple-choice items of the *Finnish Matriculation Examination* the stem is followed by three options in each of the reading items. Objective items are criticized, for example, because the answers might be guessed by the examinees, which would not test the competence in a language. However, empirical studies have shown that objective items are more reliable, fairer and that they can cover a broad subject-matter (Brown et al. 1999: 132).

The test construct should, ideally, test competences which the examinees might need in the real world (Kecker et al. 2019: 397). Such examinations are usually based on the communicative approach and test the communicative competence of examinees, which refers to their ability to deal with situations in the target language. Apart from the linguistic competence, such tests also consider the strategic, pragmatic as well as the socio-cultural competence of examinees. In fact, language testing takes different areas of applied linguistics into account in order to improve the quality of testing the learners' competences (Fulcher & Davidson 2007: xix). Performance tests try to confront examinees with real-life situations, which is why this approach is described as the real-life approach. Such tests use text sources which are relevant, authentic and close to every-day situations examinees might encounter (Glaboniat & Peresich 2018: 352-353). In the *Finnish Matriculation Examination* a wide range of different reading texts are included as the examinees might have to deal with different text types after

[1] Construct refers to the traits which a test is supposed to measure (Brown et al. 1999: 31).

school. However, it needs to be taken into consideration that the test method of multiple-choice is not a real-life situation. This is because while reading a text people, usually, do not ask themselves questions with four options and then have to find the correct one. By contrast, according to the real-life approach, testing situations should simulate realistic behavior in the target language as closely as possible (Glaboniat 2019: 412). Due to the artificiality of multiple-choice items, this aspect of realistic behavior seems hard to be reflected by using multiple-choice items. Mixing different language skills might be more authentic. For instance, with regard to reading this might include reading an email and then writing an answer to it. The disadvantage of the mixing of skills is that it is hard to measure them separately. Regarding the example just given, if there were misunderstandings while reading the email, this might lead to difficulties in the production of the answer email. In the *Finnish Matriculation Examination* reading is only tested with 25 multiple-choice items and no other test method and no other skill is involved in the testing of reading.

2.2. Testing reading

Within the different language skills, reading, along with listening, is part of the receptive skills, also referred to as *reception activities* in the *CEFR* (Council of Europe 2018: 55). Reading can only be tested indirectly by asking examinees, for instance, to complete sentences or tick boxes as receptive competences cannot be observed. On the contrary, writing, speaking and listening in interaction can be tested directly (Council of Europe 2001: 187).

In contrast to reading in the first language, the reading processes in second and foreign languages differ due to a variety of factors. For example, L2 readers have fewer linguistic resources and do not know the socio-cultural context of the target language as well as L1 speakers do (Grabe & Stoller 2002: 62-63). As the languages are stored together in the brain, it is likely that examinees also refer to their L1 reading competence when reading in the L1 (Neuner 2003: 17). This might, for example, be the case when guessing unknown words or for recognizing text types.

When reading in a foreign language, L2-knowledge is more crucial than the ability of reading in the L1. L2 readers need to know a certain amount of words, before the ability of reading in the L1 can cross over to the L2. If a task is more challenging, more vocabulary is necessary for understanding it (Alderson 2000: 39). As vocabulary plays an important part in the reading process, tests should be analysed regarding their

vocabulary and with regard to extreme lexical difficulty (Alderson 2000: 83). When reading takes place in the L2, the linguistic knowledge includes the one in the L1 as well as the relationship between the L2 and the L1 at all linguistic levels (Alderson 2000: 81).

Reading is not a passive competence, but consists of complex and dynamic processes, which are going on during the reading process (Hinger & Stadler 2018: 76). For reading successfully, text competence needs to be developed. This is especially important for texts including language usage which differs from spoken language (Portmann-Teslikas und Schmölzer-Eibinger 2008: 6). Text competence includes knowledge of different text types, the conventions of their use within a particular cultural context as well as an awareness of the differences between written and spoken language usage. Moreover, comparisons between L1, L2 and Lx text conventions might take place as well when dealing with texts (Krumm 2007: 202). Skilled readers use both bottom-up as well as top-down processes simultaneously while reading. The former refers to decoding smaller elements when constructing meaning, such as the lexical component. On the other hand, top-down processes are about background knowledge which people use while reading (Neuland & Peschel 2013: 163). Hence, top-down processes are based on hypothesis and mental concepts, while bottom up processes make use of the language presented (Hinger & Stadler 2018: 69). The different processes going on while reading interact with each other (Neuland & Peschel 2013: 162).

Different aspects concerning the reader influence the reading process as well. There are stable factors, such as the readers' age, sex or personality, and physical aspects, such as the speed of the recognition of lexical items, the automaticity of processing or eye movement. Another factor is a reader's motivation for reading (Alderson 2000: 32-33). The motivation for reading of the examinees during a test situation can be described as external motivation as they would like to pass the test, in contrast to internal motivation for reading a text in their spare time. Further aspects concerning the reader include their knowledge of the subject matter as well as their world and cultural knowledge (Alderson 2000: 39-48).

2.2.1. Reading styles

Regarding listening, Green (2017: 55-81) discusses three different listening styles, which can also be applied to reading tasks and are used for both of these skills in the

standardized Austrian *Matura*[2], for example. Her concepts were used for this analysis as they form the basis for her further recommendations concerning the development of items for language tests. Furthermore, it seemed useful to pick out items focusing only on one of the three reading styles in order to work with similar items. In the following, these three different kinds of reading behaviour are presented. Firstly, reading for gist refers to understanding the overall idea of a text. Secondly, the aim can be to search for specific information or important details (SIID) in a text, for which readers use selective reading. This includes, for instance, names, locations or numbers. The third aspect is reading for main ideas and supporting details (MISD). In contrast to SIIDs, MISDs usually include structures with a verb (e.g. to go to the airport). According to Khalifa and Weir's cognitive model of reading, the goal setter decides on the goal of reading and chooses a type of reading in order to achieve this goal (Khalifa & Weir 2009: 55-56). For the present analysis, only MISD items were taken into consideration, which will be explained in more detail in the section 4.4.2. in the methodology chapter below.

2.2.2. A framework for conceptualising reading test validity

Khalifa & Weir (2009: 3-8) developed a theoretical framework for validating exams of foreign language reading competence. Their framework is often referred to when discussing testing issues concerning reading. *High-stakes tests* need to show how the concept of validity is met in their tests. This framework helps to establish the validity of an exam and the model can be seen in *figure 1* below (Khalifa & Weir 2009: 5).

In this framework, the following main factors concerning the construct of reading are being taken into consideration: context, test-taker characteristics as well as scoring. While the first two aspects refer to the time before the actual testing, the component of scoring takes place after the exam (Khalifa & Weir 2009: 5-8).

[2] The Austrian A-levels, which are passed when leaving school before attending university.

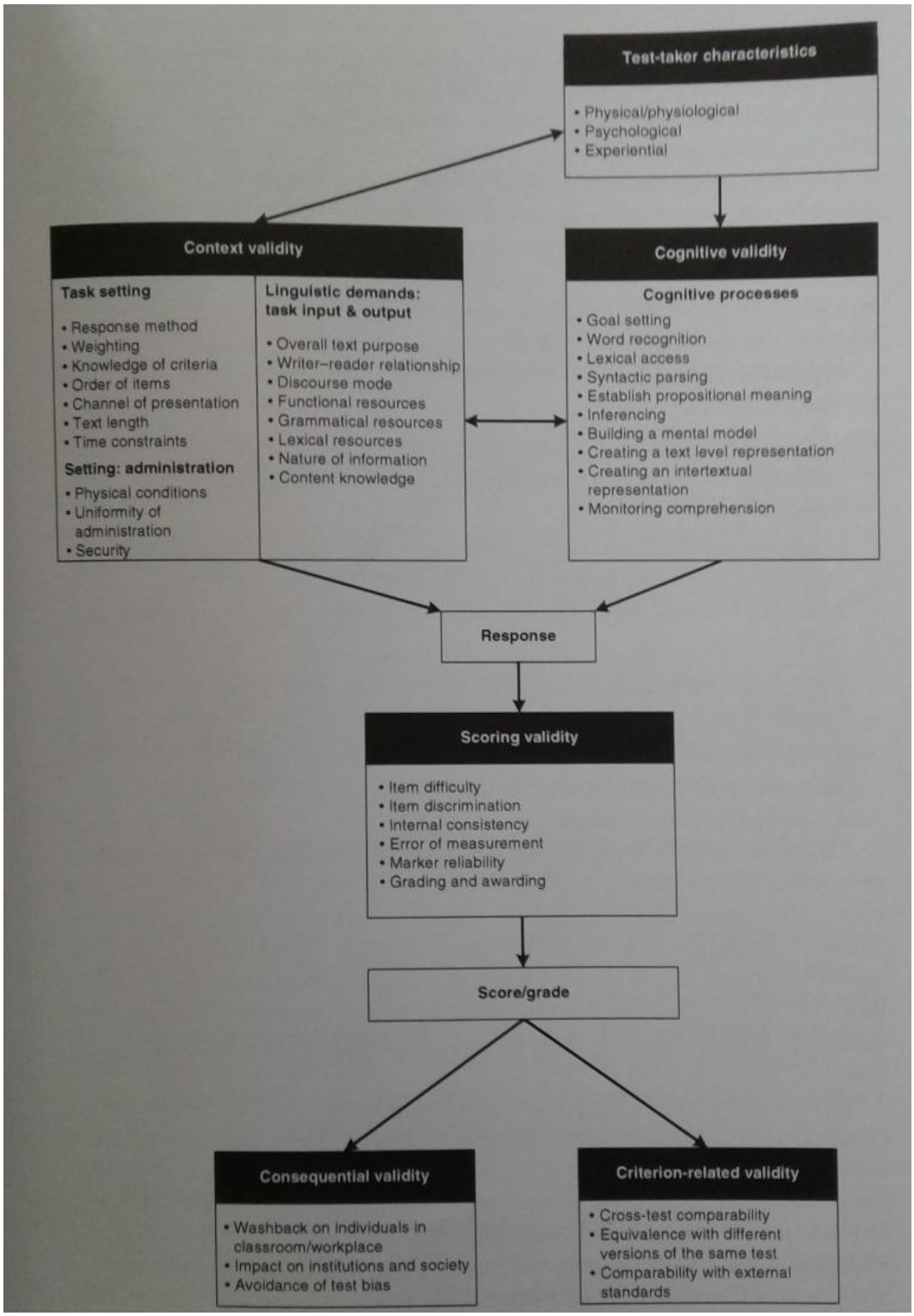

Figure 1: "A framework for conceptualizing reading test validity" (Khalifa & Weir 2009: 5

The *test-taker characteristics* at the top of *figure 1* can be divided further into three different subcategories, which might affect the performance of examinees.

1. The *physical/psychological characteristics* refer to special needs such as dyslexia.
2. The *psychological characteristics* include aspects such as a candidate's motivation or personality type, which might have an effect on the performance.
3. The *experiential characteristics* refer to an examinee's cultural as well as educational background, for instance, to their familiarity with the test (Khalifa & Weir: 2009: 6).

For this thesis, particularly aspects two and three seem relevant. An examinee's motivation is probably quite high during the *Finnish Matriculation Exam*, so that examinees will try to perform as well as possible. Regarding the third point, it can be assumed that the examinees are familiar with the cultural educational setting including the task types as the examinees are prepared for the exam within the Finnish school system.

2.2.2.1. *Context validity*

In the following, the other sub-categories of the model in *figure 1* are also presented. These different concepts are linked with each other as are the different aspects within each one of these concepts. *Cognitive, context* and *scoring validity* are often referred to as construct validity (Khalifa & Weir 2009: 5-8). These three concepts are presented in more detail in the following chapters.

In Khalifa and Weir's model the concept of *context validity* in *figure 1*, includes the linguistic content as well as the cultural and social context in which the exam is set. Therefore, this concept refers to aspect such as the response method chosen, text length and available time during exams. *Context validity* is divided into two subcategories: "Task setting" and "Linguistic demands: task input & output". An aspect that might be taken into consideration for this thesis from the subcategory "task setting" includes the response method (Khalifa & Weir 2009: 5-7). In the context of the present study, the analysed response method are multiple-choice questions. From the subcategory "Linguistic demands: task input & output" the following aspects might be considered for this research project, because the examinees needed them in particular when answering the items before the facility values were calculated. These

aspects include functional, grammatical, lexical resources, nature of information and content knowledge. *Context validity* also includes aspects concerning the reading text itself such as the overall text purpose, the writer- reader relationship, the content knowledge as well as linguistic knowledge, for instance with regard to grammatical and lexical resources (Khalifa & Weir 2009: 5-7).

2.2.2.2. *Cognitive validity*

The *cognitive validity* in *figure 1* refers to the extent to which the reading test elicits the cognitive processes which would also be found in real life reading outside of the testing situation (Khalifa & Weir 2009: 34). *Cognitive validity* includes aspects such as *inferencing*. This means that readers use background knowledge to understand information which is not explicitly stated in the text. *Inferencing* is an important strategy used by learners for coping with unknown words and expressions. This strategy is particularly important for this study because when examinees are unsure about the meaning of an expression within the item, this might make the item more difficult, although the aim is actually to test a reading passage of a text. This is why it is assumed that the formulation of the item should, ideally, not cause any difficulties for examinees. Consequently, items should be formulated a level below the target level to ensure that all examinees can understand the multiple-choice item without any problems. In order to have similar items, the ones from the higher English examination were chosen and not from the lower English examination. As there are reading items between B1 and C1 in the higher examination, it could be argued that the formulations of the items should not be higher than the B1 level or even at A2 level for the B1 items. However, if there are lexical items that examinees are not very familiar with, the examinees are assumed to guess the meaning of these unknown words. The present study analyses, among other features, whether more difficult words used in an item, tend to make a question more challenging with a lower facility value.

When guessing unknown words, new information builds on existing knowledge in the brain. However, this previous knowledge that helps readers understand a text differs greatly among readers and different cultures (Wolff 1996: 544-546). Paribakht & Wesche (1999: 199) assume that the text, the features of new words, the learners' previous knowledge and the learners' effort which they put into understanding new words all affect the understanding process of new words. Due to the fact that the knowledge of different languages in the brain is connected, the different languages

interact with each other (Brizic 2006: 38). This means that the Finnish L1 examinees, whose answers led to the facility values for this study, might, for example, make references to Swedish in order to deconstruct the meaning of less familiar English words.

Paribakht & Wesche (1999: 201-213) could gain very interesting data from their "introspective study" regarding the strategies that learners use in order to understand unknown words in a piece of writing they are reading. In all tasks, the vast majority of unknown words (97 - 99 %) were content words (nouns, verbs, adjectives), with nouns making up 60 % of the total number of unknown words. Moreover, there was a great difference in the number of words identified as unknown by the different students. Moreover, Paribakht & Wesche (1999: 201-222) point out that the students made use of three different strategies when searching for the correct meaning of unknown words, which are described below. *Inferencing* was by far the most important way of dealing with unknown words. It made up 80% of all the strategies used by students. Other strategies included rereading words several times, repeating them aloud or asking what a word means. (Paribakht & Wesche 1999: 201-222).

2.2.2.3. *Scoring validity*

Scoring validity needs to be taken into consideration together with *context* and *cognitive ability* and refers to all aspects of reliability (Khalifa & Weir 2009: 7). Within the concept of *storing validity,* there should be as little measurement error as possible and the results should stay the same over time among other aspects (Khalifa & Weir 2009: 5, 7) According to Khalifa & Weir (2009: 7) scoring validity refers to the degree

> to which test scores are arrived at through appropriate criteria in constructed response tasks and exhibit consensual agreement in their marking, are as free as possible from measurement error, stable over time, appropriate in terms of their content sampling and engender confidence as reliable decision-making indicators.

In the case of multiple-choice items in the present study, the results should be consistent over time among different raters as there should only be one correct answer among the three given options of the multiple-choice items of the *Finnish Matriculation Examination*. Consequently, there should be little measurement error. In fact, the analysis of multiple-choice items can also be done automatically by a machine, which can reduce the risk of errors even further. A subcategory which is

particularly relevant for this study is the concept of item facility, which is discussed in subsection 2.3.1.

2.2.2.4. *Consequential validity*

Consequential validity judges the validity of the scores of a test (Khalifa & Weir 2009: 5, 7). According to Khalifa & Weir (2009: 7) "it is necessary in validity studies to ascertain whether the social consequences of test interpretation support the intended testing purpose(s) and are consistent with other social values". This also takes the washback effect of a test into consideration, which refers to the teaching and learning leading up to the exam and the impact that the exam has on institutions as well as on society as a whole (Khalifa & Weir 2009: 7). However, this category does not seem to be very relevant for the present study as it does not consider the consequences of the results of the *Matriculation Exam* for the society and the Finnish educational system.

2.2.2.5. *Criterion-related validity*

Criterion-related validity includes, for instance, the comparison of exam scores with other tests or the linking of test scores to an external, standardized way of measurement such as the *CEFR* (Khalifa & Weir 2009: 5-8). The latter is relevant for this study as the *CEFR* was used by the *Finnish Matriculation Exam* to determine the difficulty of items. This is why the *CEFR* will be discussed in subchapter 2.2.4.

2.2.3. Variables affecting the readability of texts

Indices have been established in order to enable estimates regarding the difficulty of texts. It can be analysed how many words there are per sentence as longer sentences tend to be harder to understand. Moreover, all letters of a given text can be counted to get an idea of its readability. Thereby, both the lexical as well as the syntactic features are being taken into consideration.

In order to determine the difficulty of a text, the readability index by Flesch can be calculated, which is presented in the methodology section in subchapter 4.4.3. However, readability formulae offer only a simple analysis of texts. This is due to the fact that there are several different aspects apart from lexical and syntactic difficulty which affect the difficulty of a text. Further factors include the subject matter as well as the cohesion and coherence of a text. In many situations there are no absolute terms for defining the difficulty of text. Instead, test developers might want to focus on a

variety of authentic texts which the target group might have to read in real-life target language situations (Alderson 2000: 73-74). When using several texts, for example, various topics and writing styles can be tested.

Test developers need to be aware of aspects affecting a test's readability. The difficulty of a text has an effect the results of the examinees when answering the items (Alderson 2000: 83). Only texts should be selected, whose difficulty is appropriate for the target group and the target level.

Other fields such as sociology or communication studies have also conducted valuable research with regard to text variables. The factors analysed include different text types, the content, the text organisation, the syntax, the lexis, the layout, verbal and non-verbal text and the medium of a text. In addition, intrinsic factors with regard to the examinees are responsible for the individual difficulty a text might have for a certain examinee (Alderson 2000: 60-61).

There are further aspects affecting the difficulty of a reading text. Concerning the content of a text, the reader's world knowledge might help to understand it. In addition, not the content itself, but rather the style in which a text is written might contribute to its difficulty. Narrative texts tend to be understood more easily, because of less variety concerning the content among other aspects (Alderson 2000: 63-64).

