universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

"Multiple choice, true-false-not given or matching reading comprehension tests in EFL – What is the difference?"

Verfasserin

Susanne Elisabeth Hinterlehner

angestrebter akademischer Grad

Magistra der Philosophie (Mag.phil.)

Wien, Mai 2010

Studienkennzahl lt. Studienblatt:	A 190 344 353
Studienrichtung lt. Studienblatt:	Lehramtsstudium UF Englisch
Betreuerin:	Ao. Univ.-Prof. Mag. Dr. Ute Smit

To
my parents,
with love and gratitude

Acknowledgements

Writing a thesis is a challenging project in a student's life, which cannot be successfully completed without the help, support and encouragement of a number of wonderful people. Therefore, I would like to offer a few words of thanks to those who have been especially helpful to me in writing my first big paper.

First of all, I would like to thank Professor Tim McNamara for introducing me to the fascinating field of language testing. I am especially grateful that he enabled my research stay at the University of Melbourne and that he supported me in every conceivable way.

With all of my heart I would like to thank my supervisor Professor Ute Smit, who advised, supported, and inspired me in an exemplary way. Whenever I thought I had hit a dead end she always helped me sort out my ideas in a clear-sighted way. Throughout the writing process of this thesis I learned to appreciate her not only as an outstanding researcher and professor, but also as a guide and friend who listened to my problems patiently and encouraged me even out of office hours.

I am also indebted to Professor Christiane Dalton-Puffer, who first sparked the idea to conduct a replication study in me. Further, I am very thankful that she offered me to become a member of the language testing project team, where I have wonderful possibilities to plunge even deeper into the field of language testing.

Moreover, I wish to record my thanks to the 'DLE Forschungsservice and Internationale Beziehungen', Land Oberösterreich, as well as the Julius Raab Stiftung, which supported my research at the renowned University of Melbourne financially.

Thanks are also due to the head of the Baillieu Library, Sabina Robertson, who granted me privileged access to the libraries at the University of Melbourne. Further, I would like to thank the Australian librarians Vincent, Peter, and Ken who spent hours in the stacks to find the rare books and articles I desperately needed for the completion of my thesis.

Special thanks also go to the two Austrian “Gymnasien” where my reading tests were conducted as well as to the Landesschulrat für Wien, who allowed me to do research at these schools. In particular, I would like to offer my thanks to the headmaster and headmistress, to the English teachers who did everything to support my research, and last but not least to the test takers for their time and understanding.

Further, I wholeheartedly would like to thank Mag. Lucia Ashry and my dear American friend Keara for spending hours proofreading my paper and for offering very insightful comments regarding content. When it came to the statistical analysis of the test, my friend DI Bettina Länger and Professor Ute Knoch at the University of Melbourne were always more than willing to help.

Of course, I must not forget my dear friends, who were always there for me and who successfully managed to talk me into having fun and relaxing breaks. Above all I would like to thank Julia and Sara without whom my great time at university would not have been the same. Further, I am particularly indebted to Silvia not only for her friendship but also for engaging in conversations about language testing with me until late.

Posthumously, I would like to thank Kathi Kefer for sharing her fascination of foreign languages, countries, and cultures with me. R.I.P.

Above all, I owe a special debt of gratitude to my wonderful parents, who supported me in every conceivable way and who never stopped believing in me. I am eternally grateful that they shared their values, knowledge, and experience with me throughout my education. No words in the world can describe how much they mean to me and how much they have done for me.

Last, but certainly not least, I would like to express my sincere gratitude to Gert for his love, emotional support, humor, and understanding. He managed to continuously encourage me and was proud even of the smallest steps I took in the completion of this thesis. Thus, he succeeded in making me feel happy even at very stressful and challenging times. What he has done for me can never be adequately thanked for.

Declaration of Authenticity

I confirm to have conceived and written this paper in English all by myself. Quotations from other authors and any ideas borrowed and/or passages paraphrased from the works of other authors are all clearly marked within the text and acknowledged in the bibliographical references.

Vienna, May 2010 _____

Zusammenfassung

Obwohl es sich bei dem Bereich der Sprachtestforschung um einen ständig wachsenden Wissenschaftszweig handelt, gibt es bis jetzt noch wenige Forschungsarbeiten, welche den Einfluss von geschlossenen Testformaten untersuchen. Die einzigen Studien, die den Einfluss von unterschiedlichen Frageformaten auf die Lesefähigkeit der Schüler erforschen, konzentrieren sich auf einen Vergleich von geschlossenen und offenen Frageformaten. Als Beispiele für geschlossene Frageformate können Mehrfachwahlaufgaben, das Richtig/falsch/nicht im Text – Format sowie Zuordnungsaufgaben angesehen werden. Kurzantworten wiederum sind ein Beispiel für offene Frageformate. Die vorliegende Diplomarbeit stellt daher eine Neuheit dar, da sie sich zum Ziel gesetzt hat, den Einfluss von drei geschlossenen Testformaten auf die Lesefähigkeit der Schüler zu überprüfen.

In der vorliegenden Arbeit werden folgende Frageformate näher analysiert: Mehrfachwahlaufgaben, Richtig/falsch/ nicht im Text -Aufgaben, sowie Zuordnungsaufgaben. Diese Studie untersucht die Lesefähigkeit von 97 Schülern der 7. Klasse Oberstufe, welche zwei unterschiedliche Wiener Gymnasien besuchen. Insgesamt absolvieren 48 männliche und 49 weibliche Teilnehmer den Lesetest, welcher auf einem Zeitungsartikel über Studiengebühren aus einer britischen Qualitätszeitung basiert.

Der Lesetest wurde anhand von Testqualitätsmerkmalen untersucht. Um die Unterschiede zwischen den Resultaten der männlichen und weiblichen Testpersonen herausfinden zu können wurden Hypothesentests wie der Einstichproben- sowie der Zweistichproben- t-Test durchgeführt.

Die aus der Studie resultierenden Ergebnisse weisen auf einen signifikanten Effekt der Testmethode bezüglich der Lesefähigkeit der Schüler hin. Einige der drei geschlossenen Frageformate fallen den Testteilnehmern leichter, andere wiederum führen zu Schwierigkeiten. Manche Testformate weisen einen stärkeren Einfluss auf schwächere Schüler und weibliche Testpersonen auf. Im Diskussionsteil der Diplomarbeit werden die Ergebnisse diskutiert und Rückschlüsse auf den theoretischen Hintergrund, welcher der Forschung zu Grunde liegt, geschlossen.

Abstract

Despite the growing body of research in the area of language testing, little is known about how different assessment formats affect students' reading comprehension performance. While a handful of studies investigated and compared students' performance on selected-response and constructed-response items, no study to date has compared two or more selected-response formats. Thus, the present thesis attempts to fill this niche by investigating the influence of three different selected-response formats on students' reading comprehension performance.

The three selected-response formats under investigation are multiple choice, true-false-not given and matching. 97 grade 11 students of two Viennese *Gymnasien,* i.e. Austrian type of secondary school that prepares students for higher education, participated in the study. In total 48 male and 49 female testees took the reading comprehension test based on an authentic newspaper article on tuition fees. The order of the test formats remained constant: multiple choice, true-false-not given, and matching.

The reading comprehension test was analyzed according to the underlying criteria of test quality. To investigate the differences between students' performance on the three selected-response formats, difference inferential statistics with the help of independent-sample and paired-sample t-tests were carried out.

Results of this study suggest that the assessment method has a significant effect on students' reading comprehension performance. Out of the three selected-response methods some turn out to be more difficult than others and some have a greater effect on students of low-level proficiency and on female test takers. Results will be discussed and conclusions based on assumptions from the theoretical background of the study will be drawn.

Language testing is central to language teaching.

(Davies 1990: 1)

Table of contents

List of tables

List of figures

1. Introduction

1.1. Context of the problem

Reading is a very important skill not only in first but also in foreign language situations (cf. Shehadeh 1998: 1). The skill of reading is one of the four skills students most likely maintain even after having finished formal foreign language education (ibid). Rivers (1981 as referred to by Shehadeh 1998: 2) asserts that reading ability is in fact the "most retained and durable among second language skills". Given these reasons it is important for teachers and people involved in language testing to assess students' reading comprehension abilities. Here the question arises how the skill of reading is best assessed. Although there is "no one best method" (Alderson 2000: 203) some methods seem to be more adequate than others.

Along with listening reading is regarded as a receptive skill, which per se does not require a testee's productive skills. Accordingly, it is advisable to assess receptive skills by methods which do not require production on part of the students, as this could otherwise contaminate their reading ability leading to unreliable assessment (Alderson 2000: 30). This view is supported by Brown and Hudson (2002: 59) who claim that selected-response formats are "most appropriate" for testing the receptive skill of reading.

According to Popham (2002: 145) multiple choice, true-false-not given and matching tasks are "the most common kinds of selected-response test items". Here the question arises whether students' reading comprehension performance on these three selected-response formats differs. Are students equally familiar with these fixed-response formats? Do testees perceive multiple choice, true-false-not given and matching tasks to be of similar difficulty? Are students' scores on these three fixed-response formats related as they are assessing the same underlying ability, namely reading comprehension? Is there a gender bias in as much as one gender will be favored by one of the fixed-response formats? Researchers as Alderson (2000: 123 f.), Alderson, Clapham and Wall (1995: 44) acknowledge the importance of research on the comparison of different test methods, as the so-called test method effect, that is

that a certain assessment method affects a subject's scores, should be avoided as far as possible.

Until today only a handful of studies have investigated the effects of test format on reading comprehension. What all of the existing research studies have in common is that they compare the effects of selected-response formats, i.e. multiple choice items, and constructed-response formats, as for instance short answer questions, on subjects' reading comprehension performance. The present study attempts to fill a niche by comparing the following three selected-response formats: multiple choice, true-false-not given, and matching.

1.2. Purpose of the study

The purpose of the present study is to investigate the influence of three different selected-response assessment methods, i.e. multiple choice, true-false-not given, and matching, on the testees' reading comprehension performance. These assessment methods are presented in the pupils' target language English. The reading comprehension text is a newspaper article taken from the British quality newspaper, *The Guardian*, on tuition fees (cf. appendix 1).

1.3. Research questions

The main research question guiding this paper is whether any differences are to be found between three different selected-response formats assessing reading comprehension: multiple-choice, true-false-not given, and matching. On which format do the students perform best and worst? If a difference is to be found – does the test format influence test takers across proficiency levels equally or are less proficient students more affected by certain test formats than highly proficient students? Given that testees perform differently on the three selected-response assessment tasks, what are the possible reasons for these differences? How do test participants rate the three selected-response formats on a 5-point Likert-scale ranging from 1 (very easy) to 5 (very difficult)?

Additionally, the following research questions will be addressed:

1. Does male and female test takers' performance on multiple choice, true-false-not given, and matching reading comprehension test items differ significantly?

2. What role does interest play? Do students achieve better result when the reading text is more interesting?
3. Does school A and school B pupils' reading comprehension performance on multiple choice, true-false-not given, and matching formats differ? Are the differences between school A and school B reflected in the respective students' performance on the three selected-response formats?
4. Research question four, however, is only concerned with one assessment format: multiple choice. In the case of multiple choice items where more than one answer is correct, does an indication of the number of correct answers next to each item improve the testees' performance?