Regarding the language of a text, a complex lexis and syntax, obviously, contribute to the overall difficulty of understanding a text, although examinees are usually encouraged to guess the meaning of unknown words from the context (Alderson 2000: 68-71). This process of *inferencing* has been described in further detail in subsection 2.2.2.2 dealing with *cognitive validity*. Due to the interaction among lexical, syntactic, topic and discourse variables, none of these aspects can be identified as the most important one (Alderson 2000: 70-71).

2.2.4. Testing reading with the *CEFR*

The *CEFR* was developed under the commission of the *Council of Europe* between 1993 and 1996. Its original aim was to form a common basis for the development of foreign language curricula, teaching course books and exams across Europe. In addition, more transparency should be the outcome with regard to the assessment and certification of language competences (Kecker 2011: 74). Learning and teaching of foreign languages as well as language assessment should be linked closer to a real-life approach (Figueras 2012: 478). More and more centralized standardized tests have

been developed and based on the *CEFR* in order to be compared more easily. An example of these is the *Finnish Matriculation Examination.*

The new version of the *CEFR* (2018) consists of seven different levels from Pre-A1 to C2. The higher the *descriptors* get regarding the language levels, the more the descriptors refer to complex texts, which are less related to a familiar field of interest of the readership. In addition, for the higher levels a more in-depth comprehension of more detailed points as well as implications are required when dealing with texts.

The *CEFR* (Council of Europe 2018: 60-65) offers *scales* for *overall reading comprehension*, *reading correspondence*, *reading for orientation*, *reading for information and argument*, *reading instructions* and *reading as a leisure activity*. During the *Finnish Matriculation Examination* the texts are read by the examinees in order to show their reading competence by getting the necessary information from the short texts to be able to answer the multiple-choice items. This could be included under the *scales* dealing with *reading for information and argument* as the examinees have to locate the desire information in the texts for answering the items.

In the new version of the reading as a leisure activity was also put more focus on. There is a *scale* called "reading as a leisure activity" (Council of Europe 2018: 65) which also refers to literary texts along with reading newspapers, for instance. In the B2 *descriptors* there are clear references to reading literary texts:

> Can read for pleasure with a large degree of independence, adapting style and speed of reading to different texts (e.g. magazines, more straightforward novels, history books, biographies, travelogues, guides, lyrics, poems), using appropriate reference sources selectively.
> Can read novels that have a strong, narrative plot and that are written in straightforward, unelaborated language, provided that he/she can take his/her time and use a dictionary (Council of Europe: 2018).

As the analysed items date back to the years 2015-2017 and the new *CEFR* with a stronger focus on literary texts was only published in 2018, the texts in the *Finnish Matriculation Examination* do not focus on literary texts but on a wide range of other fields that B2 language users might have to deal with.

As the name '*descriptors'* for each level of the *scales* suggests, the *descriptors* are written down in a descriptive way. They are based on an activity-oriented approach of communicate language competence (Kecker 2011: 74-75). For each of the 7 language levels (incl. Pre-A1), there are *descriptors* within a *scale* such as "reading as a leisure activity" (Council of Europe 2018: 65) in order to distinguish the different

language levels from one another. For example, the *descriptors* for B1 are easier and refer to more familiar fields of the examinees than the *descriptors* at the level B2. The progression among the *scales* is influenced primarily by language activities rather than knowledge of the target languages (Kecker 2011: 74-75).

An aspect of criticism that has been raised with regard to the *CEFR* is the rather frequent usage of terms in the *descriptors* such as "simple", "complex" or "familiar", which are not specified (Kecker 2011: 77) and therefore rather vague. Further examples of unclear language in the descriptors are expressions such as "short", "long" (Figueras 2012: 483) or "everyday vocabulary". The *descriptions* with these vague formulations might be challenging to actually apply in practice (Milton n.d.: 211). For example, with regard to the cited *descriptor* concerning *overall reading comprehension above*, it is not entirely clear and open to interpretation what exactly counts as "a broad active reading vocabulary" (Council of Europe 2018: 60). Although the *CEFR* has been criticised for various aspects, it has become the standard tool for testing in Europe (Kecker 2011: 81) and it has been translated in more than 40 languages (Figueras 2012: 477).

While the lower English *Finnish Matriculation Exam* targets the lower and upper B1 level, the higher one aims at the lower B2 level. The latter one has been chosen for this study because English is often tested at a B2 level at school. For instance, in Austria almost all of the English candidates have English as their L2 and take the A-levels targeting at the level B2. The *overall reading comprehension descriptor* according to the *CEFR* at level B2 is defined as:

> Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms (Council of Europe 2018: 60).

The items of the higher *Finnish Matriculation Exam* are based around the lower B2 level. However, there are also items below and above this level in this exam, ranging from B1.2. to C1.

The reading subcategory of the *CEFR* which seems most appropriate for the selected items in this study is "reading for information and argument" (Council of Europe 2018: 63). This involves detailed as well as careful reading. As for many of the *descriptors*, also for the B2 *scale* for *reading for information and argument* there are sometimes upper and lower level descriptors for the individual levels. This is also the case in the scale *reading for orientation and argument*. The *descriptors* of the lower

B2 level are cited below as these represent the general target level of the higher examination from which the items of this study were taken.

> Can understand articles and reports concerned with contemporary problems in which the writers adopt particular stances or viewpoints.
> Can recognise when a text provides factual information and when it seeks to convince readers of something.
> Can recognise different structures in discursive text: contrasting arguments, problem-solution presentation and cause-effect relationships.
> (Council of Europe 2018: 63)

There is also a *scale* for strategies which should help the receptive processes when trying to understand a text. This *scale* is the same for listening and reading and is called *identifying cues and inferring (spoken & written)* (Council of Europe 2018: 67). This includes the competence to deduce meaning from the linguistic context and the co-text. Furthermore, examinees are expected to use a variety of strategies such as the position of the text, (sub)titles, numbers and proper nouns, prefixes and suffixes and logical and temporal connectors for understanding a text passage (Council of Europe 2018: 67).

2.3. Test items

A number of factors can be considered in order to make test items as good as possible. Green (2017: 101-103), for instance, has presented guidelines for developing listening tasks, which might as well be true for reading tasks:

- If it is possible, the formulation used in the items should be easier than the one in the text. Alderson also suggests that the language used in the item should be simple and easier than the text passage targeted. Otherwise, it cannot be said whether an examinee had difficulties understanding the text or the item (2000: S. 86). This is why in this study it is analysed whether complex item formulations correlate with the facility value of items.
- MISD and SIID questions should not be included in the same task as this makes tasks cognitively more demanding.
- When the aim is to test the comprehension of a text, in the item the same linguistic structures as in the text passage should be avoided. If the same word was repeated in the item, recognition rather than comprehension would be tested. However, if test developers can only come up with a more difficult synonym than the tested word in an exam, then the same word could be used in the item.

- Unless this is the expressed purpose of the test designers, it should be ensured that non-linguistic knowledge, such as maths, is not tested.
- Both the stem of a multiple-choice item as well as the options should not include negative formulations as these are cognitively more challenging to be decoded (Green 2017: 101-103).

2.3.1. Facility value

2.3.1.1. Description of the *facility value*

Facility values are traditionally calculated for each one of the test items as they give information with regard to the difficulty of the items (Alderson, Clapham & Wall 1995: 80-81) for a particular target group (McNamara 2000: 134). Item facility is an important element of classical item analysis to see how suitable a certain item is for an exam. The facility value shows how many per cent of examinees have answered an item correctly. It is a widely used measurement for analysing the difficulty of an item (McNamara 2000: 60).

This figure is calculated by taking the number of right answers divided by the entire number of responses (Green 2013: 26). It is important to keep in mind that unanswered items are usually considered as false (Brown & Hudson 2002: 114-115). For instance, if 60 out of 100 examinees answer an item correctly, the facility value is 60 per cent. At the same time, this means that 40 per cent of the examinees ticked the wrong answer or gave no answer. When there is a high facility value, the item can be considered to be easy for the test population (Green 2013: 26). Although Green (2017: 39) focuses on listening items in her book, her suggestions can also be taken into consideration for reading items. Referring to listening items, Green states that only the comprehension of the sound file should be tested while the items should only be used to determine the level of an examinee. However, when testing, both the sound file as well as the item together measure the level of an examinee. Consequently, item facility is a combination of the text as well as the item.

It is, however, highly unlikely that all of the items in a task are at a certain level. When having 8 items and the targeted level is B2, there is a high chance that there is at least one item which is either C1 or B1. Such items can, for example, be identified by using statistics as well as expert panels, who judge the items of a task.

Only certain results concerning the facility value of an item are considered suitable for testing the target level. After the calculations, the index may range

between 0 and 1.00 for each of the items. By moving the decimal point two numbers to the right, the facility value can be seen as the percentage of examinees answering an item correctly. For example, a facility value of 0.25 would indicate that 25% of the examinees managed to answer an item correctly. It is recommendable to list the facility value of all examinees for all items in a matrix. The facility value enables test developers to compare the difficulty of items and to see which item is the most or least difficult one (Brown & Hudson 2002: 114-115). The ideal item facility would be 0.5 in proficiency tests (McNamara 2000: 61). Khalifa and Weir (2009: 144-147) also argue that a facility value of 0.5 should be aimed at in order to get a large number of different scores. Items as close as possible to a facility value of 50% give most information concerning the individual competence of the examinees. If 50 per cent of the examinees get an item right, the facility value is often given as .5 (Alderson, Clapham & Wall 1995: 80-81). As items are often more or less difficult, an item facility between 0.33 to 0.67 is usually accepted (Khalifa & Weir 2009: 144-147). At the beginning of a test, there might be a few easier items so that the examinees might lose their anxiety and nervousness. Rather hard items could be put at the end of a test in order to get clearer scores to distinguish between the able and less able examinees (McNamara 2000: 61). Items that have a facility value of 0.75 are considered to be too easy, whereas items below 0.2 are too difficult. Items which are too easy offer little information regarding the different levels of ability of the examinees (Khalifa & Weir 2009: 144-146). Test designers have to decide whether these items will be removed from a test. Very difficult as well as very easy items are not suitable for proficiency tests as they only provide little information concerning the actual competence of the examinees (Alderson, Clapham & Wall 1995: 80-81). Many test developers use a facility value between 30 to 70 per cent for their first decision to which extent an item works for the test population. Facility values between 20 and 80 per cent also offer useful information for test developers developing proficiency tests, provided that further statistical information provides good results for an item. *P*-values above 80 and below 20 per cent can be considered to be rather inappropriate for the target test population as these items are either too easy or too difficult (Green 2013: 26). According to Brown and Abeywickrama (2010: 71) there is no absolute facility value that has to be met in order to decide whether an item should be excluded from a test, changed or dropped. However, it is suggested that appropriate items usually have a facility index between .15 and .85. An easier item could be included as a warm-up at

the beginning of an exam. Test developers have to take the facility value of the items into consideration when putting a test together.

If the aim is to show different results regarding the performance of examinees in a proficiency test, the items should neither be too challenging nor too easy. If the items are too difficult, all examinees might get them wrong and thus there cannot be drawn any conclusions concerning the differences in competence among the examinees. Conversely, if an item is too easy, everyone might get it right, and again the better examinees cannot be differentiated from the less able ones (McNamara 2000: 61).

There are various aspects which have an effect on the difficulty of an item. For example, when asking for implicit information in a text, this is usually more challenging than asking for explicit information. Furthermore, questions which do not refer to different parts of a text, but only to one passage, tend to be easier. Questions for which background knowledge is needed also tend to be more difficult. Regarding multiple-choice items, the ones containing rather implausible options will be easier to solve for examinees (Alderson 2000: 113-114). An item becomes more difficult if lexical items from the text appear in the distractors (Buck 2011: 153). This might throw examinees off the scent, because if a content word from the text appears in a distractor, examinees might assume that this is the correct answer as this word also appears in the text passage. It can be argued that such distractors are far away from natural reading, which multiple-choice is anyway, as choosing the right option out of several alternatives does not represent reading processes outside of testing situations. Additional factors making a task more demanding are the language used, the topic, the task type as well as the background knowledge of the readership (Alderson 2000: 39).

2.3.1.2. Freedle & Kostin's study of comprehension difficulty

The focus of this study is to predict the "comprehension difficulty" (Freedle & Kostin 1993: 1) of TOEFL reading items. The findings include variables, which are relevant for the formulation of the item itself, variables concerning the text and variables that are relevant for both, the text as well as the item (Freedle & Kostin 1993: 25-27). Due to a certain similarity of this study with the present research project, it seems interesting to compare the results of the two studies with regard to the significant item variables.

Due to the logo of *TOEFL* on the front page and Freedle & Kostin's publication with *Educational Testing Service* it needs to be questioned, however, whether this is an independent publication or one supporting the *TOEFL* examination, and in this particular case the quality of its reading items, which is being put in a very positive light based on their findings.

The two researchers found in their study that variables regarding the item play only a very little role, while many more text and text-item variables correlate significantly with regard to comprehension difficulty. Only a few of the item variables correlate with the comprehension difficulty of an item (Freedle & Kostin 1993: 36). As most of the significant variables of the study referred to text variables and text-item variables and as there were only three significant item-variables, it was argued that the text actually has to be read in order to answer multiple-choice items. At the same time, this means that the items can, usually, not be guessed and that multiple-choice items are, therefore, a valid measurement for testing reading according to Freedle & Kostin (1993: 1). Moreover, the identified significant variables influence weak examinees more than strong ones. Furthermore, it needs to be taken into consideration that often it is due to several variables that items become harder (Freedle & Kostin 1993: 24-26).

The data consisted of two sets of items. First of all, there were 213 nested items, which means that there was at least more than one item within a text passage. Secondly, there were 98 non-nested items with only just one item within a text passage (Freedle & Kostin 1993: 10).

With regard to the variables only referring to the item, Freedle & Kostin (1993: 36) found that three variables correlated positively with comprehension difficulty of items. Their first hypothesis concerning their research project could be confirmed:

> We expect the following variables to influence reading item difficulty significantly as determined within a multiple-choice testing format:
> Negations: the greater the number of negations, the more difficult the comprehension. (Freedle & Kostin 1991: 6)

Their research showed significant correlations for the following three item variables: negatives in the correct as well as incorrect options and the number of words in the false options. Green (2017: 101-103) also points out that negations should be avoided in item formulations as they are cognitively more challenging. The correlations concerning the significant item variables and the comprehension difficulty of items are weak. For the negatives in correct options it was .13 ($p < 0.05$, 1-tailed) in the nested sample and .16 in the non-nested sample, with no p-value given. Regarding negatives

in incorrect options, the nested items had a correlation of .11 ($p < 0.05$, 1-tailed) and the non-nested ones had a similar figure of .12, with, again, no p-value given. Both of the correlation coefficients are positive, which shows a positive relationship between the negations in the options and the difficulty of the reading comprehension of the items. This means the more negations there are in the correct as well as incorrect options, the higher the comprehension difficulty. Concerning words in incorrect options the correlation was .14 ($p < 0.05$, 2-tailed) in the nested sample and .23 ($p < 0.05$, 2-tailed) for the non-nested sample (Freedle & Kostin 1993: 36). This variable refers to the sum of all words in the incorrect options of the multiple choice items (Freedle & Kostin 1993: 11). Again, there is a weak positive correlation. This means that the more words there are in the distractors, the higher is the comprehension difficulty of the item. The variable of negations was not included in the present study of the items of the *Finnish Matriculation Examination* as hardly any of the chosen items include a negation. However, the sum of the words in incorrect options was correlated with the facility values in order to compare them to the findings of Freedle & Kostin's study.

Many more variables were found by Freedle & Kostin (1993: 36) to correlate positively and negatively with comprehension difficulty with regard to text-item overlap variables and text variables. Concerning the item-text overlap variables, the following variables correlate positively with comprehension difficulty for solving an item:

- the length of the text passage which has to be read to answer an item,
- a larger number of words which have to be read before the relevant information in the text,
- if a main idea information is located rather in the middle of a text.

In contrast, there is a negative correlation between the variables and the comprehension difficulty when words from the key text sentence or lexically related words also appear in the correct multiple-choice options. Green (2017: 101-103) describes this as *recognition*. This is the only variable with a medium significant negative correlation, while all of the other text-item variables have a weak negative correlation. This suggests that the variable concerning *recognition* is particularly significant with regard to the difficulty comprehension for answering an item correctly.

Purely text-based variables correlating with comprehension difficulty of an item were also found in Freedle & Kostin's study (1993: 36). For example, the

following features had a positive correlation with comprehension difficulty: vocabulary, the topic social science, text passages dealing with problems/solutions, the length of the first sentence of a paragraph and references across text clauses. All of these text-variables analysed in their study had a weak positive significant correlation. Concerning negative correlations with comprehension difficulty, the following three features were significant: concreteness, topics dealing with humanities and with descriptions/lists. All of these three variables have weak significant negative correlations.

2.3.2. Multiple-choice items

Concerning the test method, the concept of response format is important. An examinee will be asked to answer items by using a certain test method (McNamara 2000: 16). The response format of this analysis are multiple-choice items, which is a fixed response format as the answers are already given, in contrast to constructed response formats used in speaking or reading assessment (McNamara 2000: 136). If there are more than two options in an item, this can be referred to as a multiple-choice item (Glaboniat 1998: 96). Multiple-choice questions are also referred to as selective response since test-takers have to choose from several given responses and they usually do not have to produce any language. Multiple-choice items are a good way of testing receptive skills such as reading according to Brown & Hudson (2002: 68-69). However, although it is easy to develop a multiple-choice item, it is quite a challenge to create good ones (Brown & Hudson 2002: 68-71). In fact, the development of multiple-choice items is a highly professional skill, which requires a great amount of time in order to be done well. Moreover, it is particularly essential for multiple-choice items to be pretested as it can be hard to predict the examinees' results. After pretesting, it can be assumed that a large number of items need to be either changed or dropped (Buck 2011: 142-146).

Multiple-choice items can be divided into different parts. They have a stem, which acts as a stimulus, and there are usually three to five options, also called alternatives. Usually only one of these is the key, while the other options function as distractors (Brown & Abeywickrama 2010: 68). However, it is possible that more than one of the options is correct or none of them in a certain item (Glaboniat 1998: 96-101). Furthermore, it can be a challenge that only one of the given alternatives is the correct one (Alderson, Clapham, Wall: 1995: 47-48).

In this regard, Brown and Hudson (2002: 61) describe three rules they sum up as linguistic confounding. Firstly, the formulation of items should reflect the examinees' language proficiency of the target language. If examinees have difficulties understanding the stem, the results will be ambiguous as testers do not know whether the answers reflect the skill being tested or an inability to understand the language of the item. Both the stem as well as the options should be worded as directly and simply as possible. Also, redundancy in the options should be avoided, in order to keep them short. For example, when all of the options start with the same relative pronoun, this could be moved up to the stem and thereby the item can become shorter (Brown & Abeywickrama 2010: 69). Secondly, negative and double negative statements should be avoided (e.g. Why did the student not deny lack of punctuality as an inappropriate basis for grading?) as they may confuse examinees and are challenging to process. Thirdly, test designers should avoid ambiguous formulations so that examinees know exactly what is being asked of them (Brown & Hudson 2002: 61).

In addition, the alternatives given should be rather similar with regard to length and style because if one of the possible answers stands out, learners might be led to believe that this answer is correct or wrong (Alderson & Clapham & Wall: 1995: 49). It is a serious problem if one option is clearly shorter or longer than the others. Experienced examinees know that they have a higher chance of getting the right answer when choosing particularly short or long options (Brown & Hudson 2002: 69). Green (2017: 106) also points out that the options should be approximately of equal length as examinees tend to dismiss options seeming different than other ones. With regard to item length, different examinees might have different assumptions based on their experience. Some examinees might think that the longest option is always the correct one or that it is impossible that the same option (e.g. C) is the correct one three times in a row and that the correct options (e.g. A, B, C or D) must be distributed evenly among these four letters (Glaboniat 1998: 103). There are *high-stakes examinations* that order the options in multiple-choice items alphabetically. Hence, it happens that a certain letter (A, B or C) is frequently the answer and that one of these letters never is the answer within a task. Some examinees might be confused when, for example, C is often the answer and B is never the answer within a certain task. These are strategies that are construct-irrelevant and divert from testing the actual reading competence of examinees. In fact, the answering of multiple-choice items can be enhanced by acquiring certain techniques, which reduces the testing of the actual

reading competence (Glaboniat 1998: 103). There exist also books with tips for tackling multiple-choice items successfully.

Two basic ways for developing multiple-choice items exist: The stem either consists of a question and the options of (short) sentences, or the sentence of the stem is completed by the options (Alderson, Clapham & Wall 1995: 47). In the case of the present study all of the stems are formulated as questions.

In addition, if each item is given a separate mark, then each item should be independent from the others. So if one item is answered correctly, this should not have a negative effect on answering other items (Alderson, Clapham & Wall 1995: 47).

Furthermore, words appearing in the stem should be avoided in the distractors as this might help to lead examinees to the correct answer. In addition, if the correct answer refers to a main idea, then all of the distractors should refer to main ideas as well (Green. 2017: 106).

The advantages of multiple-choice items are their reliability. They can be easily scored because the correct answers are predetermined (Brown & Abeywickrama: 2010: 67). Khalifa & Weir (1009: 83) also stress that multiple-choice items are often chosen due to their reliability. In addition, such items tend to be good discriminators between strong and weak examinees. The difficulty of the task can be easily altered according to the level by changing the distracte3rs or the selection of the text. In addition, they are an accepted measurement for testing whether examinees have understood a certain text in detail. They also make it possible to test more sophisticated aspects of a text, such as argument, inference or opinion, in a way which is better controlled than with open ended formats.

An issue is, however, the degree of validity of multiple-choice items, especially when it comes to testing either reading or listening skills in communicative performance tests. This is because such multiple-choice items do not represent natural reactions to written or spoken language. Consequently, with regard to validity it is not clear to which extent multiple-choice items actually test what they are supposed to test. Furthermore, test developers usually want examinees to understand the key in the text and then look for the correct answer in a multiple-choice item. However, often this is not the way examinees go about when answering multiple-choice items. In an exam situation a wide range of strategies are applied by examinees. Most importantly, examinees evaluate the options according to their appropriateness for answering a multiple-choice item, until the most likely option is chosen. Those answers, which are

considered to be wrong, are no longer taken into consideration. Options are falsified until the correct answer is found. Consequently, an examinee might get the correct answer by eliminating the wrong ones. The use of these strategies can be proven by making out of each multiple-choice option a true or false item. Then the results of the true and false items are compared to the actual multiple-choice item. Such an experiment was, for example, done with a listening exercise from *Zertifikat Deutsch als Fremdsprache*, Modellsatz 05 (Teil B: Hörtext über den Spreewald)[3]. All of the multiple-choice items were also turned into true and false exercises for that research project. Two test sheets were developed. The first one consisted in the first part of the original multiple-choice items and in the second part of the true and false questions. In the second test sheet it was the other way around with the second part consisting of the multiple-choice and true and false items in the first section. Interestingly, the results showed that the examinees of both groups working with the two different sheets performed worse with regard to the true and false items. For example, there were options which were not considered to be attractive in the multiple-choice items, because the other options seemed more likely. By contrast, in the true and false questions these unattractive options were chosen more often as correct answers, although they were wrong (Glaboniat 1998: 102-103).

The criterion of practicability is given to a certain extent with regard to this test method. Multiple-choice items are easy to score for testers. However, they are challenging to develop for test developers. The production of multiple-choice tasks can be quite time-consuming. However, if used for repeated administrations in a standardized test, multiple-choice items might be very useful (Brown & Abeywickrama 2010: 67). From the examinees perspective, practicability is given, when the items are clear as well as unambiguous (Krause & Sändig: 2002: 84-85). With regard to practicability this shows that the item formulations are an important point for test developers as well as for the examinees.

Another disadvantage of multiple-choice item is that they require quite a large amount of text compared to other test methods due to the three to four options. This can be justified to a certain extent when testing reading. But when testing other skills such as listening, reading is actually not part of the construct (Glaboniat 1998: 103-104) and reading an a stem with four options takes some time while listening.

[3] Certificate German as a foreign language. Model exam 05 (part B: listening text about the Spree Forest).

There is also quite a high chance of guessing the correct answer (Galboniat 1998: 104-105). This is when examinees use their world knowledge and test taking strategies and choose the most likely option. However, these strategies are not part of the construct. Due to the high risk of guessing, researchers and test developers try to enhance the quality of multiple-choice items and to avoid the risk of guessing. This risk can be reduced when not just one answer can be correct but more than one, all of the options or none of them (Glaboniat 1998: 104-105). In the multiple-choice items of the present study there is only one correct answer for each of the items.

Although multiple-choice items have their downsides, they are often found in tests. This is also because this test method seems to have a very high face validity. However, multiple-choice items are often rejected with regard to teaching (Glaboniat 1998: 104). Teaching and skill development are better trained with other methods such as the "Fremdsprachenwachstum [foreign language growth]"[4] by Buttaroni and Knapp (1988). Test methods such as multiple-choice are more suitable for testing purposes. *High-stakes tests* such as the *Finnish Matriculation Examination* have a big washback effect on teaching. This is why past papers and similar test exercises are often used when preparing examinees for a test. However, teaching to the test should not take too much room and the focus should still be on skill development.

When test developers include multiple-choice items a few aspects should be taken into consideration. Good multiple-choice questions should represent questions a reader might have concerning a certain text. Furthermore, the questions should follow the order of the natural reading process of the reader. Moreover, the distractors should represent natural, expectations, associations and assumptions, which reflect a real situation. Negations should be avoided and if they are used, they should be underlined or highlighted somehow for the examinees. Neither the position nor the length of the

[4] In the teaching approach of the "Fremdsprachenwachstum", each individual step is clearly outlined by the researchers. Firstly, this method involves "authentisches Lesen [authentic reading]", which asks learners to read the text several times. After the first reading the learners discuss with a partner what they have understood and then after the second reading they talk about what they could take away from the text with another learner. Then a few unknown words are underlined while rereading the text. This should include lexical items that are deemed to be essential for understanding the text. The meaning of these unknown words is then also talked about with other learners, which trains the skills of *inferencing*. It is argued that the reading competences are trained by reading authentic texts repeatedly and guessing the meaning of unknown words. Secondly, during "analytischem Lesen [analytic reading]" the learners can then focus, for example on grammatical structures of a text, while reading it again several times and discussing their findings with different other learners after each time the text was read (Buttaroni & Knapp 40-43). It needs to be taken into consideration, however, that this method might be unusual for learners who are not used to working in depth with the same text and to read it several times.

options should give a hint to the correct answer. In addition, the distractors should not just be fillers but plausible alternatives to the correct answer (Glaboniat 1998: 106).

3. Context of the study

3.1. The *Finnish Matriculation Examination*

The *Finnish Matriculation Examination* is a nationwide exam, taken when finishing upper-secondary school. After passing the *Finnish Matriculation Examination*, the examinees are allowed to attend higher education.

The *Finnish Matriculation Examination* is carried out by the *Matriculation Examination Board* in all upper-secondary schools of Finland twice a year at the same time in spring and in autumn. Every year, about 35.000 examinees succeed in the exam, the majority of whom take it in spring.

The *Matriculation Examination Board* publishes guidelines, for example, regarding the contents as well as the scoring of the exam. Institutions of higher education and the *Finnish National Agency for Education* make suggestions for the around 40 members as well as the chairperson of the *Matriculation Examination Board*, who are then nominated by the *Finnish Ministry of Education and Culture*. The members of the *Finnish Matriculation Board* represent the different subjects for which exams are being produced. There are 330 associate members, with the help of which the Board prepares as well as assesses the various tests (Ylioppilastutkintolautakunta – studentexamensnamnden a: n.d.).

The examinees have to choose a minimum of four exams. The languages of the examination are Finnish or Swedish, apart from the exams for the foreign languages. Apart from the obligatory test in their first language, the examinees have to take the test in at least one further language: a foreign language or the second national language, which are both also offered at an advanced level at which the examinees have to choose one exam.

The exam must be finished within three consecutive periods of examination. When a compulsory test is failed, examinees are allowed to retake it twice and to change from the advanced to an easier syllabus level as long as the examinee takes one compulsory test at an advanced level. If a test is not passed within three examination periods, the entire exam has to be retaken. After passing the exam, examinees receive the *Matriculation Examination Certificate.*

After completing the *Finnish Matriculation Examination*, a test can be retaken once or other tests of the exam can be taken, for which they receive an additional certificate. Further tests can be taken twice without any time constraints (Ylioppilastutkintolautakunta – studentexamensnamnden b: Structure of the examination: n.d.).

Concerning the foreign languages, exams are offered in English, German, Russian, Spanish and French either at an advanced syllabus or a basic syllabus level. The *Matriculation Board of Finland* does not have any of its exams online. However, the *Finnish Broadcasting Company YLE* publishes the exams with the keys on its website, where the items and texts were also taken from for the present study. Only for the spring date of 2016 the reading key is not available, because the listening key was published twice on their website, instead. Furthermore, tests at the basic syllabus level are offered in Portuguese, Italian, Skolt Sami, Inari Sami, North Sami and Latin. The "higher level" of the foreign languages exams is referred to as "pitkä oppimäärä" in Finnish in contrast to the "lower level", "lyhyt oppimäärä". The higher exam is for those having studied English for 8-10 years at school. The lower level is aimed at examinees having studied English for 5 to 6 years. The examinees can choose either level regardless of how long they have been studying the target language. The examinations take place twice a year, in spring ("kevät") and in autumn ("sypsky"). The target level for the higher exam is B2.1. and for the lower one B1.1.-B1.2. 50% of the items are aimed at the targeted level, 25% are below it and the other 25% are above it. Consequently, there are also items at C-level in the "pitkä oppimäärä". This makes it possible to also get clearer results with regard to examinees who are above or below the target level.

For most of the languages, the test is made up of two parts: The listening exam and the written one, which take place on two separate days. Both parts need to be completed in order to pass the exam.

The written part of the test consists of reading as well as text production and is subdivided into three parts. The test methods can include, for example, multiple-choice questions, open questions, cloze tests and different forms of text production. The written tests analysed consisted all of 25 multiple-choice reading items in their first part. The reading texts consist of several short texts for each of which there are about three multiple-choice questions. This is followed by a language in use part, which is called "Grammar and Vocabulary". This consists of 25 further items, where examinees

have to fill in gaps. For each of the gaps, four options are given and the correct word or expression has to be chosen. This section is then followed by further 15 items where gaps also need to be filled in. Tips are given for each item to find the correct word. The tips include Finnish words, verbs, which have to change their form, or just the type of word is indicated, such as "preposition" and then the examinees have to find the right preposition that fits into the gap. This is followed by a section concerning text production. The listening part of the exam is in a separate file. However, for the present study only the reading items were taken into consideration.

Since 2016, digital tests are gradually being introduced, in which examinees sit the listening and written exam on the same day (Ylioppilastutkintolautakunta - studentexamensnamnden c: n.d.).

Although the skills are tested isolated from each other in the *Finnish Matriculation Examination*, Kecker et al. (2019) found in a study dealing with the university admission examination *TestDaF*[5] that the skills are usually used in combination with each other. *TestDaF* is concerned with testing German as a foreign language for examinees wanting to study in Germany. This study was conducted in order to find out in what way examinees have to use German in university contexts. The results suggest that the separation of skills in their tests did not match real-life language usage of the examinees in university settings. For example, students would have to read texts, while writing a paper or students would have to talk about what they have read in a university course. In order to prepare students better for their real-life situations, integrated exercises were included in the exams of *TestDaF* (Kecker et al. 2019: 400). The study by *TestDaF* suggests that the completely isolated test of reading in the *Finnish Matriculation Examination* does not seem to prepare examinees for all future situations, for instance, in university settings, although examinees passing the *Finnish Matriculation Examination* can then go on to university. Based on the study by *TestDaF*, it could be taken into consideration to also include exam questions which link the different skills. In fact, it is, for example, highly unlikely that examinees will have to answer multiple-choice items about a reading text outside of the educational context.

Concerning the number of examinees during each of the exam dates, there were almost 20.000 examinees in the spring dates and less examinees in the autumn dates with regard to the five exam dates, which could be analyzed. The presented figures in

[5] The exam title means „testing German as a foreign language".

this section only concern students having had Finnish as the language of instruction. The data of the students having Swedish as their language of instruction is in a separate table, which was not taken into consideration for the present study. Regions with Swedish L1 speakers are found especially on the Åland islands and other coastal areas in the south and west of the country. The numbers of examinees are based on the total number of examinees ticking the options in the multiple-choice items between spring 2017 and spring 2015. This means that for each of the five exam dates, the examinees choosing the options A, B or C could be added up in order to find out about the total number of examinees. During these exam dates, the number of examinees sitting the higher English examination are quite stable for both the spring and autumn dates. In the spring date of 2017 there were almost 19.300 examinees taking the higher English examination (YLE 2017a). In spring 2016 there were almost 19 450 examinees (YLE 2016a). Similarly, in spring 2015, there were a bit over 19 600 examinees sitting the higher English examination (YLE 2015c). With regard to the autumn dates, there were almost 14 300 examinees in 2016 (YLE 2016b). Similarly, during the autumn date of 2015 there were about 14 200 examinees taken the higher English *Finnish Matriculation Examination* (YLE 2015a).

3.2. Research question

When reading difficult formulations in items, one might wonder whether this can cause additional difficulties for answering an item, which are not part of the construct. In contrast to listening, in testing reading it could be argued that it also tests reading when an examinee has to understand the question. However, as pointed out earlier, such multiple-choice questions with three options usually do not occur outside of testing situations and therefore they do not present authentic reading texts. In fact, the focus should be on testing the understanding of a wide range of authentic texts, instead. This is why it seemed interesting to analyse the formulation of multiple choice reading items. Hence, the research question is to what extent formulations of multiple-choice items relate to the facility value of items measuring reading competence.

The Finnish school system has a very good reputation due to outstanding results in the PISA tests. Reading items from the *Finnish Matriculation Examination* could be used for this analysis thanks to the collaboration of a colleague from the *Matriculation Examination Board* who sent the facility values of five exam dates on request. It might

be interesting to see if this study leads to the same results as the study by Freedle & Kostin (1993: 36), which was presented in chapter 2.4.2.2.

3.3. Hypothesis concerning the outcomes of the research question

Apart from detailed research questions, clear hypothesis regarding possible findings are important when conducting quantitative research (Daase & Hinrichs & Settinieri 2014: 104, 109). Based on the previously presented information concerning testing, hypothesises should be formed with regard to the expected outcomes of the present study (Cohen, Manion, Morrison 2007: 79-82; Riemer 2014: 15-21). As pointed out earlier, Freedle & Kostin (1993: 36) found the following item-variables to cause difficulty: negations as well as the number of words in the distractors. It is widely accepted in the research field of testing that negations should be avoided in items, because they require more complex cognitive processes when decoding the meaning of such questions. However, negations might be used if an item becomes even more complicated otherwise. Probably due to the fact that it is suggested to avoid negations in items, there are hardly any negations in the Finnish multiple-choice items and, if they do occur, they are kept very simple. Interestingly, concerning the chosen items from the five exam dates, only in the questions from spring 2016 negations are found (YLE 2016b). In fact, with regard to the chosen MISD items, there is only a single negation in a verb construction (item 22 in the appendix A: “It did not occur.”). Apart from this, simple negations are found only in a few cases before nouns in four options (item 19 in the appendix A: “He left no will on paper. He signed no papers. He left no paper trail”; item 22: “It attracted no funding” (YLE 2016b). A single more negation occurs in an item with the negative form of the adjective “violent”, used in distractor of item 26 in the appendix A: “To parallel a non-violent revolution and the eventual outcome of the debate” (YLE 2016b). As there are so few negations and as it seems to make sense to avoid them, this was not analysed. The other significant item-variable that Freedle & Kostin (1993: 36) found in their study, the length of the distractors, was included in the present analyses, because it came out as significant in their study.

The analysis of the present study will show whether this research project comes to the same findings as Freedle & Kostin’s research project. One might question why the length is only relevant concerning the distractors in Freedle & Kostin’s study, but not regarding the stem, the key or the whole item.

In fact, the hypothesis of this research project is that length alone might not be so significant when the language used is simple. Instead, the hypothesis is that it is important to use easy lexical formulations in the items, although this aspect did not appear as significant in Freedle & Kostin's study concerning the item-variables. The main hypothesis of this study is, therefore, that there is a relationship between difficult formulations in items, especially concerning vocabulary, and the facility value of the items. In fact, it is assumed that more complex and difficult item formulations might correlate negatively with the facility value. This means that when there is a more complex formulation of the item, less examinees might get the correct answer. After the analysis, it will also be discussed to which extent this hypothesis turned out to be true.

4. Methodology

4.1. Indicators of quality of quantitative research

According to Schmelter (2014: 38-40) there can be distinguished three indicators of quality for quantitative research projects: validity, objectivity and reliability. These concepts have also been addressed from a testing perspective in the theoretical chapter, whereas now they are discussed with regard to quantitative research.

Validity asks the question whether a different measuring instrument would lead to the same result (Schmelter 2014: 40). As the results were calculated automatically with computer programmes it can be assumed that they are valid when the same definition of the concepts is used, for example, that lexical density refers to content words (*Analze My Wrting*: n.d.).

Objectivity refers to the accountability of a certain method. Independent researchers should come to the same results when calculating the data in the same way (Schmelter 2014: 39). Again, as the results were calculated automatically, other scientists should come to the same results when they work with the same definition of the concepts.