1.4. Assumptions

The present study is based on the following assumptions:

1. Multiple choice, true-false-not given, and matching formats are adequate measures of reading comprehension.
2. The students will perform the best they can on all reading comprehension assessment types: multiple choice, true-false-not given, and matching.
3. Multiple choice, true-false-not given, and matching tasks are familiar to all test participants. Nevertheless, testees might not be equally familiar with all three test formats. Possibly students are more familiar with multiple-choice and true-false-not given tasks than with matching tasks.
4. The students' scores on the three selected-response test formats: multiple choice, true-false-not given, and matching are related, as these formats assess the same underlying ability: reading comprehension. This assumption will be tested in section 4.2.
5. The newspaper article on tuition fees is unknown to the subjects.
6. The testees cooperate with the researcher. Thus, they refrain from cheating. Otherwise the test administrator prevents cheating.
7. The subjects' are non-native speakers of English. Most students' native language is German. All subjects learn English as a foreign language.

8. The test participants' mean score on the three different test formats is an accurate measure of their reading comprehension ability.
9. The scoring procedures are valid and reliable.
10. The reading comprehension test is valid and reliable. This assumption will be tested in sections 4.1.2. and 4.1.3.

1.5. Limitations of the study

The following limitations apply to the present study:

1. The multiple choice, true-false-not given, and matching tasks of the present reading comprehension test are not to a 100% comparable as they center on different parts of the newspaper article. Furthermore, the three test formats are not stem equivalent. Thus, the multiple choice items do not correspond to the respective true-false-not given, and matching items. Accordingly, the emerging differences between the three selected-response formats cannot be attributed to the test formats alone, but to the fact that some items are assessing easier and others more difficult passages within the newspaper article.
2. The results of the present study cannot be generalized to other studies with similar populations, as the statistic procedures used are sample-based statistics which are only true for the present sample of subjects.
3. In this study only one reading passage was used. Studies using different passages and text types (Shohamy 1984) obtained different results. Therefore, the results and conclusions of the present study can only be generalized to other similar passage types and topics, given that the subjects are similar to the group of testees in the present study.
4. Due to the limited scope of this paper it is impossible to explain all relevant statistical procedures in a detailed way. For a closer discussion of the statistics used see Morgan (2004).

1.6. Overview

The present thesis consists of four main parts. The first part attempts to give an overview of the existing literature on language testing, the nature of reading, as well as of the different reading comprehension test formats. Moreover, research

centering on factors as well as reader attributes, such as gender and interest, which might be influencing testees' reading comprehension performance will be presented. The second part is dedicated to an explanation of the methods used for the investigation of the research questions guiding this paper. It is up to the third part of this thesis to present the results of the research questions, which will then be discussed in the fourth part of this thesis along with insights gained from the theoretical background.

2. Review of the Literature

The purpose of this chapter is to review the existing literature related to the subject matter. It is subdivided into two main sections: language testing and testing reading comprehension, which are divided into further subsections. While the first section briefly explains the roots of language testing, purposes of language tests as well as different types of tests, the other main section centers on reading comprehension. More specifically, the second part focuses on the nature of reading and presents an overview of different second language (L2) reading comprehension selected-response test formats along with advantages and shortcomings of each type. The last part of the literature review discusses empirical studies investigating factors and reader attributes that might affect reading comprehension performance.

2.1. Language testing

Language testing is a form of measurement (Henning 1987: 1) which, as the name implies, intends to measure a language learner's knowledge of a language and ability to use it. Interestingly, many of the vital books on language testing (e.g. Davies 1990. Kitao 1999, McNamara 2000) as well as the dictionary of language testing do not offer a definition of a term so "central to language teaching" (Davies 1990: 1). This shortage of definitions is also acknowledged by Glenn Fulcher and Alan Davies, who held a competition for a definition of "language testing" suitable for a dictionary entry in winter 2009 (http://209.85.229.132/search?q=cache:I68_T92HnxAJ:languagetesting.info/whatis/lt.html+what+is+language+testing&cd=5&hl=de&ct=clnk&gl=at&lr=lang_de&client=firefox-a, 5 February 2010).

Language tests, however, are defined by Davies et al. (1999: 107) in the dictionary of language testing as being comprised by a number of "specified tasks through which language abilities are elicited". Accordingly, language testing could be defined as the act of setting and assessing specified tasks through which language abilities are elicited.

2.1.1. The roots of language testing

The roots of language testing go back to educational testing, which itself originates in measurement theory and practice in psychology and psychometrics (Davies 1990:10). Thus, it becomes apparent that the notion of language testing does not only consist of two terms: language and testing but it also encompasses two aspects, that of psychometric testing and of languages as such. Davies (1990: 10) argues that language tests can only be successful if they manage to reconcile those twin requirements.

2.1.2. The purpose of language tests

Here the question arises as to why language tests are conducted. One of the most common uses of language tests in the school context is probably to point out strengths and weaknesses in the learned abilities of a student (Henning 1987: 1). Other than that, tests are commonly used to select students for admission to universities and to place students into language programs. Additionally, language tests are used for the purpose of program evaluation (Henning 1987: 2), for evaluation as such and research (Davies 1990: 9). In the case of the present study the language test was used for research, although it had been originally developed for the purpose of a language program evaluation, i.e. to evaluate the English and French tracks of a Viennese Gymnasium. In recent decades language tests have also been increasingly used to select potential immigrants (McNamara 2009: lecture notes).

One of the most common distinctions of language tests based on test purpose is the one between achievement and proficiency tests. Researchers as Davies (1990), Henning (1987), Hughes (2003) and Kitao (1999), however, further mention placement, aptitude and diagnostic tests, which McNamara (2000: 6 f.) considers as subtypes of achievement and proficiency tests. Achievement tests focus on the past learning experience by measuring what students have learned as a result of teaching. Thus, the most typical examples of achievement tests are school tests either at the end of a period of learning or even at the end of a school career. The content of an achievement test is what has been covered during the period of instruction, i.e. what the curriculum suggests. What Henning (1987) and Kitao (1999) term diagnostic tests can be

regarded as an achievement test with the purpose of “isolating learning deficiencies” (Davies 1990: 6). In contrast to achievement tests, proficiency tests do not refer to the past but to the future situation of language use (McNamara 2000: 7). Thus, proficiency tests are often “global measures of an ability [or aptitude] in a language” (Henning 1987: 6). Therefore, they are commonly used for placement or selection purposes (ibid) as well as to assess a student’s aptitude.

The underlying test which was used for a research purpose in my study and a program evaluation purpose in the original study is a proficiency test, as it tests whether students from two different language tracks reached a certain level of language proficiency without any reference to prior instruction. These two language tracks, which are referred to as English and French track further on in this thesis differ with regard to the amount and duration of English and French classes. While the English track starts with English, pupils attending the French track begin in the first grade with French as the first foreign language. Having established the most common uses of language tests, the question arises as to what types of tests can be distinguished.

2.1.3. Types of tests

Language tests differ according to their type and their purpose. Test of the same type might be used for different purposes although the test type in some instances might be only applicable in a limited way (Henning 1987: 4). For instance, a multiple choice test may be used in a program evaluation as well as in an admission test. A portfolio assessment, however, might be used for an achievement test but not for a placement test. In the following part of the literature review the most important types of tests, which are relevant to the present reading test, will be briefly exemplified.

Tests can also be subdivided according to the standards used in grading (Kitao 1999: 6). Tests which compare a testee’s performance to that of other testees, i.e. the norm, are referred to as norm-referenced tests. Criterion-referenced tests, however, judge a test taker’s performance against a standard, “a particular defined description of the language proficiency expected” (Kitao 1999: 6). The Common European Framework of References (CEFR) can be set as a standard. The present reading test has been designed as a criterion-

referenced test, meeting the standards of the CEFR. For a more detailed discussion of the CEFR in relation to the underlying test as a whole, and the reading test in particular see Derntl (2009) and Schweinberger (2009).

Furthermore, tests can be distinguished according to the manner they are scored (Henning 1987: 4). The distinguishing feature of objective tests is a preset sample of correct answers, referred to as the scoring key. Thus, objective test can be scored by raters who do not need to have any expertise in the respective field. Multiple choice items are a common example of objective tests. Conversely, subjective tests, as for instance free compositions, do require "opinionated judgment based on insight and expertise on the part of the scorer" (Henning 1987: 4). There is, however, a possibility to "objectify" free compositions, i.e. subjective tests. This can be done with the help of precise rating scales (ibid).

A further differentiation in language testing is made between speed and power tests. While power tests allow the test taker enough time to finish the generally rather difficult items, speed tests focus on the test taker's "speed of performance rather than on knowledge alone" (Henning 1987: 8). Kitao (1999: 9) asserts that speed tests are particularly suitable for testing fluency, while power tests assess accuracy. The present reading test serves as an example of a power test.

The distinction between discrete point and integrative tests goes back to John B. Carroll's influential paper "Fundamental considerations in testing for English language proficiency of foreign students", published in 1961 (Henning 1987: 5). While discrete point tests examine only one particular skill or piece of knowledge, integrative tests focus at least on two skills at once (ibid). Carroll was an opponent of discrete point tests, as he argued that integrative tests provide a more natural and communicative representation of the testee's knowledge of the language (Carroll 1961: 36 f.).

At still lower levels tests can be classified according to the skills tested, whether a test assesses reading comprehension, listening comprehension, speaking or writing skills (see next section). A further subdivision of tests can be made according to the test format, whether a test features selected or constructed response formats. A thorough discussion of the relevant test formats of the underlying reading test, is to be found in chapter 2.2.2.

2.1.4. The four skills

Language tests are designed in myriads of different ways and they are used for a variety of purposes. To be better able to assess a learner's knowledge of language and ability to use it, language testers and language teachers have traditionally considered language ability as consisting of four skills: listening, reading, speaking and writing (Bachman 1996: 75). The origins of these four skills go back to highly influential models of language ability that have been established in 1961 by Robert Lado and J.B. Carroll (Bachman 1996: 75, Carroll 1961, Lado 1961). Standardized language tests as IELTS, Cambridge Certificates, as well as the Austrian four skills *Matura*, i.e. school leaving exam at upper secondary level, the precursor of the Austrian standardized *Matura*, assess the four skills, which are divided into receptive skills, i.e. listening and reading, and productive skills, i.e. speaking and writing, separately. TOEFL exams test the four skills of the English language in an integrated way as well as separately". (http://www.testpreppractice.net/TOEFL/toefl-testing-1.aspx, http://www.pearsonlongman.de/main/main.asp?page=exams/bookdetails&ProductID=131710, 9 February 2010). Similarly, the original study, with its reading test being analyzed in great detail in this thesis, assesses the listening, reading and speaking part separately, while the writing test assesses an integration of reading and writing skills, i.e. the task demands the students to write an argumentative essay on the topics of the two reading tasks, where arguments featured in the reading test should be used (cf. Schweinberger 2009: 92).

Although a division of language ability into the four skills might sound reasonable, Lyle Bachman (1996: 75) opposes this division as it classifies "widely divergent language use tasks and abilities together under a single 'skill'". Thus, Bachman (1996: 76) argues that instead language abilities should be subdivided into "specific activities or tasks in which language is used purposefully". Although Kitao (1999: 88 f.) argues that "in real life the different skills are not often used entirely in isolation", which accordingly makes four skills tests unauthentic, she highlights that an authentic testing situation which focuses more on authenticity than on a separation of the four skills leads to "an almost inevitable loss of reliability". Alderson (2000: 30) acknowledges the same problem. He states that integrated tests, i.e. tests integrating at least two

different skills as for instance reading and writing, contaminate a student's reading ability thus leading to unreliable assessment. Weir (1990), as referred to by Alderson (2000: 30), denotes this danger 'muddied measurement'. The present reading test only assesses the skill of reading with the help of selected-response tasks which do not require production skills on part of the students. By using selected-response tasks instead of constructed-response tasks, the present test prevents a mingling of reading and writing skills which could otherwise have led to "muddied measurement" (ibid).