Reliability refers to the degree of precision with which a method measures a certain feature. When working with two comparable groups of subjects or sets of data in this case, the same results should be reached due to a lack of measuring mistakes. This is, for instance, possible when retests are being done (Schmelter 2014: 39). It would make sense to look at the results of five further test dates of the *Finnish*

Matriculation Examination and to compare them with the findings of this study. However, in this case the *Finnish Matriculation Examination Board* would have to be asked to provide the data of five additional exam dates, which would go beyond the limits of the present research project. Also, papers from a different time span than the ones used for this analysis might differ a bit due to further development of the exam, which would also have to be checked with the *Finnish Examination Board*.

A further quality aspect concerns the reflection of ethical consideration concerning the research project. The data was emailed directly by a member of the *Matriculation Examination Board*. It was communicated by the *Finnish Matriculation Board* that the data is not published, but passed on on request, for instance, for the present analysis. At first, the results of two exam dates were sent and when asking for further results, the ones of three additional exam dates were emailed. The results are completely anonymous. Because of this, no need was seen to ask the test population for their permission to use the results for this research project. Moreover, it seems that there do not have to be any worries regarding data security when discussing the facility value of the items as the results are anonymous.

In order to make the study and its calculations as transparent as possible all the data is included in the appendix so that the reader can understand better how the results of this study were calculated. Appendix A includes the chosen 76 items from the five exam dates. Appendix B provides different kinds of information regarding the items such as the exam dates, the calculated facility values as well as the readability index for each item. Appendix C and D present the calculations of the 22 features concerning the lexis, item length and the number of phrases in the items. These features were then correlated with the facility values. The results are listed in chapter 5 in figure 2.

4.2. The data

In total, the results of five past papers were sent: from the spring date of 2017 and from the years 2015 and 2016 both the spring as well as the autumn dates were obtained. All the tables with the facility values of the different exam dates are in Finnish, but the important terms were all translated into English by the contact person of the exam board. The keys of older past papers contain less information than the more recent ones. The results of examinees having Finnish and Swedish as an L1 are separated in

different tables. It could be assumed that English is easier for Swedish learners as English and Swedish are both Germanic languages, while Finnish is a Finno-Ugric language. The tables concerning the Finnish examinees were used for the analysis as this population is much larger than Swedish-speaking group. The testing exercises for the different skills and the keys are published on a website called *YLE*, from which the reading items and the short texts from the five analyzed exam dates were taken.

4.3. The short reading texts of the exam

The rather short reading texts of the *Finnish Matriculation Examination* usually consist of a few paragraphs of texts of a wide range of different topics and sources. For instance, the spring examination of 2017 consists of nine short texts and includes topics such as business, medicine, history, tourism, the media etc. Above each of the texts there is a title written in bold letters. At the end of each text, the source and the date of publication are given, which illustrates the authenticity of the texts. This includes, for example, newspapers such as the *International New York Time* or *The Guardian* and magazines such as *The Oprah Magazine* or *Psychologies Magazines.* Test developers need to take into account that different examinees find different texts appealing, interesting and motivating to read, which the *Finnish Examination Board* seems to consider. As the texts are from different regions of English-speaking countries, the exam can be classified as pluricentric. Empirical data concerning the different standard varieties of German in the standardized Österreichische Sprachdiplom[6] (ÖSD) shows that the ÖSD does not become more difficult for examinees when more than just one standard variety of the target language is taken into consideration in an exam (Glaboniat 2019: 420). Based on this data it can be assumed that the fact that the *Finnish Matriculation Examination* is also a pluricentric exam, does not make it more difficult. For each of the short reading texts in the *Finnish Matriculation Examination* there are about three items. In total, there are 25 multiple-choice items in the reading part.

The choice of interesting texts in a test situation might make the testing procedure more pleasant for the examinees. Due to the fact that languages are often learnt involuntarily, texts need to be chosen carefully and different aspects such as the learners' competence and interest or the requirements of the curriculum need to be taken into consideration (Feld-Knapp 2005: 21-22, 48). Chosen texts might be based

[6] Austrian Language Diploma

on topics that the test population deals with and which are up-to-date for them (Feldknapp & Schoßböck 2010: 115-135). Furthermore, it might be motivating to read texts which involve learners in an emotional way such as literary texts (Peschel 2013: 173), which might support the sensitivity and understanding for others (Bredella 2007: 30-32). Difficult topics, which might involve examinees in negative emotions should be avoided and such texts are not included in the *Finnish Matriculation Examination*. An example of one of the short texts in this examination can be found in subsection 4.4.3.

4.4. The chosen multiple-choice items

4.4.1. Procedure of choosing the items

It should be ensured that the analysed items are as similar as possible with regard to aspects not referring to the item formulation. Only then it can be calculated, to which extent the item formulation is significant. This is why only MISD items from the higher ("pitkä") examination were analysed with a certain readability index concerning the target text passages. Out of the 125 multiple-choice reading items of the five reading papers, 76 similar ones could be chosen for the analysis. In the following subchapters the choice of the items is addressed in more detail.

For the purpose of this study, the facility value of the *Finnish Matriculation Examination* will be correlated with as many different features as possible concerning the item formulations in order to see to which extent the difficulty of the formulation of the items correlates with the facility value. Ideally, the relationship should be as low as possible, as the reading competence of a chosen text for a particular level should be tested. Therefore, the formulation of the multiple-choice questions should not be an obstacle for answering them.

In the following chapters, items of the exam are also discussed for illustration. An *x* is put next to the correct option of the items in the following chapters in order to indicate which of the three options is the correct one of the items.

The numbers of the items refer to the numbers 1 to 76 of the chosen items in the appendix A, where the key of each item is highlighted. In the table of the items in the appendix A, the first column orders the items from 1 to 76 for this analysis. In the second column the exam date (kevät means spring, sypsky means autumn) is given as well as the number the item had in the original exam paper. In the third column, the items are given excluding the letters A, B, C for the three options.

4.4.2. Selection of the 76 MISD items

The *Finnish Matriculation Examination* aims at testing reading in a very broad sense. There are items testing specific information, main ideas, the gist of a text passage, and in some cases examinees also have to infer meaning. The 25 items testing the reading comprehension in each exam date are below all the short reading texts. Out of the three options of the items, only one is correct as the instruction for answering the multiple choice items illustrates:

> *Read texts 1.1a–1.1g and then answer questions 1–25. Choose the best alternative for each item and mark your answers* ***on the optical answer sheet in pencil*** (YLE 2017a).

In order to be able to focus on the difficulty of the formulation of the items, as many other factors as possible which are influencing the facility value had to be ruled out. Firstly, only items from the higher examination were chosen. This should ensure that only more difficult items are included in the analysis.

Secondly, only items focusing on the same reading style, MISD, were included in the analysis in order to focus the analysis on similar items. Moreover, GIST items of the short texts might overlap with MISD items, which is one of the reasons why GIST items were excluded from the analyses. All of the chosen MISD items include a verb. However, the length of the options can differ as the following two examples illustrate. A typical MISD item consists of an element after the verb, such as a noun phrase in options B and C of the following example:

What characterizes these innovators?
A They are conventional
B They make a profit
C They provide a service x (YLE 2016a)

The MISD might also be shorter and only consist of a verb phrase without any further elements:

What is the change described in Mr. Crites's business all about?
A Settling down
B Teaming up x
C Letting go (YLE 2016a)

Although the options tend to be quite short in the *Finnish Matriculation Examination*, the MISD sometimes consists of more elements before and/or after the verb, as the following item shows:

How does research view the art of goal setting?

A It emphasizes the importance of making a plan x
B It highlights the necessity to create multiple plans
C It outlines the need for back-up plans (YLE 2016a)

GIST and SIID items were sorted out. For instance, the following item was not included:

How were optimal egg-substitutes found?

A By surveying exotic species of flora
B Through extensive research
C As a fortunate coincidence x (YLE 2015a)

The item above is one of the rare ones in which not all of the options represent the same reading style such as MISD. In this case, option A "By surveying exotic species of flora." can be classified as a MISD due to the idea expressed with a verb. By contrast, options B ("Through extensive research") and C ("As a fortunate coincidence") are SIIDs, which do not include a verb. As the majority of options are not going for a MISD idea in the text, this item was not included in the analysis.

In the *Finnish Matriculation Examination* there is also quite a large range of different items, which are not MISD and were, therefore, excluded from the analysis. The following item is an example of a SIID item, which does not include a verb and was, therefore, not included into the item corpus of this research project:

How did The Pages Project initially come into being?

A As a team effort
B As a full-time effort
C As a solitary effort (YLE 2017a)

The following item asks for the style of the text, which is why it was excluded from the chosen corpus of items:

Which type of style does this passage represent?

A One that draws on quantifiable statistics
B One that illustrates individual case examples
C One that tries to affect the reader's emotions (YLE 2017a)

The next item focuses on *inferencing*, which is why it was not included in the analysis:

What is meant by "rustout"?

A Underachieving
B Undercoming
C Undergoing (YLE 2017a)

The following question focuses on GIST by asking for the text heading of the text, which is why it was sorted out:

Opt for an alternative heading for this text.

A Averting Change
B Averting Defence
C Averting Attack (YLE 2016a)

4.4.3. Calculating the readability index of the MISD items

In order to ensure that the text passages including the MISD solutions were approximately of equal difficulty, the readability of each relevant paragraph for answering the item was analysed. An often applied formula for calculating the difficulty of a text is the readability index by Flesch, which is still in use nowadays (Alderson 2000: 71-72). According to Flesch (1948: 223-229), in order to determine the readability of a text, the following aspects need to be taken into consideration. Firstly, this concerns the average length of sentences in words. The second aspect refers to the average length of words in syllables by counting the number of syllables within 100 words. Thirdly, the average percentage of "Personal Words" needs to be taken into consideration (Flesch 1948: 223). Such lexical items include all pronouns in the first, second and third person. Only the neuter pronouns, *it, itself, its, they, their, theirs, them* and *themselves* are not part of the "Personal Words" when they are not referring to people but rather to things. "Personal Words" are lexical items having feminine or masculine natural gender such as *Jones, sister, actress or iceman*. Singular as well as plural forms should be counted. Common-gender words such as *employee* are excluded. In addition, the group words *folks* as well as *people* (incl. the plural verb) should be included (Flesch 1948: 229). Fourthly, the average percentage of what is referred to as "Personal Sentences" is considered. This category counts the percentage of sentences including spoken sentences with quotation marks, commands, questions, requests and further sentences which are addressed directly to the readership. The number of sentences whose grammar is incomplete and whose meaning therefore has to be guessed from the context are also included in the calculation of the readability index (Flesch 1948: 223).

The highest level of readability is expressed by the number 100, while the lowest is 0. A score of 100 means that a barely literate person would be able to understand a passage. A result of 0 means that a text passage does neither include "Personal Words" nor "Personal Sentences" (Flesch 1948: 224- 225). A readability index of only 0 would suggest that a text is almost unreadable (Flesch 1948: 229).

For an analysis regarding readability, firstly, each text sample should start with a new paragraph. Secondly, for each paragraph one should count up to 100 words (Flesch 1948: 228-229). The programme *Textinspector*, with which the readability of the texts were calculated in this study, also states that the readability score is only accurate on longer documents consisting of at least 100 words. Lexical elements separated by a blank space count as a word. Thirdly, the number of syllables per 100 words should be counted. Symbols without letters should be counted as they are said or written down. For instance, the dollar sign counts as one word as does the number of a year. In case there is a large proportion of numbers in a text, the result would be more accurate if these figures were excluded from the analysis. Fourthly, the average length of the sentences should be counted in the text. Fifthly, the number of "Personal Words" per 100 words should be counted. Sixthly, the number of "Personal Sentences" in the text or all samples should be taken into consideration. The number of "Personal Sentences" needs to be divided by the number of sentences within a piece of writing (Flesch 1948: 228-229). Finally, the readability of a text can be calculated by using the results of the various steps just described.

Nowadays, computer programmes such as *Textinspector* count the results automatically. This programme was used for the present analysis for calculating the readability index of the text passage of each of the 125 items. This programme enables users to copy texts into a field and by pressing *analyse* the *Flesch Reading Ease* readability score is calculated.

As the texts contained several paragraphs, each of them was analysed individually. For this, the text of each paragraph testing a MISD was copied into the field to analyse the readability of this paragraph. When a shorter text containing less than 100 words was analysed, *Textinspector* informs its users that at least 100 words are needed in order to produce accurate results. If a paragraph contained less words either the last sentence from the previous or the first sentence from the following paragraph was included, depending on which of these sentences produced a closer number to 100 words. In a few cases more than just one sentence from the adjacent paragraphs had to be included. Titles as well as information regarding the source were excluded from the analysis. In the following subchapter, the calculation of the readability index is exemplified with sample items included in the analysis.

Only items referring to paragraphs showing similar results were chosen for this analysis. The readability of all the MISD items of the five reading papers were

calculated. The results are presented in appendix B. Only items referring to text passages with a readability index between 0.35 and 0.65 were included in the analysis. Even more similar readability indexes would have been preferred for the analysis, but then the number of the chosen items would have been smaller. In this way, 76 similar MISD items could be chosen for the analysis of the item formulation. By working with MISD items with comparable readability indexes (among other similarities of the chosen items), it can then be argued that aspects of the item formulations might be significant concerning the different facility values of the chosen items.

The following text and items are a typical example for one of the short reading texts from the more difficult ("pitkä") autumn examination in 2016. The text deals with reaching one's goals. As in the exam, the items come after the text. The underlined passages indicate the answers to the correct options in chronological order of the items. Item 4 was not included in the analysis because it is a GIST item. The items can be found in the appendix A under the numbers 11-14 from the spring examination of 2016. As items 11-13 are MISD items, they were included in the corpus for the analysis. These items are MISD, because from the underlined passages in the text below it can be seen that an idea including a verb needs to be understood. The options do not refer to the entire paragraph, but only to a specific idea, which is also why these items was also classified as MISD.

Briefing: Health Matters

Now, how's this for a terrible irony: the more you want your goal, the less you are likely to plan for it, according to a forthcoming paper in the journal Behavioral Science and Policy. That is because we tend to think good intentions are enough, but an actual plan prevents procrastination, putting things off. Research shows that people with plans tend to stick to their goals way more often than those who wing it.

Yet, backup plans may backfire by zapping your desire to chase your main goal. In a series of new studies, people who were told to think up a Plan B were less likely to attain their main objective. Researchers suspect that having backup goals may make failure feel somehow more acceptable.

And you know how good it feels to tick off an item from your to-do list. Put that to work by hacking a massive goal (reading 24 books a year, say) into parts (two per month). It's more gratifying and attainable than working away at one big goal, says George Wu, professor at the University of Chicago's Booth School of Business.

Finally, think of willpower as your greatest natural resource, but know that it's also a finite one, some experts say. Every time you engage your willpower for one task – saying no to a chocolate bar – you have less energy to resist other temptations. Since willpower is the secret ingredient to meeting your goals, don't waste it.

Time, Dec. 29, 2014 / Jan. 5, 2015 (YLE 2016b)

Item 11

How does research view the art of goal setting?
A It emphasizes the importance of making a plan x
B It highlights the necessity to create multiple plans
C It outlines the need for back-up plans

Item 12

Why does it pay off to cut objectives into pieces?
A They become more updated
B They become more important
C They become more doable x

Item 13

What is said of willpower?
A It can be used quickly
B It must be used wisely x
C It should be used widely

*Item 14

Opt for an alternative heading for this text.
A How to Make Your Decisions Firm and Achievable x
B How to Make Your Objectives Relevant and New
C How to Make Your Plans Predictable and Light (YLE 2016b)

As for all of the chosen items, the readability index was calculated. First of all, the *required passage* concerning the answer was identified and highlighted in all the texts. Then it could be decided whether the paragraph including the key had 100 words or whether one or more adjacent sentences had to be included in the calculation. With regard to item 11, the first paragraph consisted of only 74 words and not the necessary 100 words. *Text Inspector* also gives a red message when less than 100 words are calculated. This is why for item one the first sentence of the next paragraph was also included in the calculation. This led to 112 words and a readability index of 66.35.

For item 12, the key "It's more […] attainable" is in the third paragraph. As this paragraph does not consist of the necessary 100 words for the calculation of the readability index either, sentences from the upper or lower paragraph had to be included. Concerning the previous paragraph, all sentences of this paragraph would have to be included in order to reach 100 words for the calculation of the readability index. In this case there would have been 112 words. Regarding the following paragraph, 104 words were counted when including two sentences of the following

paragraph. As 104 is closer to the necessary 100 words than 112 words, the third paragraph plus two sentences from the following paragraph were chosen. Then the readability index could be calculated with these 104 words. Hence, a readability index of 64.84 was calculated for the text passage of item 12.

With regard to item 13, the key is at the very end of the short text. As the last paragraph only consists of 56 words, two sentences from the previous paragraph had to be included in order to reach more than 100 words, 101 in this case. Then the text could also be entered into the readability calculator and a result of 59.02 was reached.

Item 14 has an asterisk, because it is a GIST item. The options ask for a suitable title of the text. This is why item 14 was not included in the analysis.

The same procedure was used for calculating the readability index of the necessary text passages of the other MISD items.

4.5. Calculating correlations

Correlation gives information regarding the relationship of two different variables. However, it needs to be stressed that correlation does not illustrate the causality between two variables. If there is correlation between the two variables 1 and 2, this can either be because 1 causes 2 or vice versa. Furthermore, it is possible that another variable 3 causes both 1 and 2. It always needs to be taken into consideration that another variable can be the cause for the relationship of two variables. A famous example is the strong correlation between homicides and the consumption of ice-cream in New York City. However, eating ice-cream does not lead to homicides. The relationship exists due to a third factor: During hot weather, both these variables increase. Therefore, it is advisable to analyze as many variables as possible (Pallant 2007: 122) to find out the relationship between the facility value and the formulation of the multiple-choice items.

With the help of programmes such as *SPSS* a wide range of statistical functions can be calculated (Ebermann 2010). This is why this programme was chosen for calculating the correlations.

After having calculated the correlations in *SPSS*, there appears a table with two important figures. The first one is the correlation coefficient between the two variables that were correlated. The direction can be positive or negative. In case there is a negative sign (-) before the correlation coefficient, this means that there exists a

negative correlation between the two correlated variables (Pallant 2007: 131-133). A negative correlation suggests that the less there is of variable A, the more there is of variable B or vice versa. In contrast, a positive correlation means the more there is of variable A, the more there is also of variable B and vice versa. The second figure in *SPSS* provides information concerning the significance level of the correlation (Pallant 2007: 131-133). These two figures will now be discussed in more detail.