2.1.5. The test criterion relationship

To gain an understanding of the nature of language testing it is essential to recognize that "testing is about making inferences" (McNamara 2000: 7). In language testing an important distinction has to be made between the criterion, which is referred to as "relevant communicative behavior in the target situation" (McNamara 2000: 8) and the actual test. While the test performance can be observed, the criterion is unobservable (ibid). Test performances can never account for the criterion, as they can only be a simulation of the latter. To give an example, on the basis of a listening comprehension test, i.e. students have to listen to a lecture, interferences are made on how a testee would cope with listening to lectures in the subject he is aiming to study (McNamara 2000: 8). Unjustified or wrong inferences based on a candidate's behavior do not only have serious consequences but they also constrain the validity of the test.

2.1.6. Validity and reliability

It is generally accepted that "testing is a universal feature of social life" (McNamara 2000: 3). Important decisions are drawn from test results, which can play a powerful role in people's lives as for instance in language tests for immigrants or admission tests for university students. Therefore it is vital that tests in general, and language tests in particular, fulfill the most important criteria of tests: validity and reliability. The term validity refers to the extent to which a test actually tests what it is intended to test, while reliability can be defined as consistency of measurement (Davies, 1990; Henning, 1987;

Hughes, 2003; Kitao, 1999; McNamara, 2000) . A closer examination of the "two basic factors" (Kitao 1999: 11) of tests is to be found in chapter four.

2.2. Testing reading comprehension

While the first part of this chapter aims at giving an overview of the nature of reading, the second part introduces the relevant reading test formats. The third and last part of this chapter focuses on factors and reader attributes that might be affecting reading comprehension performance.

2.2.1. The nature of reading

The present chapter is devoted to an investigation of what it means to be able to read. Further, models of reading comprehension will be explained. Additionally, the question whether there is a universal kind of reading or whether reading can be divided into different types will be explored.

2.2.1.1. What does it mean to be able to read?

> Our view of reading has a crucial influence on how we might go about testing and assessing reading (Alderson 2000: 28).

While this quote might sound logical, researchers like Alderson and Urquhart (1984), Alderson (2000), and Carr (2003) acknowledge that defining such a complex skill as reading is indeed a challenging, enormous task which risks "a dangerous level of simplification" (Alderson and Urquhart 1984: xv, xvi).

Urquhart and Weir (2002: 22) define reading as "a process of receiving and interpreting information encoded in language via the medium of print". This definition entails that the essence of the act of reading is constituted by two components: decoding, i.e. word recognition, and comprehension. Kintsch and Yarbrough (1982 as referred to by Alderson 2000: 9) agree on these two components, although they regard both as comprehension at different levels. Decoding is seen as the comprehension of words, i.e. a lower level of comprehension, while by comprehension they refer to the comprehension of the meaning of a sentence or the text as a whole, thus implying a higher level of comprehension.

Tricia Hedge (2000: 189) not only mentions the central components of reading, i.e. decoding and comprehension, but further divides these components into subcomponents. Accordingly, reading comprehension is seen to comprise at least six components: syntactic knowledge, morphological knowledge, general world knowledge, sociocultural knowledge, topic knowledge, and genre knowledge. While the first two components are concerned with the decoding process, the four latter components make up the comprehension process.

2.2.1.2. Models of reading

Generally speaking, among the most common approaches taken by readers in reading a test, also referred to as models of reading, are bottom-up and top-down approaches, schema-theoretic models as well as interactive models.

Bottom-up approaches are often referred to as serial models, as they take place in a predetermined order. They start with the smallest text unit (Urquhart, Weir: 42): the reader recognizes the graphic icons of the printed words, "decodes them to sound, recognizes words and decodes meanings" (Alderson 2000: 16). This model is typically associated with the so-called 'phonics' approach to the teaching of reading where readers are regarded as "passive decoders of sequential graphic-phonemic-syntactic-semantic systems" (Alderson 2000: 17). The top-down approach, however, does not regard readers as passive decoders but rather highlights the importance of the reader's contribution to the reading process, thereby almost disregarding the importance of the printed words (Alderson 2000: 16).

Schema-theoretic models, a subtype of the top-down approach, emphasize the centrality of the knowledge a reader brings to the text (ibid). In this approach readers interpret a text with the help of schemata, "networks of information stored in the brain which act as filters for incoming information" (Alderson 2000: 17). Reading can only be successful to the extent that the reader's schemata are relevant (ibid). This model, however, does not explain how prior knowledge is called up from memory and used for understanding, which has lead to criticism among researchers as Alderson (2000: 18).

Having explained these models of reading the question arises as to which approach most readers adopt when reading. Schank (1978, as referred

to by Alderson 2000: 17) asserts that in natural language understanding readers tend to adopt top-down processes as expectation plays a central role in understanding. Only when these expectations turn out to be useless readers get back to bottom-up processes.

Alderson (2000: 18), however, claims that neither bottom-up nor top-down processes are able to depict the reading process adequately, thus he suggests that a better model of the reading process might be the interactive model. The advantage of such a model, as compared to the bottom-up and top-down approach, is that the reading process does not follow any predetermined order but that every component in the reading process can interact with any other component regardless of its level in the processing chain. Thus, processing takes place in a parallel rather than in a serial way (Grabe 1991, as referred to by Alderson 2000: 18).

Thus, the interactive model incorporates top-down and bottom-up processes which interact in complex and until today poorly understood ways (Alderson 2000: 20). The balance between these two approaches is likely to vary with text, reader, and purpose. The easier the text for the reader, the more top-down processes will take place. In more difficult texts, however, readers will get back to bottom-up processes, in order to get to the meaning of single words they do not understand.

2.2.1.3. Different kinds of reading

Having covered the most common models of reading, the question is brought up to whether there is one universal kind of reading or whether there exist different kinds of reading. Although the research literature is primarily concerned with one type of reading, that is careful reading, there are more types of reading (Hughes 2003: 138, Urquhart and Weir 2002: 101,102).

Urquhart and Weir (2002: ibid) assert that the most common kinds of reading are careful, extensive and slow reading, which they subdivide into careful reading at local level, and careful reading at global level, and expeditious or quick and efficient reading, which is subdivided into search reading, skimming, scanning and browsing. Hughes (2003: 138), however, only mentions skimming, search reading and scanning but not browsing among the expeditious reading operations. Urquhart and Weir (2002: 108) finally admit

that browsing is the least well defined of the reading types, which makes it difficult to operationalize for teaching and testing purposes.

Skimming can be exemplified as selective reading for gist. The reader wants to get an idea of what the whole text is about without focusing on any details (Urquhart, Weir 2002: 102). Scanning can be described as an even more selective reading operation as skimming, as the reader does not want to get an overall idea of the text but just wants to spot specific words or phrases (Hughes 2003: 138,139; Urquhart, Weir 2002: 103). Finding a number in a phone directory could be regarded as an example of scanning (Urquhart, Weir 2002: 103). A further type of expeditious reading is search reading. In this kind of reading the reader's task is to locate information on predetermined topics, i.e. the reader has to answer a set of questions in a reading text (ibid).

In contrast to these expeditious reading operations, careful reading operations are not selective at all. While engaging in careful reading, the reader tries to read and understand the majority of information in the text (Urquhart and Weir 2002: 159). In careful reading readers should not only be able to understand the meaning of individual words and sentences but they should also be able to understand the text as a whole. Furthermore, readers should be able to include what they understood of the text as well as their background knowledge, from outside the text, to make meaning of a text. The interactive model probably best accounts for the approach taken by readers when engaging in careful reading. For a discussion of subskills involved in careful and expeditious reading operations see Hughes (2003: 139).

A test of reading comprehension should require careful as well as expeditious reading on part of the readers. Hughes (2003: 138) maintains that unfortunately expeditious reading skills are often disregarded in teaching and testing reading, which consequently leads to an unfavorable washback effect thus disadvantaging readers when they, for instance, study overseas and are expected to read quickly and efficiently in very limited periods of time. Therefore, the present reading test includes expeditious as well as careful reading operations as can be seen from the test specifications.

Table 1 Reading test specifications

Reading test specifications – Newspaper article (Derntl 2009: 71)	
General Description	
item description	Skill area description: READING Text type: NEWSPAPER ARTICLE A person who masters this reading test is required to demonstrate ability to comprehend advanced non-academic texts. Tasks included here are: ✓ utilizing text for study purposes: - skimming for main idea - scanning for specific information ✓ extensive reading comprehension: - summary of a single text - answering questions according to the text

Finally, the question arises whether first and second or foreign language reading are based on the same concepts. Are good first-language readers automatically good second or foreign-language readers?

Alderson (1984) addresses this question in his article entitled “Reading in a foreign language: a reading problem or language problem?”. By reviewing much of the research published at that time, he argues for the existence of a ‘language threshold’ beyond which second and foreign-language readers have to progress before their first-language abilities can transfer to the second language situation (Alderson 2000: 23). Further, he maintains that in second and foreign language reading both language knowledge as well as reading knowledge are essential, though the knowledge of the second or foreign language might constitute a more important factor than first language reading abilities (ibid).

Summing up, this chapter could hopefully show that although, at first sight, reading might appear to be the most common and “the easiest of the four skills to test” (Kitao 1999: 42), it is indeed a rather complex receptive skill which is not only difficult to explain but also difficult to assess. Nevertheless, language testers are required to have a concept of what it means to be able to read, i.e. that the reading process is comprised by decoding and comprehension processes. Unfortunately the ways in which the models of reading, i.e. top-down and bottom-up approaches, interact are still poorly understood (Alderson 2000: 20), which warrants further research. Due to the fact that readers engage in different kinds of reading, reading test constructors are advised to include in

their reading comprehension tests items which require careful as well as expeditious reading on part of the testees, so that the reading abilities of candidates can be assessed in a comprehensive way.

2.2.2. Different reading comprehension test formats under review

This chapter aims at explaining the three different test formats used in this study. Firstly, the advantages and disadvantages the three formats have in common will be described. Subsequently, each method with its advantages and shortcomings will be discussed. To begin with, a test method or test format can be described as

> the way in which candidates [are] required to interact with the test materials, particularly the response format, that is, the way in which the candidate [is] required to respond to the materials
> (McNamara 2000: 26).

In general, the testing literature uses the terms ‘test technique’ (Alderson, 2000), ‘test method’ (Bachman 1990) ‘test task’ (Bachman and Palmer 1996), ‘task type’ (Carr 2003), ‘response format’ (McNamara 2000) and ‘question format’ synonymously to refer to test format. According to Alderson (2000: 202) “the testing literature in general is unclear as to any possible difference between them”. Brown and Hudson (2002: 57), however, claim that the term ‘test question’ represents only test items in the interrogative form, which might suggest that ‘question format’ should only be used to refer to items which take an interrogative form. Except for the term ‘question format’ the terms mentioned above will be used synonymously in this paper.

A central element in defining a particular task type is the form of the expected response: selected-response, constructed-response or personal-response. Selected or fixed-response formats are also referred to as ‘recognition formats‘ (Van Blerkom 2009: 91), since the testees have to recognize the correct answer from among a set of alternatives. In constructed-response formats, however, the students cannot simply select an answer but they are required to produce or construct language themselves by writing, speaking or acting in some way (Brown & Hudson 2002: 71). Examples for constructed-response items are open-ended questions, short answer questions, fill-in exercises, essay questions, summaries, as well as cloze items (Wolf 1993:

474). Personal-response items require students to produce language as well, but as opposed to constructed-response items, they allow for different answers among students (Brown & Hudson 2002: 78). Personal-response items can take the form of oral presentations, conferences, written portfolios or self-assessment.

One aspect that the three test formats, which are compared in this study, have in common is the expected response type. All three test techniques: multiple choice, true-false-not given and matching are selected-response formats, meaning that the examinee can select the correct response from a set of supplied options (Brown & Hudson 2002: 59). Multiple choice, true-false-not given and matching tasks, "the most common kinds of selected response test items" (Popham 2002: 145), are frequently used in standardized language tests such as in the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS), the CAE (Certificate in Advanced English) and the Austrian Four-Skills Matura, to name a few.

Selected-response formats are good or as Brown and Hudson (2002: 64) claim "most appropriate" for testing the receptive skills of reading and listening as students do not have to create any language by themselves. One of the biggest advantages of selected-response over constructed-response items is that they can be easily, objectively and consistently scored, once a scoring-key has been developed (Brown & Hudson 2002: 64; Oosterhof 2001: 116; Oosterhof 2009: 95; Van Blerkom 2009: 90). If students mark responses on a separate answer sheet, their tests can be scored by a machine, which does not only simplify and shorten the scoring process even more but it also enables further analysis for diagnostic purposes provided that the machine enters the students' scores directly into a computer.

Another practical advantage of fixed-response items is that they allow teachers to sample more content during a test, since students can not only answer many more selected-response than constructed-response items in the same amount of time, but they can also answer them faster than constructed-response items (Gronlund 2003: 85; Van Blerkom 2009: 89).

Selected-response formats allow test constructors and teachers to control the range of possible answers to comprehension processes, thus enabling them to control students' thought processes to some extent when

responding. This can be regarded as another advantage of the fixed-response format (Alderson 2000: 211). This advantage is however closely linked to a disadvantage of the selected-response format. Compared to the constructed-response format they yield considerably less diagnostic information since the marker 'cannot inspect the process or path the respondent used in their attempt to solve the problem' (Robison 1989: 7). Therefore, it is impossible to determine whether the students decided for a certain answer, either the correct or the incorrect one, on the basis of knowledge or guessing.

The susceptibility to guessing is probably one of the most widely recognized disadvantages of selected-response formats. The various techniques, however, differ in their guessing factor with the true-false-not given and the multiple choice format being those techniques where the chances of getting an item correct simply by guessing are between 33.3% and 25%. The guessing factor constrains the reliability of a test (Van Blerkom 2009: 91). If a lot of guessing is involved "a student's performance will be inconsistent from item to item" (Oosterhof 2009: 96), which in turn affects negatively the generalizability of a student's performance on a test. What can be done to reduce the guessing factor is to use a large number of items in a test and to apply a correction for guessing formula where possible. Nevertheless, the correction for guessing procedure, which reduces the measurement problem induced by candidates' guessing the answers to test items (Davies et al. 1999: 35), is problematic in itself. The correction for guessing procedure assumes "that all candidates are affected by guessing to them same degree" (ibid), which probably might not be true. Thus, people involved in language testing have to decide on their own whether the application of the correction for guessing formula is more useful than problematic for their purpose.

Another limitation of a selected-response format is that the process involved in the selection of the correct answer(s) "bears little relation to the way language is used in most real-life situations" (Heaton 1988: 27), in which stimuli are produced rather than selected from among a set of different options. Certainly, this constraint to authenticity only holds true for items that aim at testing productive skills, i.e. the ability to use language, none of which appear in the present reading test. Therefore, the test constructor's choice to use selected-response formats to assess the receptive skill of reading is not only

perfectly justifiable but according to Brown and Hudson (2002: 64) even “most appropriate”.

2.2.2.1. Multiple choice

Multiple choice tasks are the most popular and highly regarded means of assessing students’ reading comprehension abilities (Alderson 2000. Brandtmeier 2005, Carr 2003, Heaton 1988, Robison 1989, Shehadeh 1997, Shohamy 1984, Van Blerkom 2009, Wolf 1993). This almost universally familiar test format is not only used in language tests but also in myriads of other types of educational examinations: driving tests, IQ tests, personality tests. Multiple choice items are commonly used as they are flexible and can be used to test a variety of content material (Van Blerkom 2009: 90). According to Oosterhof (2001: 115) “many group-administered standardized tests consist entirely of multiple choice items” because “they possess some psychometric qualities that make them highly desirable with organizations that produce standardized tests” (Van Blerkom 2009: 89), e.g. they can be scored easily, objectively and consistently.

Multiple choice items consist of two parts: a stem, which presents a problem situation in the form of a question or an incomplete statement, and a set of several, usually three, four or five alternative answers, options or choices, which provide possible solutions to the problem (Gronlund 2003: 60. 82). Students are required to select among the set of options the answer or the answers that provide(s) the best solution(s) to the problem. The correct option is generally known as the key (Heaton 1988: 28), while the remaining incorrect alternatives are referred to as distracters (Gronlund, Oosterhof), distractors (Alderson, Popham) or foils (Oosterhof, Van Blerkom). The aim of the distractors is, as the name implies, to distract testees who are unsure about an answer.

One virtue of multiple choice items is that they can contain several answers with differing degrees of relative correctness (Popham 2002: 134). Thus, students are asked to make subtle decisions among options, several of which might be somewhat correct. This helps teachers to control the thought processes test takers engage in to a certain extent. Van Blerkom (2009: 91) names an additional advantage of the multiple choice format. He claims that it is

more sensitive to partial knowledge than other techniques, since it is a recognition format, i.e. the correct answer is among the possible answers. Popham (2002: 134), however, criticizes that in selected-response and multiple choice tasks students only need to tick off a correct answer from a set of given answers. As they are not expected to generate a correct answer themselves the guessing factor might thus be increased.

The high guessing factor of multiple choice tasks is probably the most widely known criticism of this fixed-response format. According to Oosterhof (2009: 93) multiple choice items have even been nicknamed 'multiple-guess'. One possibility to reduce the guessing factor is to employ multiple choice items with four or more options, as in the present reading test, since this reduces the guessing factor to about 25%, compared to binary choice or multiple choice items with only three options (Brown & Hudson 2002: 68). The guessing factor, however, can also be reduced when the number of items is large. If students have to complete a multiple choice test consisting of 20 items, it is rather unlikely that they will get a score of 100% just by guessing correctly (Popham 2002: 130).

A further point of criticism from a test constructor's point of view is that multiple choice items are difficult to construct. It is often hard to find distractors which are plausible but clearly incorrect (Gronlund 2003: 67; Van Blerkom 2009: 91). The problem of partially correct distractors is that they could attract proficient students, who would have otherwise known the correct answer, thus reducing the reliability of the test (Van Blerkom 2009: 91).

One of the frequently heard problems of multiple choice questions is that some of them are not passage dependent, which means that testees could answer them based on guessing or their knowledge of the world, without having to read the text before (Shehadeh 1997: 16). Bernhardt (1991: 198) argues that a lack of passage dependency cannot be attributed alone to poor test construction since "even formally and professionally developed ones, fall into the passage independency category". The problem of passage independency is linked to another shortcoming of the most popular selected-response format: sometimes students can answer multiple choice questions without reading the text by simply matching words and phrases from the passages with words and phrases in the stem or alternatives (Cohen 1988: 71; Shehadeh 1997: 18; Wolf

1993: 475). Thus, it is rather questionable if such items that can be answered by surface matching are testing reading comprehension at all or rather test taking strategies (Wolf 1993: 475). To ensure that multiple choice tasks are indeed testing reading comprehension, test constructers are advised to write passage dependent items with a wording that differs from the one used in the passage.

2.2.2.2. True-False-Not Given

The true-false-not given format also known as 'alternate-choice format' (Oosterhof 2001, 2009: 139, 112), an elaboration of the dichotomous format (Alderson 2000: 222), binary choice format (Brown & Hudson 2002: 65) or true-false test, is one of the most popular and most widely used tests of reading comprehension (Heaton 1988: 113). According to Hughes (2003: 144) true-false-not given items are a variety of multiple choice with only two distractors. Students are presented with a statement about the target text and have to decide on the basis of the text whether this is true, false or whether the text does not say, i.e. it is not given. The advantage of the true-false-not given format against the true-false technique is that the chance of guessing of 50% can be reduced to about 33.3%. The disadvantage, however, is that the third category 'not given' can lead to confusion, especially if the items test the examinees' ability to infer meaning (Alderson 2000: 222).

An advantage of true-false-not given items is that they are typically so terse that students can answer numerous items within a short time. This makes it possible to test a large amount of content within one short assessment session (Popham 2002: 130). One virtue of true-false-not given tasks is that they are easier to construct than multiple choice and matching items, as the test constructor simply takes a salient point out of a text and asks the students whether this is true, false or not given. This advantage entails a further advantage:

> true-false and true-false-not given tasks can be used as a valuable teaching device with which the students' attention is directed to the salient points in the [text] by means of the true/false items (Heaton 1988: 114).

Brown & Hudson (2002: 66) claim that one advantage of binary choice items is that they focus on assessing the test taker's ability to discern between two alternatives, thus checking on whether a particular point has been understood or not. Therefore, the question arises whether the concept of something not given in the case of true-false-not given items does not blur the test taker's ability to distinguish between the alternatives. A further limitation of the true-false-not given technique is that items can only be developed for material which is unambiguous – which is either true, false or not given. The problem is, however, that in many fields there are relatively few materials which are absolutely true (Van Blerkom 2009: 106).

One shortcoming of true-false-not given tasks is that some items can be deceptive as they rely too much on the meaning of a single word or phrase or depend on some ambiguity (Brown & Hudson 2002: 66). When comparing the true-false-not given format to other selected-response formats such as multiple choice or matching, another disadvantage of true-false-not given items emerges: students are probably more influenced by guessing than with the other two tasks, as their chance of guessing a true–false-not given item correctly is 33.3%, whereas they only have a chance of 25% to guess a multiple choice item with 4 alternatives correctly (Gronlund 2003: 85).

What could be regarded as the biggest disadvantage of true-false-not given items is that the underlying concept of something not given, something non-existent can mislead and distract test takers and thus might blur their comprehension abilities and result in a non-accurate measure of their understanding.

2.2.2.3. Matching

The matching technique, which is also referred to as 'multiple matching' (Alderson 2000: 215) is an alternative objective testing technique, which is, according to the comparatively small number of literature available that goes into detail with this test format, less popular than multiple choice and true-false-not given tasks. Matching tasks require test-takers to match "two sets of stimuli against each other" (ibid), i.e. to match prompts to options (Brown & Hudson 2002: 67), as for instance, matching headlines for paragraphs with their corresponding paragraphs. Thus it is a selected or fixed-response format, as

the reader is not required to construct a response on his own. According to Alderson (2000: 218), Oosterhof (2001: 132) and Van Blerkom (2009: 100) matching items can be defined as a special case of multiple choice items, since there is usually a common set of choices of options for each item, where all but one act as distractors.