Correlation provides information to which extent two different sets of results are in agreement with each other and this is an important concept for the analysis of tests. The most important figure when dealing with correlations is the correlation coefficient. When there is a correlation of +1.0, there is a perfect positive correlation between two sets of results. By contrast, when there is a correlation of -1.0, this is described as a perfect negative correlation and the two sets of scores are as different from one another as possible. In both cases there is a very strong agreement between the two sets of results. While in the first case it is a positive relationship in the second one it is a negative relationship (Kent State University: n.d.). The correlation coefficient gives information concerning the strength of the relationship. If the correlation is +.05 or below, it is so close to .00 that there is no correlation between the scores and in case there was some relationship this might just be by coincidence. If the correlation is +.70, this suggests that there is a pretty strong relationship between the two results (Kent State University: n.d.). Cohen (1988: 79-83) proposes the following interpretation of the correlation coefficient: small = .10 to .29, medium = .30 to .49 and large = .50 to 1.0. However, different authors propose different interpretations of the correlation coefficient (Pallant 2007: 132). Similarly to Cohen's interpretation, Dancey and Reidy (2007, cited in Green 2013: 85) give the following numbers: 0.7 to 0.9 as strong, 0.4 to 0.6 as mediocre and 0.1 to 0.3 as weak. Correlations below 0.1 are so weak that they can be considered as irrelevant. For this study, Cohen's interpretation of the correlation coefficient was used.

The most widely used correlation coefficient is the Pearson product moment correlation, which is usually calculated with the help of statistical programmes, such as *SPSS*. The bivariate Pearson Correlation calculates a correlation coefficient, r, which indicates the strength of the relationship of two continuous variables. This correlation method assumes that there is the same difference between each score (Alderson, Clapham & Wall 1995: 77-80), which is why it is not always suitable for the calculations (Kent State University: n.d.) as it is the case for this research project. The

Spearman rank order correlation is applied when using ranked or ordinal level data (Pallant 2007: 126). During this kind of correlation the scores are converted into ranks before calculating the correlation (Green 20133: 84) and the data is not normally distributed (Green 2013; 82), which is the case for the data of the present study. For this reason, the Spearman procedure was chosen over the Pearson procedure for calculating the correlations with *SPSS*.

After having chosen the Spearman's correlation regarding the test of significance, a change can be made in SPSS from the already selected two-tailored box to the one-tailored one. If a researcher is already quite sure concerning the direction of the relationship between two variables, the tick would be changed to the one-tailed box (Green 2013: 83). As the direction is not clear, the two-tailed box has been left as it was for the present analysis.

Furthermore, the scores with regard to the variables need to be independent from each other (Green 2013: 82). With regard to this study, the results of the calculations of the different items are independent from each other, because each MISD item tests another idea in the texts.

Exceptions regarding the numbers might have a strong impact on the correlation coefficient. Consequently, the data should be checked for exceptions and whether they occur because of an error. Outliners should not be included in the analysis (Green 2013: 82).

The computer programme *SPSS* measures the significance of a correlation automatically, which is the second relevant number indicated. The level of confidence provides information concerning the per cent to which one can be confident that the results are not because of chance, but that there is an actual relationship between two variables (Green 2013: 84). The significance level does not provide information regarding the strength of the relationship between two variables, which is indicated by the correlation coefficient (Pallant 2007: 133). The probability of the significance level is indicated on a scale of 1, which stands for a strong probability, to 0, which shows no probability. The degree of probability for the results happening to pure chance depends on the situation. Before a football match it seems fair to toss a coin to see which team gets the ball first. There is a 50 per cent chance of probability for each of the two football teams, which seems suitable for such a situation. However, in language testing

it should be aimed at a probability level of at least 95 per cent. At the same time this means that there is a probability of 5 per cent that the wrong decision was made. Depending on the situation, for example, a probability of only 1% that the wrong decision was made could also be chosen. This would be indicated by .01. When there is no probability of making the wrong decision, this is called the null hypothesis (Green 2013: 89). In the case of the present study, it seems suitable to aim for the 5 per cent usually used in linguistics. In *SPSS* the box to get the indication of the significance was ticked in order to obtain automatically information concerning the significance of the results. This figure is indicated with one, two or three stars after the correlation number in *SPSS*. One star means that there is a 5% chance or less that the correlation was just calculated by pure chance. Two or three stars mean that the probability of error is below 1% (Ebermann: 2010). The significance is indicated by the letter 'p', e.g. $p < 0.05$.

The following aspects concerning the results of correlations should be presented. Firstly, the two features[7] whose relationship was investigated need be addressed. Secondly, the way of calculating the correlation needs to be stated, which might be Pearson product moment correlation or Spearman's, with the latter being indicated by *rho*. Thirdly, it needs to be described whether the correlation is positive or negative and whether it is strong, medium or weak. The following numbers should be included: the correlation coefficient (r for Pearson product moment or rho for Spearman's), the number *n* of the samples and *p* (e.g. $p < .05$) with regard to the significance to the result. Furthermore, the relationship of the two features can be stated explicitly. For example, that the numbers of feature 1 are associated with lower levels of feature 2 (Pallant: 2007: 133).

4.6. Definition of features for the calculation of the correlations

This section deals with the chosen features concerning the item formulations for calculating the correlations. The aim was to define as many features as possible for the analysis in order to get a broader range of results with regard to the correlation of the features with the facility value of the items. Twenty-two different features were defined. All of them were correlated in an isolated manner from each other with the facility value of the items.

[7]Scientists such as Freedle and Kostin (1993) worked with the term "variables", whiles for this research project "features" was used. Both terms can be seen as synonyms.

The features can be divided into two groups. Firstly, the ones focusing on lexical aspects and secondly, features dealing with the length of the item formulations.

4.6.1. Lexical features

The aim of taking lexical features of the item formulations into consideration was to find out to which extent more difficult words and expressions relate to the facility values. The programme *Compleat Lexical Tutor* was used in order to calculate most of these features, which are described in the following subchapters. Sample items are used as illustrations of the features. The items can be found in appendix A and the results concerning the lexical features are presented in a table in appendix C.

On the website *Compleat Lexical Tutor* the section "Vocabprofile" was used for the analysis of the lexis. This programme divides the words of a text into four categories based on the degree of frequency at which lexical items are used in English. The first two categories are made up of the 1000 and 2000 most frequently used words. The third category are the academic words and the fourth one are the remaining lexical items, which are not found on any of the previously mentioned lists. The list for the fourth category is called the "off-list" (*Compleat Lexical Tutor*: Vocab Profilers). For instance, proper names might be found in the "off-list". The words from the off-list were, however, not included in the analysis, as a bit more than the majority of the 76 items does not include any words from the off-list and hence more complicated calculations would have been necessary in order to find out about how the words from the off-list correlate with regard to the facility value of the items. The first thousand words are being referred to as K1, while K2 stands for 2000 words. Concerning the remaining lexical items, they are classified from K-3 (3000 words) until K-25 (25 000 words).

4.6.1.1. Type-token ratio and tokens per type

One of the lexical features analysed is the 'type token ratio'. When analysing texts, all the words of a piece of writing put together are being referred to as 'tokens'. The different words within a text are called 'types'. Common words such as 'the' tend to occur more often in texts. The type token ratio is the number of types to the amount of tokens. For instance, there might be 1.000 types in a text. As some words occur several times, there might be 2.000 tokens. In this case, the type-token ratio would be 1/2 or 0.5 (Matthews 2015).

‘Tokens per type’ means how many tokens appear for a certain type. Frequently used words might appear more often. Such words are, for example, pronouns or articles. The *Compleat Lexical Tutor* calculates the tokens per type automatically. For example, item 1 below consists of 29 tokens and 27 types. In this case, the pronoun *she* appears at the beginning of each option and is underlined in the following item. Hence, *she* counts as three tokens, but only as one type. The calculated tokens per type are 1.07, which means that for each type there are on average 1.07 tokens.

> What do we learn about the person described?
> A She comes from a long line of academics
> B She holds a master’s degree in education
> C She is dedicated and goal-oriented x (YLE 2017a)

4.6.1.2. K-1/K-2/K-3 words

The next three categories refer to the frequency of words, which is based on the “Vocab profile” (*Compleat Lexical Tutor*: Vocab Profilers. For example, item 1 of the spring date of 2017, can be classified according to the following K-levels after having copied the item into *Compleat Lexical Tutor*:

> What do we learn about the person described?
> A She comes from a long line of academics
> B She holds a master’s degree in education
> C She is dedicated and goal-oriented (YLE 2017a)

This item consists of

- 21 K1-types: what, do, we, learn, about, the, person, she, comes, from, a, long, line, of, holds, master, degree, in, education, is, an
- 2 K2-types: described, goal
- 3 K3-types: academics, dedicated, oriented
- No K-4 or more challenging types.

By contrast, there are also items, such as item 8, which consist of several types above K-3:

> What motivated the experiment?
> A The urge to provoke x
> B The need to soothe
> C The necessity to conform (YLE 2017a)

The following types are above K-3 level:

- K-4 types: necessity, conform
- K-5 types: soothe

Interestingly, the difficult lexical items appear in the distractors and could be avoided by simplifying the false options.

The following item 75 should also illustrate the usage of off-words. Apart from 3 K-3 types there is also a K-6 type in this item:

> How is the type of ferrofluid invented by Dr Hawkett's team important for Dr King's project?
> A It has already been applied in a number of clinical trials
> B It may provide the key to successful performance x
> C It can hamper the project's funding prospects (YLE 2015c)

- K-2 types: project, applied, trials, provide, successful, performance, funding
- K-3 types : invented, clinical, prospects
- K-6 types: hamper
- Off-list: ferrofluid, Hawkett

With regard to the "off-list", 'ferrofluid' is a lexical item not found in dictionaries. 'Hawkett' is the name of the doctor, which is why it appears in the "off-list".

4.6.1.3. Lexical density

This concept refers to the percentage of lexical items in a text which can be classified as lexical and not as grammatical (Matthews 2014). The lexical density was calculated automatically with the help of the website *Analyse My Writing.* This site refers to the lexical words also as content words. They are made up of the following parts of speech: substantive, verbs, adjectives and adverbs. Other words such as articles are more grammatical and carry less meaning with regard to the content of a text. These words are referred to as non-lexical or functional words of a text. Auxiliary verbs are also considered to be functional words, as they do not offer an additional meaning. The lexical density is, therefore, the percentage of the words giving the important aspects of the meaning of what is being communicated. Concerning writing, lexical density is an indicator with regard to how informative a certain text is (*Analze My Wrting*: n.d.). As functional words are short and occur frequently, it can be assumed that examinees are familiar with them. A text with a higher lexical density might, usually, be harder to process than a text containing a lot of non-lexical words. This is also because the number of content words is much larger, while the amount of grammatical words is quite limited and therefore the latter are well-known to examinees.

The calculation of lexical density is illustrated by item 12 from appendix A:

> Why does it pay off to cut objectives into pieces?
> A They become more updated
> B They become more important
> C They become more doable x (YLE 2016b)

Working with the PC programme *Analyze My Writing,* the following content and grammatical tokens were identified:

- 13 content tokens: pay, cut, objectives, pieces, become, more, updated, become, more, important, become, more, doable
- 9 grammatical tokens: why, does, it, off, to, into, they, they, they

This illustrates a majority of content types in this item. The lexical density for this item amounts to 59.09%.

4.6.2. Item length and number of clauses

The following features refer to the length of items. Sample items are also included as illustration.

4.6.2.1. Number of words, syllables and characters of the multiple-choice items

The number of words, syllables and characters was calculated for the following five parts of each item in order to find out about the significance of these aspects:

1. the entire item
2. the stem
3. all three options
4. the key
5. the two distractors

The last aspect was included, because it turned out to be significant according to the findings of Freedle & Kostin's study (1993: 36).

In order to get the results, the relevant part of the item (e.g. the stem) was copied into the programme *How Many Syllables*, which then provided information concerning the number of words, syllables and characters. The letters A, B and C of the options of each of the multiple-choice items were excluded from this analysis.

The calculations were made as follows. Words are separated by space characters. This means that the expression "term's" counts as a single word in the following item 5. In the same way, the contracted form "she's" would count as well as a single word and only as one syllable. Spaces between words and sentences also count as characters.

Punctuation such as the question mark in each of the items is also counted as a character. There are no dots at the end of the options in the items.

The calculation of the words, syllables and characters are also illustrated by item 12, which can be found in the appendix. The results of the features with regard to the item length can be found in appendix D. Due to the fact that the letters *A, B* and *C* of the items were excluded from this analysis, as they are not relevant for the content of the item, the entire item was pasted in the following form into the programme *How Many Syllables*:

Why does it pay off to cut objectives into pieces? They become more updated They become more important They become more doable (YLE 2016b[8])

According to the automatic calculation by *How Many Syllables,* the results for the entire item 12 are as follows:

- 22 words,
- 35 syllables,
- 126 characters

When only the stem of the item above is analysed (Why does it pay off to cut objectives into pieces? (YLE 2016b)), the following results are reached:

- 10 words,
- 14 syllables,
- 50 characters

When analysing the three options of the item above (They become more updated They become more important They become more doable (YLE 2016b)), there are

- 12 words,
- 21 syllables,
- 75 characters

With regard to the key C (They become more doable (YLE 2016b)), the results are as follows:

- 4 words,
- 7 syllables,
- 23 characters

[8] The original layout oft he item has been adapted for the analysis.

Finally, also the two distractors A and B (They become more updated They become more important (YLE 2016b)) were pasted into *How Many Syllables*, which led to the following results:

- 8 words,
- 14 syllables,
- 51 characters

All of the 76 items were analysed in this manner in order to get the necessary data for then correlating the features with the facility values of the items.

4.6.2.2. Number of clauses

A clause consists of a verb plus the elements which accompany the verb. The main clause is seen as the entire sentence in which a subordinate clause can be included. This can be illustrated by using brackets (Matthews 2014) as the following example shows: [She explained [she saw him]]. In a subordinate clause, the verb is a subordinate verb, the subjects is a subordinate one, etc. Such clauses are also referred to as *dependent* or *lower clause* (Matthews 2014). Clauses are linked by connectives (Mathews 2014). Generally speaking, it can be assumed that an item becomes harder to process if it contains more clauses. In fact, most of the items analysed are short and contain four clauses: one in the stem and one in each of the three options. This is illustrated with item 17 below:

> What characterizes these innovators?
> A They are conventional
> B They make a profit
> C They provide a service x (YLE 2016b)

In this item the stem as well as all of the three options consist of only a single clause. This means that the entire item with the stem and the three options consists of four clauses. The calculation of the number of clauses for each of the items can be found in the last column of appendix D.

Some items consist of more clauses, however. This is, for instance, the case with regard to item 26:

Why does Waugh quote Mahatma Gandhi?
A To parallel a non-violent revolution and the eventual outcome of the debate x
B To emphasize the importance of a significant thinker who studied Shakespeare's works
C To demonstrate his knowledge of relevant changes as regards literary history (YLE 2016a)

In item 26 both, the stem and option A are made up only of one clause. In A, two NP are linked with the connector *and*. Each of the options B and C consists of two clauses. In B, the relative clause is the subordinate clause to the main clause. Similarly, option C also consists of two clauses. Hence, this item is made up of six clauses in total.

4.7. Calculation of the data

In their post-test analysis, the *Finnish Matriculation Examination Board* collects a considerable amount of data of the results of the examinees of all the different subjects. Concerning the data of the Finnish L1 examinees, the results are separated with regard to the weaker and stronger performers and concerning all put together. From this last figure of all the Finnish L1 examinees the facility value was correlated with the 22 features concerning the item formulations.

Due to the high number of examinees during the five exams dates, the facility values of the *Finnish Matriculation Examination* can be argued to be representative. This is because the data includes the examinees from the entire country from the areas where Finnish is used as an L1.

At first, the calculated data was entered into *Excel*. The items were ordered vertically at the very left. The calculations of each feature were put horizontally one after the other in a different column. For each of the items, the 22 features were calculated such as the lexical density, the number of words, syllables, characters of an item, etc. The calculated features as well as the facility values and the readability index for the items are found in appendix B. Then the result of each calculation of the feature was typed into *Excel* for the 76 items.

In *SPSS* the 22 features were defined as variables. The following suggestions for naming the variables were taken into consideration. The variables must all have a different name. Furthermore, they cannot include any punctuation such as full stops,

question marks or exclamation marks and no spaces. Also, all of the variables have to start with a letter and not with a number. The variables cannot have more than 64 characters. And finally, the variable names cannot consist of *SPSS* commands such as 'by', 'or', 'and', 'not', 'with' or 'all' (Pallant 2007: 12).

It is possible to transfer the data from *Excel* directly into *SPSS*, which was done for calculating the correlations in *SPSS*.

5. Results

With the help of *Excel*, the mean value of the chosen 76 items of the *Finnish Matriculation Examination* was calculated, which is 72,48. This shows that almost three quarters of the examinees answered the items correctly. The facility values of the chosen items cover a spectrum from 35 (item 17) to 95.3 (item 62). The results for each item, also concerning the readability index, can be found in appendix B.

The readability index of the selected items ranges from 35.3 (item 35) to 69.2 (item 48). Items with higher or lower readability values were excluded from the analysis in order to work with similar items to find out about the significance of the item formulations.

The following figure 2 presents Spearman's correlations between the facility value of the 76 items and the 22 features. The first column numbers the features and the second one shows the name of the analysed features. In the third column, the correlation coefficient is given. In the fourth column, the significance of the correlation is indicated. In the fifth column, there are comments concerning the level of significance. At first, the features dealing with regard to lexical aspects are listed (number 1-6). This is followed by feature 7 concerning the number of clauses. The rest of the features deal with the length of the item.

No.	Feature	Correlation coefficient rho	Significance p	Comment
1	Type- token ratio	0.041	.724	
2	tokens per type	-.048	.680	
3	K-1 words	.096	.409	
4	K-2 words	-.045	.701	

5	**K-3 words**	**-.298****	**.009**	****p < .01 (2-tailed)**
6	**Lexical density**	**-.235***	**.041**	***p < 0.05 (2-tailed)**
7	Number of clauses	-.159	.171	
8	Number of words	-.020	.861	
9	Number of syllables	-.130	.264	
10	Number of characters	-.148	.201	
11	Number of stem words	-.127	.275	
12	**Number of stem syllables**	**-.231***	**.045**	***p < 0.05 (2-tailed)**
13	Number of stem characters	-.182	.116	
14	Number of option words	-.002	.989	
15	Number of option syllables	-.096	.411	
16	Number of option characters	-.098	.400	
17	Number of key words	.001	.994	
18	Number of key syllables	-.088	.448	
19	Number of key characters	-.102	.381	
20	Number of words in false options	-.030	.794	
21	Number of syllables in false options	-0.122	.292	
22	Number of characters in false options	-.084	.472	

Figure 2: Results of the correlations between the 22 features and the facility values

As figure 2 illustrates, only three out of the 22 analysed features have a significant correlation coefficient. These are written in bold in figure 2. Two of the significant features are the K-3 words and lexical density with quite similar results. For the K-3 words the correlation coefficient rho is -.298. The rho of lexical density is -.235. Both of these correlations are negative. This means that the more K-3 words there are and the higher the lexical density is, the lower the facility value. In both cases

there is a weak correlation as it is below 0.3. Especially with regard to the K-3 words, the .3 level for a medium correlation coefficient is almost reached. Both of these features have significant results. The one for the K-3 words has two asterisks, which indicates that the risk the correlation happened only due to chance is less than 1 per cent (p = 0.01). The significance value for the K-3 words p is .009. Concerning lexical density, there is only one asterisk next to the correlation coefficient. This means that there is a 5 per cent chance that the correlation occurred to chance. The significance value p is .041.