Like any other item type, matching items have advantages as well as shortcomings. Compared to multiple choice items, the guessing factor of matching items is comparatively low (Brown & Hudson 2002: 65).
While in 4 options multiple choice items with one correct answer students' chances of guessing an item correctly are 25%, their chances of guessing a matching item correctly are much smaller as students are required to select one answer from a common, large set of answers. What is important for matching items is that there are more options than prompts, i.e. that the number of alternatives is larger than the number of items (Popham 2002: 141), otherwise the students could improve their chances of guessing correctly by eliminating options (Brown & Hudson 2002: 68). Another virtue of matching items is that

> their compact form takes little space on a printed page, thus making it easy to tap a good deal of information efficiently (Popham 2002: 140) .

Certainly, matching items also have shortcomings. One major point of criticism of the matching format from a test constructor's point of view is that the items are difficult to construct. No choice should be unintentional (Alderson 2000: 218), and there should be only one option that is true for each prompt. Another disadvantage of the matching task is that it is "restricted to measuring students' abilities to associate one set of facts with another" (Brown & Hudson 2002: 65). This is especially crucial as these facts sometimes only test students' memorization of low-level factual information, which is not always a desired aspect when testing students' reading comprehension skills (Popham 2002: 140). Another shortcoming of matching items, which is also true for multiple choice items, is that some candidates might be distracted by choices they would otherwise not have considered (Alderson 2000: 219), which might prevent them from choosing the correct answer.

One disadvantage that was only discovered in the course of this study was that the test takers were not too familiar with this test format, a fact which

might not only have led to an increased nervousness but which might also have confused students, thus resulting in a not completely reliable and accurate measure of their reading comprehension abilities (cf. Popham 2002: 127).

In conclusion, this section could hopefully exemplify the respective advantages and shortcomings of the three selected-response formats. Test constructors aiming at creating a valid and reliable test should not only consider the general advantages and shortcomings of selected-response formats, but also the particular pros and cons of the subtypes of selected-response formats, i.e. multiple choice, true-false-not given, matching.

2.2.3. Factors affecting test performance – Test formats

This research paper explores whether various test formats, i.e. multiple choice, true-false-not given, and matching items, may influence differently a test taker's performance on reading comprehension tests. Thus, it follows that it is the aim to investigate in the following the effect that different types of questions can have on the testee's performance on reading test items.

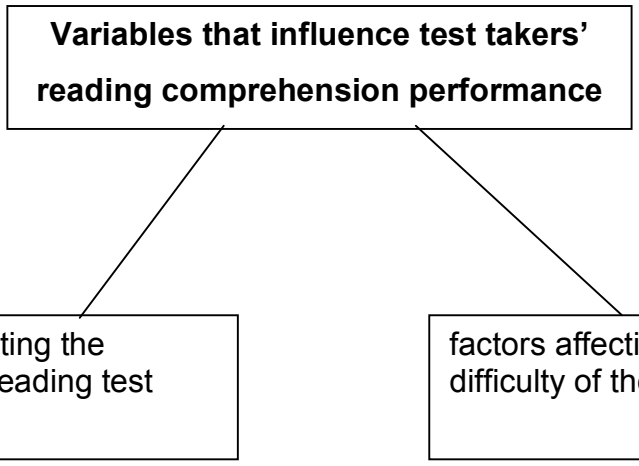
Broadly speaking, the variables that affect the nature of reading in general, and reading test performance in particular, can be grouped into three different variables: reader variables, text variables and test variables. The focus of the present chapter is on test and text variables. The following chapter 2.2.4, however, is devoted to reader variables that are especially relevant to the research assignment. As a thorough explanation of all these variables would go beyond the scope of this thesis, only variables that are particularly relevant to the present study will be explained in detail.

Alderson (2000: 60 f) lists as main text variables that affect the nature of reading: text topic and content, text type and genre, text organization, typographical features, and syntax, a traditional linguistic variable. Further text variables are text readability, i.e. syntactic complexity, and lexical density, verbal and non-verbal information (pictures) and the medium of text presentation (paper, overhead slides, computer screens, TV screens). For a detailed explanation of these text variables as well as research on them see Alderson (2000: 60-84).

Variables that influence test takers' performance on reading comprehension tests are further subdivided into factors affecting the difficulty of

reading test items and factors affecting the difficulty of reading test texts (Alderson 2000: 86 f.), which is exemplified in table 2 below.

Table 2 Variables that influence test takers' reading comprehension performance



• language of questions	• background knowledge
• types of questions	• presence of text while answering questions
• testees' language skills	• text length
• use of dictionaries	

Among different types of question Alderson (2000: 87) distinguishes between items that focus on one part of the text or on a whole passage. Further, he categorizes item types according to the format of their expected response. Items focusing on one part of the text, where both the information to the question as well as the correct answer are to be found in the same sentence, are referred to as textually explicit questions (ibid). Conversely, textually implicit questions require the testee to synthesize information across sentences or the whole text. A further distinction which relates to the location of the answer to the items can be drawn between local and global comprehension. Among the formats of the expected responses a distinction can be made between selected-response and constructed-response items, which has already been explained in more detail in chapter 2.2.2. The remainder of this chapter focuses on studies

investigating the effect of 'question type' on testees' reading comprehension performance[1].

Researchers like Alderson (2000), Clapham and Wall (1995) acknowledge the importance of studies comparing different test formats for the investigation of the so-called test method effect. The test method effect, that is that a method used for testing a language ability affects the testee's score, should be avoided as far as possible (Alderson, et al. 1995: 44; Alderson 2000: 123-4). Language testers are interested in finding out about a testee's reading ability rather than whether a candidate is good at multiple-choice tests, or can do matching tests better than other candidates (Alderson, et al. 2005: 44). Brandtmeier (2005), Shehadeh (1997), and Wolf (1991, 1993) argue that despite the importance of research on this topic, to date only a handful of studies compare test types and explore their effect on the reader's performance.

Research on the comparison of reading comprehension assessment methods is not only scarce, but also disparate with regard to the following aspects. While some studies compare multiple choice and open-ended tasks (Elinor 1997; Shohamy 1984), others focus on multiple choice and cloze tasks (Bensoussan 1984) or multiple choice tasks, open-ended tasks and cloze tasks at the same time (Wolf 1991, 1993). Despite the variety of test formats compared, what all of the relevant studies have in common is that they compare the testees' performance on selected-response formats, i.e. multiple choice, to constructed-response formats, i.e. open-ended tasks, cloze tasks.

A further disparity of the studies is the language in which comprehension is assessed. While the majority of studies took place in an EFL context (Bensoussan 1984; Elinor 1993; Shohamy 1984) other studies investigated foreign languages other than English, i.e. Spanish (Wolf 1991, 1993). Some studies compare the test takers' L1 to their FL reading comprehension performance (Shohamy 1984; Elinor 1997; Wolf 1991, 1993). Other studies, like the present study and Benoussan's (1984), however, only compare the testees' EFL reading comprehension abilities across different test formats (Bensoussan 1984).

[1] In the present paper the terms 'test technique', 'test method', 'test task', 'task type', item format' and 'response format' will be synonymously used (cf. chapter 2.2.2. pg. 19).

Another difference is to be found in the alphabet of the testees' native language. While in some studies the testees share the same alphabet, i.e. the Latin alphabet (Wolf 1991, 1993), other studies investigate contexts where the students' native language relies on a different alphabet. The research projects by Shohamy (1984), Bensoussan (1984) and Elinor (1997) took place in Israel where testees' native languages, i.e. Hebrew and Arabic, have a different sign system (alphabet). The problem with learners whose native language is based on alphabets other than Latin is that they tend to need more time to be able to read proficiently in a foreign language with a different sign system. This fact leads to another disparity among the scarcely available studies: the target language experience of the subjects tested.

While some studies test prospective college students (Bensoussan 1984; Elinor 1997) others focus on advanced college students (Wolf 1991, 1993) or grade 12 high school students (Shohamy 1984). The present study investigates the reading comprehension performance of grade 11 high school students. It is important to consider the proficiency levels of the test takers in the studies as they might exert an influence on the students' familiarity with test formats, which could consequently improve their performance on otherwise rather unfamiliar task types, i.e. matching.

Taking all the above mentioned aspects into consideration it is still difficult to make a comparison between the few existing research studies on the comparison of test formats. The problem is that a further disparity is related to their differing research designs. The majority of studies (Elinor 1997; Shohamy 1984; Wolf 1991, 1993) employs between-subjects designs, i.e. one group of students is tested on the selected-response formats, the other group is assessed on open-ended formats and then these two groups are compared. Other studies (Bensoussan 1984, the present study), however, investigate within-subjects designs, i.e. all subjects perform the same test tasks.

Another diversity among the research studies is related to the number of texts testees have to read. In most studies (Bensoussan 1984; Elinor 1997; Wolf 1991, 1993) all students were required to read the same and only one text. In a study conducted by Shohamy (1984), however, students were divided into two groups and each group read a different text. As previously mentioned, each student just had to do one test format. Shohamy (1984: 153) then compared the

test takers' results on the respective test formats and texts and found that students' performance on one text was significantly higher than on the other text, leading her to conclude that one text was more difficult. The order of the difficulty of the test formats, however, remained constant across the two texts. Despite the division of test takers into many groups, Shohamy's study is sound due to the large number of test takers, i.e. 655. Due to the fact that the present study is a replication study that aimed at comparing an already existing reading comprehension test only one text was used, as in the original study. Further, each test taker had to do all three tasks, again in line with the original study.

The large number of different research designs among the available studies exemplifies that there is no one best method to compare test formats, though some methods might be more suitable than others.

Having covered the different research designs the question arises as to how the different question formats are constructed. Wolf (1991, 1993) posits that

> to compare directly the effects of different comprehension assessment tasks on test takers' performance on those tasks […] [they] must assess the same information.

Elinor (1997) compares in her study EFL university students' performance on multiple choice and open-ended tasks. To accomplish this she constructed multiple choice questions first and then used the stem versions without the alternatives for the open-ended questions. Shohamy (1984) proceeded in the creation of open-ended tasks for her study in a similar way. In contrast to Elinor (1997), Wolf (1991, 1993) constructed open-ended questions first, which she then rewrote so that the responses required for a particular item in one task corresponded to an item in each of the other tasks. Rational deletion cloze tasks were conducted as a post-reading assessment task which took the form of a summary of the passage. Wolf paraphrased the information so that what had to be filled into the gaps corresponded to the respective answers to the multiple choice and open-ended items. Bensoussan (1984) compared the effects of multiple choice and cloze formats on students' performance on an already existing test, the Haifa University English reading comprehension examination, and does not offer any information neither on how these items were constructed nor on whether they assess the same passage of the

underlying reading text. Having covered how the different researchers proceed in the comparison of various test formats the question arises as to how students' performance on selected-response or constructed-response reading comprehension items differs.

In'nami and Koizumi (2009) conducted a meta-analysis of test format effects on reading and listening test performance with the focus on multiple choice and open-ended formats. The advantage of a meta analysis over a narrative literature review is that a meta analysis reanalyzes and reinterprets previous studies by taking the study characteristics into account. Thus, researchers employing meta-analysis do not have to blindly rely on conclusions drawn from authors of the original studies (In'nami, Koizumi 2009: 220. 222). In'nami and Koizumi (2009: 219) did not find any overall format effects in L2 reading. Nevertheless, they discovered that multiple choice formats were easier than open-ended formats under any of the following conditions: between-subjects design, random assignment, stem-equivalent items, learners with a high L2 proficiency level.