The third feature showing a significant correlation is "number of stem syllables". It also shows a negative correlation. This means that the higher the number of syllables is, the lower the facility value of the items. The correlation coefficient of this feature rho is -.231, which indicates a weak correlation as it is also below .3. There is one asterisk, which shows that there is a certain level of significance. The p-value of this feature is .045, which is just a bit below .05. This means that there is a 5 per cent chance that the correlation happened only by pure chance. Two more features were analysed regarding the significance of the stem: the number of stem words as well as the number of stem characters. By contrast to the number of stem syllables, these two features did not show any significant results. The correlation coefficient rho for the number of stem words is -.127 (p = .275) and for the number of stem characters it is -.182 (p = .116). No asterisk indicated by *SPSS* that there would be a significant correlation.

All of the further 19 features did not show any significant results. Concerning the lexis, this concerns the K-1words (rho = .096, p = .409), K-2 words (rho = -.045, p = .701), the type token ratio (rho = .041, p = .724) and the tokens per type (rho = -.048, p = .680). With regard to the features dealing with the length of the items, the number of stem syllables was the only one with a significant correlation. All of the other features concerning length have very low correlation coefficients. The following features have a correlation coefficient below 0.1 and have therefore no correlation according to Cohen (1988: 79-81): number of words (rho = -0.20, p = .861), number of option words (rho = -.002, p = .989), number of option syllables (rho = -.096, p = .411), number of option characters (rho = -.098, p = .400), number of key words (rho = .001, p = .994), number of key syllables (rho = -.088, p = .448), number of key words in false options (rho = -.030, p = .794) and number of characters in false options (rho = -.084, p = .472). Apart from the already mentioned two features number of stem words

and number of stem characters, a correlation coefficient above 0.99 with an insignificant p-value concerns the following features with regard to item length: number of clauses (rho = -.159, p = .171), number of syllables (rho = -.130, p = .861), number of characters (rho = -.148, p = .201), number of key characters (rho = -.102, p = .381) and number of syllables in false options (rho = -.122, p = .292).

More detailed results regarding all the calculations done for this analysis can be found in the appendices.

6. Discussion of the results

In this section the described results of the study are discussed. The reported findings of Freedle & Kostin's study (1993) as well as relevant theoretical aspects are also taken into consideration.

Only three out of the 22 features show significant results. Two of the features concern lexical difficulty. The weak negative correlation of both K3-words as well as lexical density with the facility values supports the hypothesis concerning this research project: There is a relationship between the usage of more difficult words in item formulations and the facility value. This means that item writers need to be aware that the use of more difficult words in an item may have an influence on the facility values. This seems to make sense intuitively and has now been confirmed by this very limited set of data, with only a weak correlation coefficient, however. In contrast, the facility values of the items are representative with regard to the Finnish L1 speakers, as the results of all the Finnish L1 examinees were included in the calculation of the facility values. The K-3 figure showed already significant results, while the figures concerning more frequently words (K1- and K2-words) did not do so. The rho for K-1 words is only .096 (p = .409) and for K-2 words (p = .701) there is an insignificant negative correlation of -.045. However, K-3 words only refer to the 3.000 most frequently words, which is not yet a high level. More difficult lexis in the items could not be analyzed in this study as the majority of items did not contain more difficult words than K3 words. Furthermore, the analysis of more infrequent words would have required more complex calculations. However, based on the results of this study, it can be assumed that K-4, K-5, K-x words lead to even higher negative correlations regarding difficult words in items and the item facility value.

It is also noteworthy that lexical difficulty did not turn up as a significant variable in Freedle & Kostin's research (1993). Especially with regard to nested items their corpus of items was larger with n=213, which means that their results are probably more representative. This suggests that further research in this field with a larger corpus than the one of this study might be necessary in order to see why there is a discrepancy between Freedle and Kostin's study and this research project with regard to more difficult words used in item formulations.

In order to illustrate lexical difficulty in the items with an example, the following item 12 (according to the numbering in the appendix A) includes two K-3 words.

Why does it pay off to cut objectives into pieces?
A They become more updated
B They become more important
C They become more doable x (YLE 2016b)

The two K-3 words are "objectives" and "updated". Furthermore, there is an "off-word", which does not appear in any of the other lists between the K-1 and K-25 words. This lexical item is "doable." It could be tried to simplify these two K-3-words due to the relationship between difficult lexis in items and the facility value according to the present study. However, there might be certain reasons why test developers chose these formulations, which outsiders cannot know. The usage of other words might have created other issues with this item. Clearly, the test developers tried to avoid the words from the required passage containing the necessary information for understanding the item, which is: "It's more […] attainable" (YLE 2016b). It could be assumed that most B2 examinees are able to guess the meaning of "doable" as it consists of the simple verb "do" and the frequent productive suffix "-able", which examinees will be familiar with from other words such as "available". Having this in mind, "doable" might be easier than the understanding of the probably more challenging word tested in the text, which is the K-4 word "attainable". The K-3 word "updated" appears in distractor A and could be avoided by simplifying the formulation of the distractor.

It was not hypothesized beforehand that the number of stem syllables would show a weak correlation with regard to the facility value in this study. This number is, however, not supported by the other two features analyzed concerning the stem (number of stem words and characters), which both did not show a significant correlation. Concerning the number of stem words the correlation coefficient is -.127 ($p = .275$) and regarding the number of stem characters the correlation coefficient is -

.182 (p = .116). Apart from the number of syllables per stem, all the other 12 features dealing with the length in this study did not show any significant results either. This means that the weak correlation of the syllables per item is not backed up by any of the other 14 features dealing with item length. Also Freedle and Kostin's study (1993: 36) did not find this feature to be important in their study. This is why the figure of the syllable correlation was also checked to see whether an error occurred, but none was found. Due to these reasons it seems that the weak correlation of the stem syllable with the facility value could be neglected. Based on the findings, the assumption could be made that it might be better that the item is a bit longer than including difficult lexis. Hence, if an exam developer has to decide between a more complex or a simpler but longer formulation, it could be suggested that it might be better to go for the longer and simpler formulation, having the relationship between difficult lexis and the facility value in mind.

Apart from the negations, which could not be analyzed with this set of data, because there appear hardly any negations, there is only one further feature that Freedle & Kostin (1993: 36) found to be significant with regard to the items. This refers to the number of words in the distractors. However, the correlation was very weak with rho = .14 ($p < 0.05$, 2-tailed) for the nested items and .23 ($p < 0.05$, 2-tailed) for the non-nested sample. The sample size of the non-nested items with n=98 is a bit larger, but quite similar to the sample of the present study (n=76). Because of the findings in Freedle & Kostin's study, the number of words, syllables and characters in the distractors was also included in this research project. Yet, no significant results were found concerning the length of the false options. In this study, for the number of words in the distractors the rho = -.030 (p= .794), for the syllable in false options rho = -0.122 (p= .292) and for the characters in the distractors rho = -.084 (p= .472). This is backed by the fact that all of the other features concerning length did not show any significant results either, apart from the number of syllables in the stems with a weak correlation. One might wonder why the length of the distractors showed a correlation in Freedle & Kostin's study, although it is only weak. Further research could go into this direction to find out more about the discrepancy between the significance of the number of words in the distractors in Freedle & Kostin's study and this research project, which came to different findings.

The usage of difficult lexis in item formulations might have an impact beyond the exam itself, which refers to teaching and test preparation. A test can have positive

or negative consequences, which is being referred to as washback effect. This concerns individuals such as teachers and students learning for tests but also institutions such as schools or the Ministry of Education. A negative effect is also given, for instance, when teaching has the only aim of preparing students for a test, which is referred to as teaching to the test. A positive effect exists when learners have the feeling of being treated fairly and that they had to acquire useful competences for a test (Krause & Sändig 2002: 90). Paying attention to easy formulations in the development of school-leaving examinations might have a positive washback effect on foreign language learning in the classroom, when teachers, for example, also pay more attention to make use of easier item formulations in the tests that they put together.

7. Conclusion

This final section should contain the most important findings of the study and a discussion of their relevance for this research field. Moreover, further research gaps should be addressed (Esselborn-Krumbiegel 2010: 145-148).

After having excluded as many factors as possible that influence the facility value, 76 similar MISD items could be identified for the analysis. Having chosen comparable items, the relationship between difficulty item formulations and the facility value could be analyzed. It turned out that only three out of the 22 features concerning the item formulation correlate with the facility values. Two features show that there is a weak negative correlation between difficulty lexis in the items and the facility value. Furthermore, the number of syllables in the stem also showed a weak negative correlation concerning the facility values. However, this figure was neither backed by the other two features dealing with length of the stem nor by any of the other 12 features dealing with item length. This indicates that the significance of this feature might be neglected. The findings suggest that item writers need to be aware of the weak relationship between difficulty item formulations and the facility value. It is also noteworthy that none of these three features was found to be significant in a similar, yet bigger, previous study by Freedle & Kostin (1993). In contrast, they found in their research project that the length of the distractors showed a weak positive correlation with comprehension difficulty. However, this was not supported by any of the three different features dealing with the length of distractors in the present study. This discrepancy regarding the two studies asks for further research in this area. It needs to

be stressed, however, that Freedle & Kostin's research was bigger and consequently also more representative. The limited number of only 76 items in this study is hardly even representative for Finnish L1 speakers. As there were only three weak significant item features in both Freedle and Kostin's (1993: 36) and the present research project, this suggests that item-text and text-only variables are far more significant concerning the facility values, as Freedle and Kostin's study showed. Freedle & Kostin (1993: 36) found many more variables to correlate positively and negatively with comprehension difficulty with regard to text-item overlap variables and text variables as presented in in subsection 2.3.1.2. Therefore, Freedle & Kostin (1993: 1) draw the conclusion that examinees actually have to read the text to be able to answer multiple-choice items. This also implies that the multiple-choice items are, usually, not guessable and that they are a valid measurement for testing reading. This argumentation is supported by the fact that only three out of the 22 item features in the present study correlate weakly with the item facility. This suggests as well that other text-item and text features might be more significant. However, text and text-item features were not analyzed in the present study as the focus was on the item formulation. Hence, based on the results of Freedle & Kostin's (1993: 1, 36) and this study the reading text passage seems to be much more important for answering the item, while the formulation of the items plays a minor role. However, certain features might want to be taken into consideration concerning the item formulations. Test developers might want to avoid negations, which were identified as a significant variable in the study conducted by Freedle & Kostin's (1993: 36). Furthermore item writers might also want to keep in mind that more complex item formulations correlate weakly with item facility. However, it needs to be kept in mind that correlations only show the relationship between two features and do not provide any information concerning the effect of one feature on another, as explained in further detail in subsection 4.5.

The item corpus shows that some items could be formulated in an easier way. In particular in some of the distractors of the multiple-choice items complex formulations appear, which might want to be avoided by rewording the ideas or coming up with alternative MISD distractors. This would also have a positive washback effect on the preparation for the *Finnish Matriculation Examination* at schools as explained in chapter 6. However, there might be certain reasons why test developers chose these more difficult formulations, which outsiders cannot know. It also needs to be stressed that too much teaching to the test might want to be avoided as the preparation for

dealing with test items might not be the most suitable way for developing one's language competence, which was dealt with in subsection 2.3.2. about multiple-choice items.

Some features could not be analyzed in this study, although they might have been relevant. Lexical items being more difficult than K-3 words as well as off-tokens were not included in the present study, because such words only appeared in a small number of items. Hence, further t-tests would have been necessary, which was not possible for this research project. However, as both K-3 words and lexical density showed significant results, the hypothesis can be formulated that even more difficult words in the items might show even higher negative significant correlations with the facility values. Similarly, the significance of connectives could not be analyzed for the same reason as they appeared only in a limited number of items. The analysis of the significance of these features might, however, be interesting for future research projects.

What is representative in this study, are the facility values of all the Finnish L1 speakers taking the *Finnish Matriculation Examination* in English. Since the *Finnish Matriculation Examination Board* calculates the results of the examinees having Finnish or Swedish as a language of instruction in different tables, in a further study it might also be interesting to see whether the results of the Swedish L1 speakers are better. This could be assumed to be the case because Swedish is a Germanic language, like English, although Swedish is a North-Germanic language, whereas English is a West-Germanic language like German.

Multiple-choice items often appear in *high-stakes tests*, which might have far-reaching consequences for the examinees, who might therefore be under great pressure in such test situations. This is why research in the development of test methods such as multiple-choice seems important in order to provide item writers with guidelines about how to develop items that can measure the competence of examinees as successfully as possible.

8. References

Alderson, Charles. 2005. *Assessing reading.* Cambridge: Cambridge University Press.

Alderson, Charles; Clapham, Caroline; Wall, Dianne. 1995. *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Analyse My Writing. http://www.analyzemywriting.com/index.html [6 December 2019]

Brown, Douglas; Abeywickrama, Priyanvada. 2010. *Language assessment: principles and classroom practises.* White Plains: Pearson Education.

Brown, James Dean; Hudson, Thom. 2002. *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Bredella, Lother. 2007. „Textrezeption und Textproduktion im Rahmen rezeptionsästhetischer Literaturdidaktik“ [“Text reception and text production in the context of aesthetic literary didactics“]. In Bausch, Karl-Richard; Burwitz-Metzer, Eva; Königs, Frank; Krumm, Hans-Jürgen. *Textkompetenzen [Text competences]*. Thübingen: Narr. 26-38.

Brizic, Katharina. 2006. “Das geheime Leben der Sprachen. Eine unentdeckte migrantische Bildungsressource [The secret life of languages. An undiscovered migratory educational resource]“. www.kurswechsel.at 2. 32-43. [16 October 2018]

Brown, Annie; Davies, Alan; Elder, Cathie; Hill, Kathryn; Lumley, Tom; McNamara, Tim. 1999. *Dictionary of language testing.* Cambridge: Cambridge University Press.

Buck, Gary. 2011. [10th edition]. *Assessing listening.* Cambridge: Cambridge University Press.

Buttaroni, Susanna; Knapp, Alfred. 1988. “Fremdsprachenwachstum”. In: Fernkurse der Wiener Volkshochschulen. 34-43, 68-70.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioural sciences.* Hillsdale: Lawrence Earlbaum Associates.

Cohen, Louis; Manion, Lawrence; Morrison, Keith. 2007. “Planning educational research“. In Cohen, Louis; Manion, Lawrence; Morrison, Keith. *Research methods in education*. London: Routledge. 78-99.

Compleat Lexical Tutor. https://www.lextutor.ca/ [13 January 2020]

Compleat Lexical Tutor - Vocab Profilers: some research uses of vocabprofile (VP). https://www.lextutor.ca/vp/research.html [13 January 2020]

Council of Europe. 2018. *Common European framework of reference for languages: learning, teaching, assessment – companion volume with new descriptors*. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989 [15 December 2019]

Council of Europe. 2001. *Common European framework of reference for languages: learning, teaching, assessment*. Strasbourg: Cambridge University Press. https://rm.coe.int/1680459f97 [3 October 2018]

Daase, Andrea; Hinrichs, Beatrix; Settinieri, Julia. 2014. ”Befragung” [“Questioning]”. In Settinieri, Julia; Demirkaya, Sevilen; Feldmeier, Alexis; Gültekin-Karakoc, Nazan; Riemer, Claudia. *Einführung in empirische Forschungsmethoden für Deutsch als*

Fremd- und Zweitsprache [Introduction to empirical research methods for German as a foreign and second language]. Paderborn: Schöningh UTB. 103-121.

Ebermann, Erwin. 2010. *Grundlagen statistischer Auswertung [The basics of the analysis of statistical data.* Institut für Kultur- und Sozialanthropologie. https://www.univie.ac.at/ksa/elearning/cp/quantitative/quantitative-108.html [15 November 2019]

Dancey, Christine; Reidy, John. 2007. *Statistics without maths for psychology.* Harlow: Prentice Hall.

Davis, Larry; Xi, Xiaoming. 2016. „Quality factors in language assessment". In Banerjee, Jayanti; Tsagari, Dina. *Handbook of second language assessment.* Boston: Walter de Gruyter Inc. 12. 61-76.

Douglas, Dan. 2010. *Understanding language testing.* New York: Hodder Education.

Esselborn-Krumbiegel, Helga. 2010 *Richtig wissenschaftlich arbeiten: Wissenschaftssprache in Regeln und Übungen [Working scientifically correctly: rules and practice of scientific lamguage]*. Paderborn: Ferdinand Schöningh.

Feld-Knapp, Ilona; Schoßböck, Judith. 2010. „Textwelten erkennen lernen. Zu notwendigen Lehrendenkompetenzen bei der Arbeit mit aktueller österreichischer Gegenwartsliteratur im DaF-Unterricht" ["Recognizing text worlds: about the necessary competences of learners concerning current Austrian literature in the German as a foreign language classroom]. 13. 115-135.

Feld-Knapp, Ilona. 2005. *Textsorten und Spracherwerb. Eine Untersuchung der Relevanz textsortenspezifischer Merkmale für den "Deutsch als Fremdsprache"-Unterricht [Text types and language acquisition. A survey of the relevance of text type characteristics for the teaching of German as a foreign language]*. Hamburg: Dr. Kovac.

Figueras, Neus. 2912. "The impact of e CEFR". In *ELT Journal.* 66. 4. 477-485.

Flesch, Rudolf. 1948. "A new readability yardstick". In *Journal of Applied Psychology* 32. 3. 221-233.

Freedle, Roy; Kostin, Irene. 1993. *TOEFL research reports: the prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: main idea, inference and supporting idea items.* Report 44. Princeton: Educational Testing Service.

Fulcher, Glenn. 2010. *Practical language testing.* New York: Routledge.

Fulcher, Glenn; Davidson, Fred. 2007. *Language testing and assessment - an advanced resource book.* New York: Routledge.

Fulcher, Glenn; Davidson, Fred. 2006. *Language testing and assessment.* London: Routledge.

Glaboniat, Manuela. 2019. „Sprachprüfungen ‚made in Austria' als Beitrag zur Förderung der deutschen Sprache [Language exams ‚made in Austria' as a contribution to fostering the German language]". In Ammon, Ulrich; Schmidt, Gabriele. *Förderung der deutschen Sprache weltweit [Fostering the German language globally]*. Berlin/Bosten: Walter de Gruyter. 407-421.