Indeed, studies using a within-subjects design, e.g. Shohamy (1984) and Wolf (1991, 1993), showed that multiple choice tasks are easier than open-ended formats. The only exception is Elinor's (1997) study in which the two formats prove to be of similar difficulty. In Bensoussan's (1984) study again no significant differences between students' performance on the multiple choice and the cloze tasks of the reading comprehension test are to be found, but in contrast to Elinor she used a within-subjects design.

Shohamy (1984) also investigated whether the testing methods affect high-proficiency and low-proficiency students differently. In her study it turned out that low-level learners were more affected by the testing method, while for high-level students different test formats did not matter. Shohamy (1984: 158) posits that advanced level testees might not be affected by the testing method, since they are able to manipulate the language better than low-proficiency test takers. Wolf (1991, 1993), however, contradicts Shohamy, as in her study advanced-level students did not perform equally well regardless of the test formats.

Considering all aspects covered in this part of the literature review, it is apparent that the present study attempts to fill a niche by comparing testees'

reading comprehension performance on three selected-response formats, i.e. multiple choice, true-false-not given, and matching. Further, it will be investigated as to what extent high-proficiency and low-proficiency students are influenced differently by these test formats, which will then be compared to Shohamy's and Wolf's research results.

2.2.4. Reader attributes affecting reading comprehension performance

While the previous section has examined factors that affect test performance, the present one aims at investigating reader attributes that may affect reading comprehension performance. Alderson (2002: 33) established the following main reader variables which can be divided into subcategories: reader's knowledge, reader's motivation to read and reader's interest in the text, strategies adopted by readers when processing text. Furthermore, Alderson mentions stable reader characteristics such as sex, age and personality, as well as physical characteristics like eye movements, speed of word recognition and automaticity of processing. The present chapter is divided into two subchapters. While the first subchapter intends to give an overview of literature examining gender differences in reading comprehension performance and test format, the second subchapter is dedicated to another reader variable: interest.

Although some research has been devoted to either gender differences in test formats or in reading comprehension performance, rather less attention has been paid to studies focusing on these three aspects: test format, gender, and the EFL context at the same time (cf. Yazdanpanah 2007: 69,75). Thus, the present study attempts to fill this niche. Concerning response formats, the research has tended to focus on a comparison of selected and constructed response formats, rather than on gender performance in various selected response formats. More specifically, the only selected response format that has been paid attention to in the relevant studies is the multiple choice format. Other fixed-response formats as true-false-not given and matching seem to have been neglected so far.

Another problem of the research is the lack of EFL specific studies. While there are a few studies examining gender differences in reading comprehension in the EFL context, no study to this date has investigated gender differences on

test formats in EFL (cf. Yazdanpanah 2007: 69). More specifically, the majority of studies on gender differences in test formats focuses on disciplines such as natural sciences, i.e. mathematics or medicine. Due to the range of different disciplines it is little surprising that a further disparity of studies on gender differences is the test takers' native language.

While the majority of studies were conducted in countries where English is the Native Language, i.e. Australia, Great Britain, United States, fewer took place in countries where English is used as a Foreign Language (EFL). To be more precise, the EFL studies that are relevant for the purpose of this paper took place in Cyprus (Yazdanpanah 2007) and Thailand (Phakiti 2003). Therefore, the question arises as to what extent studies on gender differences which have taken place in different countries are comparable, since the positions of men and women in society, their role models as well as what are regarded as appropriate or inappropriate male and female activities and disciplines of study, differ from culture to culture.

Given these differences not only in disciplines but also in the underlying language (English as a Native Language, English as a Foreign Language), and culture it is little surprise that studies on gender differences present inconsistent findings.

2.2.4.1. Gender

2.2.4.1.1. Gender differences in test format

The effects of test format on female and male test takers' performance in disciplines like mathematics, science, history and English as a native language, have been widely studied with rather conflicting results. One aspect that most studies have in common is that they examine the difference between multiple choice as selected response format, and free-response formats, as for instance short answer or extended response items. The present study will be one of the first in examining gender differences in more than one selected-response format, i.e. multiple choice, true-false-not given, and matching.

Interestingly, research focuses on contexts other than EFL within which comparisons across even rather different disciplines are drawn. To be more precise, some studies compare whether male and female test takers perform differently on multiple choice and free-response tests in Mathematics and

English. This non-uniformity of disciplines as well as the period of time when the studies took place, which encompasses more than 20 years, might also have contributed to diverging results.

Broadly speaking, studies investigating gender differences in test format can be subdivided into the following categories according to their outcomes:

- Girls are disadvantaged by multiple choice tests
- Boys outperform girls
- Boys and girls perform the same
- Girls surpass boys

Geering (1993: 25) cites a report of a study by the University of London School Examinations Board undertaken in 1985 devoted to discovering biases of examination components. The report claimed that there is considerable evidence that females are disadvantaged by the multiple choice format compared to other test formats. Clark and Grandy as referred to in Geering (1993: 24) expanded on the results of this report in their research on sex differences in academic performance of women and men by examining their grades on the SAT, i.e. Scholastic Aptitude Test, and their college freshmen year grades. The SAT is a standardized college admission test in the United States. The majority of questions in the SAT take the multiple choice format (http://sat.collegeboard.com/home, 21.01.2010). According to Geering (1993: 24) Clark and Grandy argued that first-year college grades of female test takers were slightly underpredicted by their test scores. The same was discovered in a recent study by Christiane Spiel (2008: 39) investigating male and female test takers' scores on the admission test for Austrian Medical Universities, entirely consisting of multiple choice tasks, and their scores on science subjects featured in the admission test at school. Girls outperformed boys on science subjects at school. Their superiority, however, was not reflected in their score on the admission test, where they were not only underpredicted but even outperformed by the male test takers (ibid). Thus the question arises whether female test takers' performance can be adequately measured by means of multiple choice tests.

Geering (1993: 23) comments on a highly interesting study undertaken by Murphy in 1980. in which a Geography Examination before and after the introduction of multiple choice tasks were compared. Before the introduction of

multiple choice tasks male and female candidates were said to perform the same. After the introduction of the multiple choice tasks, however, boys achieved better scores than girls. This suggests that multiple choice tasks favor boys but underpredict girls. Geering (1993: 27) further reports on work of Breland who compared two advanced placement examinations: United States History and European History consisting of free-response and multiple choice parts. While in the free-response parts no gender differences could be found, boys were significantly and to a large effect size favored by multiple choice tasks.

Bolger and Kellaghan as referred to in Geering (1993: 32) even considered three different school subjects: Mathematics, Irish, and English, to investigate gender differences in results of multiple choice and free response tests. The researchers found that male test takers generally surpassed females in all subjects and on both test formats. In language subjects, however, gender differences diminished and boys only performed slightly better than girls. They concluded that male participants performed significantly better on the multiple choice than on the free-response formats, whereas for girls the opposite was true.

The disadvantage of female test takers on multiple choice tests in all of the above mentioned studies raises the question why boys seem to be favored by this selected-response format. One argument, cited in numerous studies (Spiel 2009, Freeman 2007, Barboza 1993, Geering 1993, Hardcastle 1991), might be that boys are more likely to guess and to employ risk taking strategies than girls. Barboza (1999: 33) expands on this further by adding that while boys tend to risk a guess, girls tend to leave a blank. Stobart, Elwood and Quinlan as referred to in Geering (1993: 39) argue that the superiority of boys is related to their employing of a so-called “eyes down” approach, which they claim is better suited for selecting a correct answer out of a set of options. Girls, however, appear to be inhibited by seeing the relative “rightness/wrongness” of the items. Ryan, as referred to by Barboza (1999: 33), points out that female test takers tend to see “unintended nuances” in the set of answers which in turn makes it harder for them to guess. A further aspect that might contribute to male test takers’ superiority on multiple choice tasks is that females are said to have higher levels of test anxiety than males (Phakiti 2003: 670). Phakiti (2003: 670)

claims that while males are more likely to see a test situation as a “personal challenge”, females tend to regard a test situation as a threat, which according to Phakiti “leads to states of fear, and worry” (ibid).

While numerous studies claim that boys are favored by multiple choice tests, other studies argue that there is no significant difference between male and female test takers’ performance. Geering (1993: 27,28) refers to a study by Huntley in which no significant differences between males and females on a multiple choice geometry test, administered to 3,000 testees, could be detected. Interesting results also emerged from a recent study undertaken by Freeman (2007: 89) on gender differences in the reading test part of the Ohio Graduation Test (OGT), which contains multiple choice, short answer, and extended response items. Freeman discovered that male and female test takers’ scores did not significantly differ, which he suggests is contrary to most of the literature found on gender differences in multiple choice tests (ibid).

In contrast to the above mentioned researchers, Doolittle and Welch as quoted by Geering (1993: 28) found that female testees outperformed male testees in the multiple choice writing test of the Collegiate Assessment of Academic Proficiency (CAAP). In the mathematical part of the CAAP test, however, male test takers surpassed female test takers. This outcome poses the question whether gender differences are related to the subject content and not so much to the test format. An ESSSA (Equity in Senior Secondary Schools Assessment) project cited by Geering (1993: 38), however, contradicts this assumption. Geering quotes that the ESSSA report in the subject English indicated that female testees surpassed male testees on all types of essay questions, whereas boys outperformed girls on multiple choice questions, with the exception of questions about people and personal relationships. This result again highlights that boys might be favored by the multiple choice format, even in disciplines like languages where they could be outperformed by girls.

2.2.4.1.2. Gender differences in reading comprehension

Previous research on gender differences in EFL reading comprehension performance is relatively scarce and has produced rather conflicting outcomes so far (cf. Phakiti 2003: 652). While in some studies gender differences could not be detected, other studies found that males outperformed females on

reading comprehension tasks or that the contrary proved to be true: female testees surpassed male testees.

Yazdanpanah conducted a study investigating the interaction of a reading comprehension test with gender in a School of Foreign Languages in North Cyprus. He used three different reading comprehension passages: two out of which could be identified, following Bügel's and Buunk's classification, as 'male' topics, i.e. 'the latest technology used in the design of houses' and 'space travel'. The third passage featured a neutral topic, i.e. 'how to make changes in life'. Yazdanpanah (2007: 71, 73) found that females scored slightly, though not significantly, higher than men on the reading test, despite the fact that two out of three reading passages featured male topics. This result made him conclude that text topic does not influence male and female performance on the reading test (Yazdanpanah 2007: 64). Yazdanpanah (2003: 68, 69) reports on a similar outcome by referring to a study by Lin & Wu, wherein performance differences of male and female Chinese university graduates on the reading comprehension part of an English proficiency exam, modeled after the TOEFL test, were examined. In this study again no differences in the performance of both genders could be detected.

These findings conform to a study by Phakiti undertaken in 2003. Phakiti examined gender differences in the context of an official EFL reading comprehension test in Thailand, administered to 384 first year undergraduate students in order to make high-stakes decisions regarding students' achievement at university. Phakiti (2003: 668) found that males and females did not differ in their reading comprehension performance as assessed by a multiple choice reading comprehension test. Phakiti grouped testees into three proficiency levels according to their overall result. But male and female test takers of even the same achievement levels did not differ in their reading performance (Phakiti 2003: 672).

Yazdanpanah (2007: 68) refers to a highly interesting study by Brandtmeier, which took place in a non-EFL context, i.e. Spanish as a Foreign Language, in the United States. Brandtmeier's study, investigating the effect of gender on reading comprehension of intermediate and advanced level students studying Spanish, produced divergent as well as interesting results. Her study consisted of two different reading passages: one about boxing, which is

considered a ‘male’ topic, and another one on housewifery, a ‘female’ topic. While advanced level male and female test takers’ performance on the two texts did not significantly differ, intermediate level male and female test takers’ performance proved to be significantly different. Male testees outperformed their female counterparts on the text about boxing, whereas female test takers surpassed males on the female topic on housewifery.

An instance of male and female superiority on their respective topics in a reading comprehension test is also reported by Yazdanpanah (2007: 68) who refers to a study by Bügel and Buunk investigating the reading performance of Dutch EFL students. In this study females scored significantly higher on ‘female’ topics as ‘midwives, a sad story, marriage dilemma, and talks about style’, whereas their male counterparts outperformed them on male topics about ‘laser thermometers, volcanoes, cars, and football players’.

Still other studies found that female testees generally outperform males on reading tests. Phakiti (2003: 652) refers to a study conducted by Wen and Johnson in 1997 in the context of a standardized national proficiency test for tertiary-level English majors in China. They discovered that females scored significantly higher than males on the reading part of this standardized national proficiency test. These results correspond to those of a study conducted by Chavez in 2001, referred to by Phakiti (2003: 652). Chavez found that regardless of topics females demonstrated significantly higher levels of performance than males on a multiple choice reading test.

In conclusion, research has produced rather inconclusive evidence of gender differences on both reading comprehension performance and test format. Reasons for the non-uniformity of the results could be that especially in EFL studies the language background differs from study to study. Thus, cultural influences as well as models of what activities and disciplines are regarded as appropriate for males and females might play a role. Some 30 years ago scientific subjects might have been considered as truly male domains, while subjects as languages were seen as female subjects. This strict division into male and female domains, however, seems to have been loosening since then. Another reason for the differing results is that test formats across diverse subject domains, i.e. languages, mathematics, and medicine, were compared. Furthermore, the testing conditions under which the studies were conducted

differed from low-stakes to high-stakes contexts. Due to the fact that testing conditions might extensively influence the performance of learners (cf. Phakiti 2003: 68, Yazdanpanah 2007: 69), one reason for the differing results could be found in this non-uniformity of contexts. Thus, the present study is devoted to shed some light on gender differences in reading comprehension performance and test formats, a domain which is according to Yazdanpanah (2007: 69) in need of further research.

2.2.4.2. Interest[2]

This chapter is devoted to an investigation of the relationship between test takers' interest and their reading comprehension performance. Although the present research paper can only examine this relationship to a rather small extent, it is important to be aware of the role that interest can exert.

To begin with, a definition of what interest actually is should not be missing. Shirley (1992), as quoted by Le Loup (1993: 3), defines interest as a reader-specific, internal characteristic that the reader brings to the reading situation. Commonly, the concept of interest is subdivided into individual and situational interest. While individual interest can be exemplified as "a psychological state within a person" (Le Loup 1993: 2), situational interest is, as the name implies, a feeling or reaction triggered by an outside stimulus, as for instance a highly interesting text (ibid).

The available studies investigating the role of interest on reading comprehension focus on a number of different test formats, i.e. cloze test, free recall test, and free response format. Nevertheless, there seems to be a shortage of studies focusing on constructed response formats, such as multiple choice. Another problematic aspect of the existing studies is that some lack a specification of the concept of interest (Le Loup 1993: 4). Therefore, differences in research design are manifold.

While some researchers predetermined the interest level of a test without any prior survey of their students' interests (Bernstein 1955; Stanchfield 1967 as referred to in Le Loup 1993: 12), others gave the test takers the chance to

[2] For reasons of practicality the concepts of gender and interest are treated separately in this thesis. Nevertheless, the researcher acknowledges that interest and gender are interrelated. Interest could of course also be the relevant factor underlying gender.

choose what they considered as the most and least interesting topics from among a pool of different topics (Asher & Markell 1974; Belloni & Jongsma 1978; Le Loup 1993). The actual reading comprehension test then featured each student's individual high and low interest material (Asher & Markell 1974; Belloni & Jongsma 1978; Le Loup 1993). Still other researchers asked students to assess the level of interest of texts only after they had completed the reading comprehension test (Alexander et al. 1994). This method, however, appears to be of a rather questionable nature, since interest might be interrelated with difficulty. If students perceive a text they had previously identified as highly interesting, as extremely difficult and hard to read this could make them lose interest in the topic. On the other hand, if students read a less interesting but easily readable text and can answer the reading comprehension tasks to their satisfaction, this could enhance their interest in the text and the topic as such.

Another central disparity of the research is concerned with the underlying language. While most studies focus on L1 contexts, where English is used as a native language, others center on the EFL or the ESL context, while still others investigate the role of interest in languages other than English, i.e. Spanish (cf. Le Loup 1993). Taking into consideration the aspect that "the L2 reading process is not merely the L1 process with different words" (Le Loup 1993: 30), results from the L1 context are not completely transferable to other language contexts. Nevertheless, cautious comparisons between studies on the role of interest in the various language contexts seem to be warranted as pupils' interest in certain topics is supposed to remain the same across languages.

Furthermore, the variety of the test takers' proficiency levels ranges from L1 elementary school children, fourth to tenth graders to EFL university students. Regardless of the numerous differences in research design the results are consistent. All studies agree that interest influences reading comprehension performance positively (Belloni & Jongsma 1978, Alexander et al. 1994; Asher & Markell 1974, Le Loup 1993, Oakhill & Petrides 2007).

As early as 1955 Bernstein (cf. Le Loup 1993: 12) found that in a L1 context grade 9 pupils, who had read a high interest and a low interest story, performed significantly better on what they had rated as the highly interesting text. Le Loup (1993: 13) cites a study by Estes and Vaughan (1973), which showed similar results and underscored the assumption that interest could be a

powerful factor in determining students' reading comprehension performance. In Le Loup's study investigating the role of interest on L2 reading comprehension in Spanish, testees had to rank order 5 Spanish newspaper articles according to their level of interest, ranging from high to low. In the reading comprehension test, which featured each student's individual high and low interest topic, i.e. newspaper article, students were asked to write a recall protocol in their first language, English. Results are in accordance with the aforementioned studies and show that while a high level of interest facilitates reading comprehension, little or no interest in the topic hinders reading comprehension (Le Loup 1993: 99).

Here the question arises in how far a high level of interest can influence comprehension positively. Theories about this link between interest and performance are highly speculative. Nevertheless Oakhill and Petrides (2007) offer a theory which appears to be sound. They assume that a high level of interest in a text is interrelated with a motivation to understand a text covering this topic. It is this motivation which then activates cognitive processes, i.e. text- and knowledge-based inferences, that are central to reading comprehension (Oakhill and Petrides 2007: 232).

In contrast to some other researchers, Asher and Markell (1974: 680. 681) did not only consider the variable of interest but also gender differences as well as various proficiency levels in investigating testees' reading comprehension performance, which they assessed with a cloze procedure. First of all, they applied a rather innovative technique to specify the 5th grade students' interests: the picture rating technique. Furthermore, their testees were asked to read not only one high and one low interest text, as it is common in other studies, but those three high and low interest passages which they had previously and individually selected from among a set of 25 topics. Asher and Markell's L1 study produced valuable results. They found that boys performed as well as girls on the high-interest material, whereas male test takers were significantly outperformed by females on the low-interest passages. This led the researchers to the conclusion that boys appear to be more affected by the level of interest than girls (Asher and Markell 1974: 684). Thus, they averred that sex differences in reading performance were only significant in low interest material (Asher and Markell 1974: 685). Considering the various proficiency levels, the

level of interest influenced high and low achieving students to the same extent. While reading comprehension performance of both high and low achieving students was significantly enhanced by high-interest material, low-interest material had a negative effect on reading comprehension. Le Loup (1993: 14, 15), as well as Oakhill and Petrides (2007: 224) refer to a study by Anderson, Shirley, Wilson & Fielding, affirming Asher and Markell's claim that male test takers are more influenced by the effect of interest on L1 reading comprehension than females.

The question then arises as to why the interest level of the material effects male test takers' reading comprehension performance more than females. Research outcomes by Ainley et al. (cf. Oakhill and Petrides 2007: 224 f.) show that one reason why boys are more influenced by the level of interest could be found in girls' perseverance. Ainley et al. (ibid) claim that while girls try hard to be able to understand a text even under low- interest conditions, boys tend to immediately put down what they regard as an uninteresting text and give up.

Oakhill and Petrides (2007: 231) and Asher and Markell (1974: 685) agree on another possible explanation. They claim that one of the reasons could be that reading is often regarded as a feminine activity: sex-appropriate for girls but sex-inappropriate for boys. Consequently, girls might read well regardless of the interest level of the task. Boys, however who see reading as a sex-inappropriate activity might thus need an additional incentive, such as a highly interesting text, which helps them disregard the sex-inappropriateness. Additionally and by referring to a recent longitudinal questionnaire study on voluntary reading conducted by the University of Sussex among grade 3, 4 and 6 students, Oakhill and Petrides (2007: 231) suggest that the reason why girls are left rather unaffected by a text's level of interest might also result from them being more engaged in voluntary out of school reading. Their exposure to more reading material outside of school could help them acquire familiarity with vocabulary in a wide range of topic areas, which consequently might put them at an advantage even in low-interest topics. Boys, however, might only perform well on high interest tasks containing the vocabulary they are familiar with.

In contrast to the aforementioned studies, research projects by Shnayer, Vaughan and Walker did not focus on the effect of interest on gender but on

whether pupils across various proficiency levels were affected differently by the text's level of interest. Shnayer (1969), Vaughan (1975), and Walker et al. (1979), as referred to in Le Loup (1993: 16), found that less proficient readers were significantly more affected by the level of interest than highly proficient readers. Walker et al., as referred to by Le Loup (1993: 16), investigated three different proficiency levels: highly proficient, proficient and low proficient readers. Walker, as referred to in Le Loup (1993: 16), found that while the impact of interest on above average readers was of a rather questionable manner, average and below average readers were significantly affected by the level of interest. A possible explanation why less proficient students were significantly more affected by the topic's level of interest could be again that highly proficient students might tend to engage in a broader range of reading activities in their leisure time. This extended exposure to reading material could in turn help them acquire a familiarity with a wider range of topics and their respective vocabulary, an aspect that might advantage them in reading low-interest material.

In short, studies on the role of interest in reading comprehension differ widely concerning their research designs. Nevertheless a large number of studies produced conclusive results. The majority of research studies suggests that interest plays a vital role and influences students' reading comprehension performance positively (Asher & Markell 1974, Belloni & Jongsma 1978, Le Loup 1993, Oakhill & Petrides 2007). Other researchers do not only agree on the vital role of interest in reading comprehension, but they also found out that the level of interest influences boys and girls differently. More specifically, they averred that boys are significantly more affected by the level of interest than girls. Asher and Markell assume that this difference based on gender might be related to children's sex stereotypical views of reading. Furthermore, it was demonstrated that the effect of interest is greater for low proficiency than for high proficiency students, a difference which can probably be traced back to the respective amount of leisure reading activities high and low proficiency students engage in.