Glaboniat, Manuela. 1998. *Kommunikatives Testen im Bereich Deutsch als Fremdsprache: Eine Untersuchung am Beispiel des Österreichischen Sprachdiploms*

Deutsch [Communicative testing in German as a foreign language: an investigation using the example of the Austrian language diploma German]. Innsbruck: Studien Verlag.

Glaboniat, Manuela; Peresich, Carmen. 2018. "Österreichisches Sprachdiplom Deutsch" ["Austrian language diploma German"]. In Sigott, Günther. *Language testing in Austria: taking stock – Sprachtesten in Österreich: eine Bestandsaufnahme.* Berlin: Peter Lang. 349-368.

Grabe, William; Stoller, Frederika. 2002. *Teaching and researching reading.* Harlow: Pearson Education.

Green, Rita. 2017. *Designing listening tests: a practical approach.* London: Macmillan Publishers.

Green, Rita. 2013. *Statistical analysis for language testers.* New York: Palgrave Macmillan.

Hinger, Barbara; Stadler, Wolfgang. 2018. *Testen und Bewerten fremdsprachlicher Kompetenzen [Testing and assessing foreign language competences]*. Tüningen: Narr Francke Attempto.

How Many Syllables. https://www.howmanysyllables.com/ [6 December 2019]

Kecker, Gabriele; Depner, Günther; Marks, Daniela; Schwarz, Leska; Zimmermann, Sonja. 2019. „Die deutsche Sprache weltweit fördern: Was können Sprachprüfungen dazu beitragen?" ["Promoting the German language globally: How can this be supported by language tests?]". In Ammon, Ulrich; Schmidt, Gabriele. *Förderung der deutschen Sprache weltweit [Fostering the German language globally]*. 393-404.

Kecker, Gabriele. 2011. *Validierung von Sprachprüfungen: Die Zuordnung des TestDaF zum Gemeinsamen europäischen Referenzrahmen für Sprachen. [Validation of language examinations: assigning TestDaF to the Common European Framework of Reference for Languages]*. Frankfurt am Main: Peter Lang - Internationaler Verlag der Wissenschaften.

Kent State University. "SPSS tutorials: Pearson correlation". https://libguides.library.kent.edu/SPSS/PearsonCorr [7 December 2019]

Khalifa, Hanan; Weir, Cyril. 2009. *Studies in language testing: examining reading.* Cambridge: Cambridge University Press.

Krause, Wolf-Dieter; Sändig, Uta. 2002. *Testen und Bewerten kommunikativer Leistungen im Unterricht Deutsch als Fremdsprache: Linguistische Grundlagen und didaktische Angebote [Testing and assessing performances in German as a foreign language: linguistic basics and didactic approaches]*. Frankfurt am Main: Peter Lang - Europäischer Verlag der Wissenschaften.

Krumm, Hans-Jürgen. 2006. "Lernen lehren – Lehren lernen: Schwierigkeiten und Chancen des autonomen Lernens im Deutschunterricht [Teaching learning – learning teaching: difficulties and chances of autonomous learning in the German classroom]. In Feld-Knapp, Ilona. *Budapester Beiträge zu Deutsch als Fremdsprache [Budapest contributions to German as a foreign language]*. 1. 60-76.

Matthews, Peter. 2015. *A dictionary of psychology*. [4th edition]. Oxford: Oxford University Press.

Matthews, Peter. 2014. *The concise Oxford dictionary of linguistics.* [3rd edition]. Oxford: Oxford University Press.

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

Messick, Samuel. 1989. Validity. In: Robert L. Linn (ed.). *Educational measurement,* 3rd ed. 13-103. New York: Macmillan. Quoted in Davis, Larry; Xi, Xiaoming. 2016. „Quality factors in language assessment". In Banerjee, Jayanti; Tsagari, Dina. *Handbook of second language assessment.* Boston: Walter de Gruyter. 12. 61-76.

Milton, James. n.d. "The development of vocabulary breath across the CEFR levels: A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe. In *Eurosla monographs series 1: communicative proficiency and linguistic development.* 211-232.

Moss, Pamela A. 2003. Reconceptualizing validity for classroom assessment. *Educational*
Measurement: Issues and Practices 22(4): 13–25. Quoted in Davis, Larry; Xi, Xiaoming. 2016. „Quality factors in language assessment". In Banerjee, Jayanti; Tsagari, Dina. Handbook of second language assessment. Boston: Walter de Gruyter. 12. 61-76.

Moss, Pamela A. 2013. Validity in action: Lessons from studies of data use. *Journal of Educational Measurement* 50(1): 91–98. Quoted in Davis, Larry; Xi, Xiaoming. 2016. „Quality factors in language assessment". In Banerjee, Jayanti; Tsagari, Dina. *Handbook of second language assessment.* Boston: Walter de Gruyter. 12. 61-76.

Nation, Ian. 2001. *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Neuland, Eva; Peschel, Corinna. 2013. "Textverstehen" ["Text comprehension"]. In *Einführung in die Sprachdidaktik [Introduction to language didactics]*. Stuttgart: J.B. Metzler. 159-193.

Neuner, Gerhard. 2003. „Das Konzept der Mehrsprachigkeitsdidaktik" [„The concept of the didactics of plurilingualism]". In Hufeisen, Britta; Neuner, Gerhard. Mehrs*prachigkeitskonzept – Tertiärsprachen - Deutsch nach Englisch [Concept of plurilingualism - tertiary languages - German after English]*. Kapfenberg: Bachernegg.13-34.

Pallant, Julie. 2007. *SPSS survival manual: students' favourite.* New York: Open University Press.

Paribakht, Sima; Wesche, Marjorie. 1999. "Reading and 'incidental' L2 vocabulary acquisition - an introspective study of lexical inferencing". In *Studies in second language acquisition.* 21. 195-223.

Portmann, Paul; Schmölzer-Eibinger, Sabine. 2008. „Textkompetenz" ["Text competence"]. In *Fremdsprache Deutsch: Zeitschrift für die Praxis des Deutschunterrichts [German as a foreign language: journal for the practise of teaching German]*. München: Klett. 39. 5-17.

Riemer, Claudia. 2014. "Forschungsmethodologie Deutsch als Fremd- und Zweitsprache" [„Reasearch Methodology in German as a foreign and second language"]. In Settinieri, Julia; Demirkaya, Sevilen; Feldmeier, Alexis; Gültekin-Karakoc, Nazan; Riemer, Claudia. *Einführung in empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache. [Introduction to empirical research methods for German as a foreign and second language]*. Paderborn: Schöningh UTB. 15-30.

Schmelter, Lars. 2014. „Gütekriterien" ["Indicators of quality"]. In Settinieri, Julia; Demirkaya, Sevilen; Feldmeier, Alexis; Gültekin-Karakoc, Nazan; Riemer, Claudia.

Einführung in empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache [Introduction to empirical research methods in German as a foreign and second language]. Paderborn: Schöningh UTB. 33-45

SPSS. Version 26.

Tsagari, Dina, & Banerjee, Jayanti. 2016. *Handbook of second language assessment*. Boston: De Gruyter Mouton.

Text Inspector. https://textinspector.com/ [13 January 2020]

Wolff, Dieter. 1996. "Kognitionspsychologische Grundlagen neuer Ansätze in der Fremdsprachendidaktik" ["Cognitive-psychological basis of new approaches in foreign language didactics]. *Info DaF [Info German as a Foreign Language]*. 23. 541-560.

Ylioppilastutkintolautakunta - studentexamensnamnden a. "Matriculation Examination". https://www.ylioppilastutkinto.fi/en/matriculation-examination [29 January 2019]

Ylioppilastutkintolautakunta - studentexamensnamnden b. "Matriculation Examination: structure of the examination". https://www.ylioppilastutkinto.fi/en/matriculation-examination/the-examination/structure-of-the-examination [29 January 2019]

Ylioppilastutkintolautakunta - studentexamensnamnden c. "Matriculation Examination: description of tests". https://www.ylioppilastutkinto.fi/en/matriculation-examination/the-examination/description-of-tests [29 January 2019]

Data and primary sources

YLE. English written part pitkä oppimäärä [Long course]. 2017a. https://drive.google.com/file/d/0Bw3oPkjh-TYLUFhMVDV3YnF5QkE/view [4 December 2019]

YLE. Englanti, pitkä oppimäärä, kirjallinen osa [English long course, written part].

2017b. Key. https://drive.google.com/file/d/1goRcZ9E64KkL8NQ7TGWMqWG404Lg1iwP/view [4 December 2019]

YLE. English written part pitkä oppimäärä [Long course]. 2016a.[9] https://drive.google.com/file/d/1jiGJJmd5iD9zyE21y5CAvlFSZ6B5WJ3b/view [4 December 2019]

YLE. English written part pitkä oppimäärä [Long course]. 2016b. https://drive.google.com/file/d/0Bw3oPkjh-TYLakF0VWdmYm55LTA/view [4 December 2019]

YLE. Englanti, pitkä oppimäärä, kirjallinen osa [English long course, written part]. 2016c. Key. https://drive.google.com/file/d/1eiCKQxTG4hAon9vxvtvgHi8m5_YkcPoQ/view [4 December 2019]

[9] Document of the reading key contains the listening key, which is why the key of this reading paper is not listed here.

YLE. English written part pitkä oppimäärä [Long course]. 2015a. https://yle.fi/progressive/fynd/oppiminen/oppiminen.yle.fi/yo-kokeet/eat-s15.pdf [4 December 2019]

YLE. Englanti, pitkä oppimäärä, kirjallinen osa [English long course, written part]. 2015b. Key. https://yle.fi/progressive/fynd/oppiminen/oppiminen.yle.fi/attachments/2015_s_eat_sabl.pdf [4 December 2019]

YLE. English written part pitkä oppimäärä [Long course]. 2015c. https://yle.fi/progressive/fynd/oppiminen/oppiminen.yle.fi/yo-kokeet/eat-k15.pdf [4 December 2019]

YLE. Englanti, pitkä oppimäärä, kirjallinen osa [English long course, written part]. 2015d. Key. https://yle.fi/progressive/fynd/oppiminen/oppiminen.yle.fi/attachments/2015_k_eat_sabl.pdf [4 December 2019]

9. Appendices

9.1. Appendix A: 76 chosen MISD items from the higher *Finnish Matriculation Examination*

Item no.	Exam date	Item
1	2017 kevät 1	What do we learn about the person described? She comes from a long line of academics She holds a master's degree in education She is dedicated and goal-oriented
2	5	What is often typical of marginalia? That they tend to represent parts of conversations That they seem to be meant for substance matter experts only That they are often quite straightforward and matter-of-fact
3	7	What motivates Erik Schmitt to keep doing this? Conserving historical publications Honoring something marginalized Discovering what needs to be marginalized
4	13	What can be said of the cinema goers? They are becoming fewer They appreciate fiction They are well catered for
5	17	What is said about the future of this course of events? An update of the term's history is most likely Extensive research on related terms' history is neededAll the information on the term's history is conclusive
6	19	What describes the role of technology? It won't offer any practical solutions It won't overcome emotional bonds It won't substitute for historical details
7	21	What is said of the exhibition? It was frequented by visitors It hardly attracted viewers It was relatively modestly received
8	23	What motivated the experiment? The urge to provoke The need to soothe The necessity to conform

9	24	Who got to the bottom of the issue? A group of experts Those with a discerning eye Everyone who gave a try
10	25	What's the curator's final opinion of the counterfeit? He grew rather accustomed to it He refrained from taking a stand He continued to dislike the object
11	2016 sypsky 1	How does research view the art of goal setting? It emphasizes the importance of making a plan It highlights the necessity to create multiple plans It outlines the need for back-up plans
12	2	Why does it pay off to cut objectives into pieces? They become more updated They become more important They become more doable
13	3	What is said of willpower? It can be used quickly It must be used wisely It should be used widely
14	5	According to this text, what were the working conditions first like in Sierra Leone? Skilled personnel were necessary to operate the preinstalled systems Expertise was needed to prevent things from collapsing Creativity was required to get everything running
15	6	What is said of the sample handling procedures in Sierra Leone? They are in accordance with specific national criteria They take into account both the scientists and samples They fail to reach the strict standard set
16	7	What is the scientist's take on being called a hero? She is quite overwhelmed by the praise She is rather sensible in her reaction She is very grateful for the attention
17	12	What characterizes these innovators? They are conventional They make a profit They provide a service
18	13	According to this text, who are the intended end-users of the wheelchair? Those who can afford it Those who are in need Those who order one

19	14	On the basis of the first paragraph, why is William Shakespeare to blame? He left no will on paper He signed no papers He left no paper trail
20	15	What is typical of the documents relating to Shakespeare? They date back to the 16th century They represent only one text-type They are representative in their contents
21	16	What is the key point in Alexander Waugh's book? That Shakespeare was a pen-name That Shakespeare was well educated That Shakespeare was one-of-a-kind
22	17	What happened to the debate suggested by SAC? It attracted no funding It went by unnoticed It did not occur
23	18	What does Professor Stanley Wells point out as regards evidence? That the documents concerning Shakespeare seem exhaustive That the case of Shakespeare is typical of its era That Shakespeare's remaining records are unconvincing
24	19	According to Sir Derek Jacobi, how would the end-outcome of the debate make a difference? By contributing to the plays' relevance that defies time By inspiring new playwrights to produce contemporary pieces By emphasizing the texts' atypical contents and datedness
25	20	What does Professor Wells say is worth remembering? There can be researchers who will figure this out There could be studies that have remained unnoticed There may still be things to be found and discovered
26	21	Why does Waugh quote Mahatma Gandhi? To parallel a non-violent revolution and the eventual outcome of the debate To emphasize the importance of a significant thinker who studied Shakespeare's works To demonstrate his knowledge of relevant changes as regards literary history

27	22	What is said about Mr. Crites's career? It has resulted in one-of-a-kind facilities It has involved solitary work It has meant working as an employee
28	23	What is the change described in Mr. Crites's business all about? Settling down Teaming up Letting go
29	24	Why is the label "Ruby Tree" used in this context? To refer to an appreciation of beauty To cite a well-known narrative To emphasize the beauty of stone
30	25	What characterizes Ms. Gray's current view on artwork? She is looking for new co-workers She appreciates craftsmen to a certain extent She is inspired to keep innovating
31	2016 kevät 2	Why did bananas inspire her? They aided her understanding of inequality They provided her an easy career option They offered her a means to undermine locals
32	3	What is said about Rachel Lichte's diamond business? It thrives as an African co-operative Its profits continue to increase It depends on her connections
33	4	What's Rachel Lichte's take on success? It's brought about by optimal settings It requires determination It tends to come effortlessly
34	5	How does the described gadget function? By producing audible pulses By emitting unpleasant pulses By creating undermined pulses
35	11	How is it possible to get all the residents involved in power generating? By showing that the local schools benefit a lot from the results By emphasizing that shareholders will get their due profits By demonstrating that people can impact the end-outcome

36	12	What characterizes the beginning paragraph of this book review? It starts frankly like the novel it describes It exhaustively describes the characters in the novel It introduces the novel's sequence of events thoroughly
37	13	What is said of the novelist's narrative style? It reflects the novel's contents It reveals the plot early on It contradicts the actions described
38	14	What is typical of the issues that the novel deals with? They tackle problems present in Western society which are caused by gender They demonstrate how the deeds of earlier generations may affect the offspring They center on cultural and religious themes highly relevant only in North America
39	16	Why do MOOC providers want to maintain low dropout rates? To recruit optimal tutoring staff To sustain academic standards To continue attracting funding
40	17	What is potentially revolutionary about the MOOC master's degree? Both campus and online degrees are accepted on equal terms Both campus and online degrees are becoming more popular Both campus and online degrees are going to be affordable for everybody
41	18	Why do top-level universities remain unaffected by the described changes? Because of the social standing, respect and connections they offer Because of the high quality of tuition available on campus Because of the qualifications their academics demonstrate
42	23	What does the text say about the origins of singing? They seem thoroughly researched They are clear as demonstrated by research efforts They remain vague despite the research

43	24	According to this text, why is the book's second thesis surprising? Music is becoming increasingly composer-centred Music is reclaiming its composer-free roots Music is continuing to attract young composers
44	2015 sypsky 1	What is seahorses' movement in the water usually believed to involve? Swimming in quick bursts Floating with a clear goal Being rather slow
45	2	What makes the seahorse an effective hunter? The ability to dive deeper than its prey The ability to sneak up on its prey The ability to swim parallel to its prey
46	3	What is said to be typical of seahorses' main food source? It reacts to still water It often feels threatened It remains hard to catch
47	4	What is mentioned about the clinic's structure? It features glass ceilings It floats on water It is easily changeable
48	5	How is the environment said to affect the patients? It creates a feeling of calmness It prevents them from letting go It enhances their unpredictability
49	6	Why does the architect refer to the boat metaphor? To emphasize stability To belittle harmony To highlight equality
50	7	What does the architect say about the use of space in the clinic? It functions as a means of preservation It aims at promoting a sense of liberty It concentrates on preserving values
51	8	What is said about Josh Tetrick's profession? He works as an entrepreneur He specializes in innovative design He runs a family business

52	9	According to the text, why are egg-substitutes necessary? To support farmers To sustain consumption To preserve nature
53	11	What is primarily implied about the immediate future of the business? It will continue to attract funding It will rapidly expand globally It will constantly reduce food prices
54	14	Which one of the following summarizes the third study? Being outgoing intensifies age-related illnesses Developing one's mind requires constant training Socializing aids in giving one's mind a workout
55	17	What is said about technology in brain research? It has its limitations It provides a cure It comes to the rescue
56	18	According to the text, what is the main contribution of this publication? It provides an overview of academically applicable studies It highlights the real-life accounts of informants It bases its argumentation on hands-on experimentation
57	19	What is special about the plant's common name? It originally referred to a particular region It is derived from a plant's Latin-based name It features references to its believed mystical qualities
58	21	What does this text base its treatment of historical details on? It relies on data collected by Harvard Business Review journalists from historical sources It relies on information provided by relevant institutions It relies on knowledge acquired while interacting with peers
59	22	What characterizes the beginning of the use of @? The urge to impress the intended audience The need to produce legible texts The strive for overall efficiency
60	23	What is said of the symbol's development in Europe? It became part of business negotiations It acquired a practical meaning in trade It retained its original commercial meaning

61	2015 kevät 3	According to this text, what characterizes Mr Farinetti's career? He has swapped areas of specialization He chose this line of business already as a young man He has managed to make a high number of profitable career moves
62	4	What does the "slow food movement" mainly concentrate on? Changing how individuals feel about dieting Affecting attitudes towards the origins of ingredients Influencing people as to how often they opt for eating out
63	5	What is said about the future of the company? Business will pick up slightly Business will continue as before Business will bloom
64	7	What is said to be typical of the most famous physicists? They are household names They become infamous They tend to be egoistic
65	8	Why is the comic format primarily preferred for these types of scientific topics? To entertain the readership optimally To honor natural sciences in particular To make the contents easily accessible
66	9	How is Dr Hawking's behavior on TV described? He appears matter-of-fact He comes across as modest He is easily irritable
67	10	What is said of Dr Hawking's personal development over time? He became a hermit He developed narcissistic tendencies He overcame physical difficulties
68	12	What is said about the works featured in this exhibition of the Frick Collection? They are from an American collection They are part of a private collection They are on loan from a national collection
69	14	How is the special hanging of one of Vermeer's works in the text? The painting should have been placed in another space in the museum All the paintings on display deserve to be optimally repositioned Some contenders may surpass the painting in quality

70	16	What characterizes the art work they sell? They are cheap They copy earlier works They appear very exclusive
71	17	What does the text say about the owners reaching their objective? They've taken their time They are still working on it They made it in a relatively short time
72	18	Why does this text start with a riddle? To meet the demands of scientific precision To introduce the writer's main idea To pass the author off as an intellectual
73	20	Why does Dr King mainly do this research on nanosatellites? To extend their usability To make them inexpensive To derail their development
74	21	Why did an "oily hedgehog" shape seem optimal for Dr King's work? It appeared to avoid the third law of motion It allowed escape from magnetic fields It enabled the miniature rocket's movement
75	22	How is the type of ferrofluid invented by D r Hawkett's team important for Dr King's project? It has already been applied in a number of clinical trials It may provide the key to successful performance It can hamper the project's funding prospects
76	23	What is said about the future of Dr King's project? Fundraising is likely to be discontinued The results meet all the expectations The outcome remains yet to be seen

9.2. Appendix B: data regarding the items

item	exam date	number in the original exam	facility value	readability index
1	2017, kevät	1	84,4	59,12
2	2017, kevät	5	84,4	52,76
3	2017, kevät	7	40,9	47,6
4	2017, kevät	13	87,2	55,62
5	2017, kevät	17	73,7	40,78
6	2017, kevät	19	86,8	53,05
7	2017, kevät	21	70,9	34,14
8	2017, kevät	23	60,8	42,16

9	2017, kevät	24	85	50,64
10	2017, kevät	25	86,2	60,78
11	2016, sypsky	1	81,1	66,35
12	2016, sypsky	2	87,9	64,28
13	2016, sypsky	3	86,2	59,02
14	2016, sypsky	5	75,5	59,84
15	2016, sypsky	6	77,9	59,84
16	2016, sypsky	7	79,5	59,84
17	2016, sypsky	12	35	48,15
18	2016, sypsky	13	92	48,15
19	2016, sypsky	14	74,1	56,15
20	2016, sypsky	15	49,1	56,15
21	2016, sypsky	16	78,2	41,74
22	2016, sypsky	17	79,6	41,74
23	2016, sypsky	18	73,6	63,6
24	2016, sypsky	19	56,6	44,13
25	2016, sypsky	20	83,8	59,21
26	2016, sypsky	21	75,6	68,18
27	2016, sypsky	22	69,3	45,97
28	2016, sypsky	23	68,4	45,97
29	2016, sypsky	24	44,2	45,89
30	2016, sypsky	25	89	49,91
31	2016 kevät	2	86,9	47,74
32	2016 kevät	3	64,2	54,29
33	2016 kevät	4	86,9	54,29
34	2016 kevät	5	52,4	56,31
35	2016 kevät	11	77	35,3
36	2016 kevät	12	66,4	39,74
37	2016 kevät	13	54,1	61,9
38	2016 kevät	14	74,8	40,41

39	2016 kevät	16	70,4	46,55
40	2016 kevät	17	63,9	47,64
41	2016 kevät	18	56,1	43,64
42	2016 kevät	23	76,3	43,86
43	2016 kevät	24	65,2	43,86
44	2015 sypsky	1	79,6	57,73
45	2015 sypsky	2	87,5	57,73
46	2015 sypsky	3	74,9	57,73
47	2015 sypsky	4	77,4	51.76
48	2015 sypsky	5	88,6	69.20
49	2015 sypsky	6	77,4	68.78
50	2015 sypsky	7	73,8	58,58
51	2015 sypsky	8	83,2	66.27
52	2015 sypsky	9	88,4	54.02
53	2015 sypsky	11	37,5	55,4
54	2015 sypsky	14	66	40,82
55	2015 sypsky	17	79,6	63,1
56	2015 sypsky	18	68,2	63,1
57	2015 sypsky	19	90,1	51.18
58	2015 sypsky	21	66,5	49,66
59	2015 sypsky	22	65,7	49,66
60	2015 sypsky	23	73,2	49,66
61	2015 kevät	3	64,5	38,8
62	2015 kevät	4	95,3	38,8
63	2015 kevät	5	72,5	56,9
64	2015 kevät	7	62,5	42,17
65	2015 kevät	8	74,9	42,17
66	2015 kevät	9	61,8	42,17
67	2015 kevät	10	67,2	42,17
68	2015 kevät	12	70,4	59,03
69	2015 kevät	14	61	66,52
70	2015 kevät	16	69,8	45,63
71	2015 kevät	17	86,6	45,63
72	2015 kevät	18	70,4	47,83
73	2015 kevät	20	76,7	47,2
74	2015 kevät	21	55,4	64,94
75	2015 kevät	22	85,9	42,93
76	2015 kevät	23	88,6	42,93

9.3. Appendix C: lexical variables

Explanations of abbreviations in the table:

- t = tokens
- t.t.r. = type token ratio
- t.p.t. = tokens per type

item	K1-t	K-2-t	K3-t	t.t.r.	t.p.t.	lexical density
1	24	2	3	0,93	1,07	46,43
2	25	2	3	0,82	1,22	53,13
3	14	0	5	0,9	1,11	65
4	16	1	1	0,9	1,11	45
5	30	4	2	0,7	1,42	51,35
6	16	7	3	0,81	1,24	52,17
7	14	2	4	0,9	1,11	50
8	9	4	2	0,75	1,33	50
9	17	0	0	0,89	1,12	44,44
10	20	2	0	0,85	1,18	50
11	24	3	5	0,82	1,22	56,25
12	19	0	2	0,77	1,29	59,09
13	20	1	0	0,77	1,29	40
14	27	7	4	0,88	1,14	60,53
15	25	4	5	0,89	1,12	55,56
16	24	3	4	0,81	1,24	41,94
17	10	2	3	0,87	1,15	46,67
18	22	2	1	0,81	1,24	36
19	22	2	2	0,74	1,35	48,15
20	19	4	3	0,89	1,12	48,15
21	23	1	0	0,78	1,29	52,17
22	15	3	1	0,9	1,11	50
23	22	6	2	0,85	1,18	57,58
24	26	3	7	0,86	1,16	60
25	32	2	0	0,89	1,13	38,89
26	25	6	10	0,86	1,17	58,54
27	23	2	1	0,89	1,12	52
28	15	1	0	1	1	58,82
29	21	2	5	0,83	1,21	50
30	20	3	4	0,9	1,11	59,26
31	18	6	1	0,85	1,18	50
32	19	3	1	0,95	1,05	52
33	15	4	0	0,95	1,05	60
34	6	4	2	0,83	1,2	66,67
35	34	3	7	0,84	1,18	50
36	18	8	7	0,7	1,43	54,55
37	14	2	8	0,83	1,2	54,17
38	33	5	7	0,88	1,14	54,17
39	10	6	3	0,91	1,1	73,91
40	29	2	2	0,72	1,38	65
41	24	10	2	0,82	1,23	51,35
42	18	5	6	0,86	1,16	48,28
43	22	5	4	0,81	1,24	65,52
44	19	3	0	0,65	1,55	60,87
45	23	1	2	0,65	1,55	48,39
46	19	4	1	0,88	1,14	60
47	14	3	2	0,89	1,12	52,63
48	17	5	3	0,92	1,09	48

49	9	3	3	0,82	1,22	55,56
50	23	2	8	0,82	1,22	45,45
51	17	2	0	0,95	1,05	54,55
52	12	1	5	0,89	1,12	58,82
53	17	7	4	0,86	1,17	53,57
54	22	5	1	0,9	1,11	75
55	14	6	1	0,95	1,05	42,86
56	24	5	5	0,83	1,2	52,94
57	22	7	1	0,81	1,23	54,84
58	28	7	10	0,8	1,24	58,54
59	15	4	5	0,77	1,3	53,85
60	18	5	4	0,86	1,17	57,14
61	30	4	2	0,79	1,27	57,89
62	24	6	2	0,97	1,03	54,55
63	17	20	2	0,82	1,22	45,45
64	17	2	1	0,87	1,15	47,83
65	20	4	3	0,9	1,11	60
66	17	1	2	0,91	1,1	55
67	15	3	1	0,95	1,05	63,64
68	30	3	1	0,71	11,4	40
69	34	2	5	0,8	1,26	45,45
70	15	2	1	0,89	1,12	61,11
71	27	0	3	0,83	1,2	44,83
72	22	2	4	0,9	1,12	48,28
73	17	3	1	0,92	1,09	50
74	24	2	3	0,94	1,06	63,64
75	28	8	3	0,86	1,17	54,76
76	23	6	1	0,84	1,19	50

9.4. Appendix D: variables concerning item length and phrases

Explanations of abbreviations in the table:

it = item number in the present study

w = number of words

s = number of syllables

c = number of characters

sw = number of words in the stem

ss = number of syllables in the stem

sc = = number of characters in the stem

ow = number of words in the options

os = number of syllables in the options

oc = number of characters in the options

kw = number of words in the key (correct option)

ks = number of syllables in the key

kc = number of characters in the key

wfo = number of words in the distractors (the two false options per item)

sfo = number of syllables in the false options

sfo = number of characters in the false options

nc = number of clauses in the item

it	w	s	c	sw	ss	sc	ow	os	oc	kw	ks	kc	wfo	sfo	cfo	nc
1	29	46	160	8	11	44	21	35	116	6	12	34	15	23	82	5
2	33	53	223	6	13	37	28	42	171	8	13	50	19	29	120	4
3	20	31	125	8	12	47	12	32	109	3	9	32	9	23	76	5
4	20	31	127	20	31	111	12	20	73	5	6	25	7	14	47	4
5	37	59	234	11	14	55	26	45	157	8	16	54	17	32	110	4
6	23	44	169	6	10	38	17	34	115	5	10	33	12	24	81	4
7	20	37	125	6	9	31	14	28	97	5	9	29	9	19	63	4
8	16	27	110	4	10	30	12	17	63	4	5	19	8	12	43	4
9	18	24	102	8	10	35	14	19	70	5	7	27	9	12	42	4
10	26	42	171	8	15	54	18	27	99	6	9	31	12	18	67	4
11	33	49	185	9	11	47	24	38	137	8	14	45	15	24	91	4
12	22	35	126	10	14	50	12	21	75	4	7	23	8	14	51	4
13	20	25	97	5	7	29	15	18	70	5	6	22	10	12	47	4
14	38	70	285	14	22	84	38	70	258	7	15	49	17	33	124	4
15	36	58	233	11	18	63	25	40	156	9	14	54	16	26	97	4
16	31	46	186	10	14	53	21	32	120	7	12	38	14	20	77	4
17	15	28	111	4	11	38	11	17	63	4	6	22	7	11	40	4
18	26	35	156	13	19	73	14	16	64	4	6	20	8	11	43	4
19	27	38	163	13	18	73	15	18	73	5	6	25	10	12	45	4
20	28	45	187	9	16	57	19	29	114	6	9	33	13	20	76	4
21	27	36	172	9	12	48	18	24	110	6	7	31	9	17	69	4
22	20	30	123	8	12	45	12	18	65	4	5	16	8	13	44	4
23	33	58	249	10	16	64	23	42	162	10	14	50	13	28	112	4
24	41	73	263	16	25	89	25	49	176	9	15	56	16	34	117	4
25	35	49	206	8	13	51	27	36	154	10	12	52	17	24	101	7
26	42	80	307	6	9	36	36	71	237	13	25	75	23	46	161	6
27	28	44	167	7	11	39	21	33	114	9	14	45	12	19	68	4
28	17	25	117	11	16	64	6	9	35	2	3	10	4	6	24	4
29	29	43	175	29	43	161	19	30	102	7	13	37	11	17	66	4
30	28	46	188	8	14	54	20	32	115	6	10	34	13	22	79	4
31	26	46	170	5	8	29	21	38	126	7	12	39	15	26	87	4
32	26	42	170	8	12	52	17	29	100	5	8	29	11	21	70	4
33	20	34	148	6	8	39	14	26	94	3	8	25	11	18	68	4
34	18	36	142	6	9	29	12	17	87	4	9	29	8	18	57	4
35	43	69	280	13	22	73	30	47	182	9	15	55	21	32	124	4
36	33	60	235	9	18	63	24	42	155	8	11	45	16	31	109	5
37	24	37	162	8	12	47	15	25	97	4	8	31	11	17	65	4
38	48	81	325	11	15	56	37	66	236	12	21	78	25	45	157	7
39	23	42	166	10	14	57	13	28	94	4	9	30	9	19	63	4
40	40	74	282	9	20	67	31	54	187	10	16	58	21	38	128	4

41	38	75	291	11	22	73	27	53	183	10	16	58	17	36	116	4
42	29	44	192	10	14	52	18	29	121	6	9	38	12	20	82	4
43	31	58	227	11	17	67	20	41	138	7	12	43	12	29	94	4
44	23	36	158	11	20	69	12	16	69	3	5	17	9	11	51	4
45	31	47	178	7	11	44	24	36	117	8	11	35	16	25	81	4
46	25	34	148	11	15	58	14	19	75	5	6	24	9	13	50	4
47	19	30	131	7	11	47	12	19	69	4	5	18	8	14	50	4
48	25	43	107	9	14	51	16	29	100	6	9	32	10	20	67	4
49	18	36	131	9	14	50	9	22	64	3	7	21	6	15	42	4
50	33	50	202	13	17	65	20	33	116	8	12	39	12	21	76	4
51	22	39	149	7	11	45	15	28	89	5	8	27	10	20	61	4
52	18	32	132	9	16	57	9	16	60	3	5	18	6	11	41	4
53	18	51	201	11	21	69	17	30	105	6	10	35	10	20	69	4
54	29	56	216	9	15	54	20	41	145	8	13	47	11	28	97	4
55	21	31	127	8	13	48	13	18	64	4	7	22	11	11	41	4
56	36	71	265	12	20	73	24	51	164	8	13	50	9	38	113	4
57	32	54	212	8	11	46	24	43	149	8	16	57	15	27	91	4
58	41	76	303	11	16	64	30	60	210	8	19	58	15	31	115	4
59	26	45	174	8	14	49	18	31	109	5	10	33	22	21	75	4
60	28	50	192	9	14	51	19	36	124	7	11	40	13	25	83	4
61	38	63	243	9	20	65	29	43	156	6	11	38	12	32	117	4
62	33	61	232	9	13	57	24	48	157	7	17	54	8	31	102	5
63	22	33	144	9	13	45	13	20	83	3	4	19	12	16	63	5
64	23	35	143	12	19	70	12	19	70	4	5	24	21	14	45	4
65	30	61	214	13	23	81	17	38	116	6	12	38	17	26	77	4
66	22	37	134	8	14	45	14	23	74	5	7	25	10	16	48	4
67	22	46	164	10	18	60	12	28	89	4	11	33	8	17	55	4
68	35	54	220	14	21	81	21	33	118	8	12	43	11	21	74	4
69	44	69	298	13	16	65	30	50	185	8	14	51	7	36	133	4
70	18	29	129	7	11	42	11	18	69	4	7	23	8	11	41	5
71	29	39	179	11	16	65	18	23	93	8	11	39	13	12	53	4
72	29	46	182	8	9	39	21	37	121	6	11	35	22	26	85	4
73	22	41	155	10	17	60	12	24	78	4	9	25	7	15	52	4
74	33	52	215	12	17	65	21	35	126	6	13	42	10	22	83	4
75	42	68	271	16	26	92	26	42	153	8	13	48	15	29	104	4
76	30	46	186	10	14	51	19	31	113	7	9	34	8	22	78	4

English abstract

The aim of this study was to find out about the relationship between complex reading item formulations and the facility value of these items. Firstly, 76 similar reading items with regard to the construct and difficulty were selected from the *Finnish Matriculation Examination*. After having identified comparable items, the focus could be put on the relationship between difficult item formulations and the facility value. In total, 22 item features concerning the item formulation were defined. These referred to lexical aspects as well as to the length of the items and the number of clauses. Each of these features was correlated with the facility values of the items. Only a few features showed a weak negative correlation. Two of these were the ones dealing with difficult lexis. Hence, based on the present study, it seems important that item writers are aware of the relationship between difficult words in items and the facility value. The fact that there were only three item features out of 22 with a weak significant correlation suggests that features involving the text are more significant for the item facility value. Certain features could not be tested within this research project (e.g. even less frequently used words in items) and another much larger study came partly to different results, which is why further research in this field might be necessary.

Deutsche Zusammenfassung der Diplomarbeit / Deutsches abstract

Das Ziel der vorliegenden Studie war es, die Beziehung zwischen komplexen Formulierung von Lese-Items und deren Lösungshäufigkeit zu erforschen. 76 vergleichbare Leseitems hinsichtlich Konstrukt und Schwierigkeitsgrad der finnischen Maturaprüfung wurden dafür ausgewählt. Nachdem somit besonders ähnliche Items herausgefiltert waren, blieb nur noch der Aspekt der komplexen Item-Formulierung als relevanter Faktor hinsichtlich der Lösungshäufigkeit bestehen. 22 Variablen hinsichtlich der Item-Formulierung wurden definiert, welche sich sowohl auf die Lexik als auch auf die Item-Länge und deren Satzteile bezogen. Jede dieser Variablen wurde mit der Lösungshäufigkeit der Items korreliert. Nur wenige dieser Variablen zeigten eine leichte negative Korrelation auf. Zwei signifikante Variablen hinsichtlich komplexer Lexik in den Items unterstützten einander in ihrer Aussagekraft. Folglich erscheint es wichtig, dass sich Testentwickelnde der Beziehung zwischen schwierigen Item-Formulierungen und der Lösungshäufigkeit bei der Aufgabenentwicklung bewusst sind. Da nur drei der 22 Variablen eine leichte negative Korrelation aufzeigten, ist die Hypothese aufzustellen, dass Variablen, welche auch Bezug auf die

relevanten Textstellen nehmen, deutlich signifikanter für die Lösungshäufigkeit sind, wie eine ähnliche deutlich umfangreichere Studie herausgefunden hatte. Weitere relevante Item-Variablen konnten in diesem Forschungsprojekt nicht erfasst werden (z.B. noch selten gebräuchlichere Lexik in Items) und eine vergleichbare Studie kam zu teils anderen Erkenntnissen, weswegen weitere Forschungen in diesem Feld sinnvoll erscheinen.