The present chapter was devoted to a review of the relevant literature. While the first section (2.1) introduced the reader to the most important concepts of language testing, the second section (2.2.1) outlined the nature of

reading. More specifically, questions like what it means to be able to read were explored and different models of reading were exemplified. It was up to the third section (2.2.2) to present the three selected-response test formats under investigation with all its advantages and disadvantages. While the first sections of the literature review were concerned with the theoretical background of the study, sections 2.2.3 and 2.2.4 focused on studies with similar research interests, which can be compared to the present study further on in this thesis. Owing to the fact that the aim of the present paper is to investigate the influence of three different selected-response formats on students' reading comprehension performance, section 2.2.3 reviewed research studies exploring the influence that test formats can have on students' test performance. The last section 2.2.4 gave an overview of literature investigating how the reader attributes gender and interest are affecting testees' reading comprehension performance.

Taking all aspects of chapter two into consideration which are important for the present research study, it is apparent that the present study attempts to fill a niche as no study up to this date has investigated the differences between three different selected-response test formats and their influence on students' reading comprehension ability. Although there are studies investigating the influence of test format on reading comprehension (Bensoussan 1984, Elinor 1993, Shohamy 1984, Wolf 1991, 1993), no study to date has investigated three selected-response formats. Additionally, students will be divided into two proficiency levels in order to check whether highly proficient students are less influenced by test format than low proficient ones. These outcomes will be compared to Shohamy's (1984) and Bensoussan's (1984) studies which investigated whether various reading comprehension test methods affect high- and low-proficiency students differently. Further, any gender differences regarding students' reading comprehension performance on the three different test formats, i.e. multiple-choice, true-false-not given, and matching will be explored. Research on gender differences in reading comprehension either investigated gender differences on various test formats (Barboza 1993, Geering 1998, Freeman 2007, Hardcastle 1991, Spiel 2008) or gender differences in reading comprehension more generally (Brandtmeier 2005, Phakiti 2003, Yazdanpanah 2007). The results of these studies differed widely, thus it is

interesting what the outcomes of the present study will be. Additionally, the variable of interest and its influence on students' reading comprehension performance will be investigated. The outcomes of the present study will be compared to studies by Asher & Markell (1974), Belloni & Jongsma (1978), Le Loup (1993), Oakhill & Petrides (2007) which all suggest that interest plays a vital role and influences students' reading comprehension performance positively.

3. Methodology

3.1. Materials

3.1.1. The original study

The present study is a replication study. The original study was conducted from 2008 until 2010 by a group of students of the English and American Studies department of the University of Vienna. They worked in the context of an ongoing language program evaluation under the lead of Prof. Christiane Dalton-Puffer. The aim of this language program evaluation, which had been commissioned by the parents' association of a Viennese high school, was to independently evaluate the learning outcomes of two different tracks by the time of the "Matura", i.e. the Austrian school leaving exam. These tracks differ with regard to the first foreign language taught at school. While one track starts with English in the first grade, pupils in the other track begin with French. To find out about any differences in language competence between the two tracks in the final year, the project group developed a comprehensive test of the four skills (listening, reading, writing, speaking).

3.1.2. The present study

The purpose of the present study is to examine the effects of different selected-response assessment task types on EFL students' ability to demonstrate reading comprehension. The following assessment tasks were used to examine task effects: multiple choice, true-false-not given, and matching. Therefore, it was useful to use one part of the reading test developed by the aforementioned project group, as it featured exactly multiple choice, true-false-not given, and matching items. The reading test will be described in more detail in sections 3.1.3, 3.1.1.2, as well as in appendix 1.

To test the effects of test types on performance all students took the same reading test. In an attempt to investigate whether the number of correct answers given on the multiple choice task affects students, one group of students knew how many answers to each item were correct, while the other group did not. Further, the present study wants to find out whether students are affected by the level of interest, although this can be only accomplished to a

rather small extent. In order to examine whether the text on a topic that was not highly relevant at the time of the examination, i.e. tuition fees, is more interesting when the publishing year 2003, i.e. six years ago, is left out, one group of students had the publishing date given while the other had not. A more detailed description of the research questions guiding the present study is to be found in chapter 4.

3.1.3. The reading test

The reading test used in the original study consisted of two reading texts: one newspaper article about tuition fees, and a literary text entitled “The Bully”. While the questions to the newspaper article were set in multiple choice, true-false-not given, and matching formats, the response formats in the literary texts were matching and short answer questions. Corresponding to the purpose of this paper, namely to compare three selected response formats, only the first part of the original reading test, the newspaper article on tuition fees, was in use. The reading test is to be found in the appendix (cf. section 7.1.).

3.1.3.1. The test development process

Although the reading test was taken from an already existing study, a conscientious researcher should always investigate the test development process of the original test and compare it to what research literature suggests regarding test development. Handbooks on language testing, as for instance McNamara (2000), Hughes (2003), Davies (1990), Bachman & Palmer (1996) all comment on the test development process. These researchers all group the test development process into various stages, with differing degrees of complexity. Bachman & Palmer (1996: 87-91) classify the test development process into thee main stages: design, operationalization and administration. McNamara (2000: 25 ff.), however, proposes four stages of the test development process: establishing test content, establishing test method, writing test specifications, trialling and trying out. Similarly, Derntl (2009: 60 ff.), who was part of the project team and whose MA thesis describes the development of the original test, supports McNamara’s categorization of the test development process.

3.1.3.1.1. Test content and test formats

Following McNamara's classification, the test content of the reading test is, as the name implies, the reading competence. The test method, i.e. "the way in which the candidate will be required to respond to the materials" (McNamara 2000: 62), used in the present reading test were three selected-response formats. The item types were multiple choice, true-false-not given, and matching. The project group had decided on these conventional test formats because they are both not only widely used and accepted and they also facilitate reliable scoring (cf. Derntl 2009: 62).

A more detailed discussion of these three selected-response formats can be found in chapter 2.2.2. Below illustrative examples of the three item formats are given.

Multiple choice format group A (number of correct answers given)

TASK: Multiple Choice Questions Tick the correct option(s) in the following Multiple Choice Tasks. Bear in mind that more than one statement will be correct in some tasks. If more than one statement is correct, the number of correct statements is given in brackets.

2. According to the text, in Sweden there are no fees (2) a) although they have been on the agenda. b) as the government does not want to charge students c) due to disagreement among the political parties. d) due to financial support from the government.

Example 1: Multiple choice item group A (number of correct answers given)

Multiple choice format group B (number of correct answers not given)

TASK: Multiple Choice Questions

Tick the correct option(s) in the following Multiple Choice Tasks. Bear in mind that more than one statement will be correct in some tasks.

2. According to the text, in Sweden there are no fees

a) although they have been on the agenda.

b) as the government does not want to charge students

c) due to disagreement among the political parties.

d) due to financial support from the government.

Example 2: Multiple Choice Item group B (NOCS not given)

True-False-Not Given

TASK: True – False – Not given

Read through the statements 1-9. Are they "true" or "false"? If there is not enough information to answer, choose "not given".

1	The German government has long contemplated tuition fees.	□ True □ False □ Not given

Example 3: True-False-Not Given Item

Matching

TASK 3: Matching

The following statements are summaries of the single paragraphs. Match the most appropriate statement to each of the paragraphs by indicating the letter of the statement next to the number of the paragraph in the grid. There are more statements than paragraphs but match only one statement to each paragraph! One example has already been done for you.

paragraph	**statement**	**paragraph**	**statement**
1		**4**	
2		**5**	
3		**6**	

A	America as model for tuition fees
B	Amounts of tuition fees
C	Are tuition fees a step in the right direction?
D	Depressed about tuiton fees
E	Drastic dropout rates
F	Anxious prospect of increasing tuition fees

G	Negative preview of tuition fees
H	No need for tuition fees

Example 4: Matching

3.1.3.1.2. Test specifications

The third stage of McNamara's (2000: 25ff.) test development guidelines is concerned with test specifications. Brown and Hudson (2002: 88 ff, referring to Popham 1981) mention three different types of descriptors. While the general descriptor focuses on the overall purpose and aim of the test, specific test descriptors center on the skills tested. Item specifications are concerned with the next smaller unit, the item. Thus they are exemplified and defined by a general description, a sample item, prompt attributes, and response attributes.

Once the test and item specifications have been written, they should be evaluated with regard to item quality and content by a group of experts. The revised version is then normally given to item writers, who develop a test on the basis of these test specifications. In the original study, however, the writers of the test specifications were at the same time the actual item writers (Derntl 2009: 75). Thus, it remains the task of the present researcher to find out whether the specifications correspond to the actual reading test; comments will be written in italics (see Table 4).

The project group combined the general and the specific test descriptors in their specifications, otherwise they followed Brown and Hudson's (2002: 87 ff.) guidelines. The general test descriptor of the project group was slightly modified to fit the present reading test, which consisted of only one reading text.

Table 3 General test descriptor

GENERAL TEST DESCRIPTOR (adapted from Derntl 2009: 70 f.)	
subtitle	✓ Test of English Reading Comprehension Competence
description	✓ The test is designed to compare grade 11 students' reading comprehension performance on three different selected-response tasks ✓ The test, a paper-and-pencil test, lasts 25 minutes. ✓ The level of the test corresponds to the B2 level of proficiency of the Common European Framework of Reference according to which students "can read articles and reports in which the writers adopt particular attitudes and viewpoints. [They] can understand contemporary literary prose". (Common European Framework of Reference for Languages: Learning, teaching, assessment, 2001: 27, as referred to in Derntl, 2009: 71). ✓ The test covers the English language skill area of reading and in particular reading comprehension competence.
structure	✓ The test consists of a newspaper article from a quality newspaper. ✓ Examinees have to apply certain reading strategies in order to cope with texts and the responses (e.g. skimming, scanning). The examinee demonstrates mastery of academic reading abilities, such as understanding the core context of a specific text type (newspaper article), being able to detect detailed information as well as to gain a broad overview. The global and detailed understanding will be tested in different [selected] response test formats.

As previously mentioned, item specifications are comprised by three components: a) general description, b) sample item, and c) prompt attributes (Brown & Hudson 2002: 90 ff.). Below the four components of the item specifications are briefly outlined.

a) General descriptor

Table 4 General description

Reading Test Specifications – Newspaper Article

GENERAL DESCRIPTION (cf. Derntl 2009: 71 f.)	
item description	Skill area description: READING Text type: NEWSPAPER ARTICLE A person who masters this reading test is required to demonstrate ability to comprehend advanced non-academic texts. Tasks included here are: ✓ utilizing text for study purposes: - skimming for main idea - scanning for specific information ✓ extensive reading comprehension: - summary of a single text - answering questions according to the text

b) Sample reading test items

can be seen in chapter 3.1.2.1.1. "The test formats".

c) Prompt attributes

Table 5 Prompt attributes / text specifications

PROMPT ATTRIBUTES/ TEXT SPECIFICATIONS	
source	✓ the text is chosen from a quality newspaper from an English speaking country, broadsheets like for example: The Guardian, The Observer, The Times, The New York Times, etc. *[actual reading text taken from The Guardian]*
topic	✓ the text will, in general, be unfamiliar to the test takers but the topic might be more or less familiar depending on the test takers' own interest in the news ✓ the test shall not require any special or former knowledge from the test taker concerning vocabulary on the topic itself ✓ the text shall be controversial so that test takers can produce an opinion-based piece of writing afterwards *[The newspaper article on tuition fees fulfills all these criteria.]*
length	✓ 800-1000 words ✓ the text is presented in an unsimplified version ✓ to shorten the text it is possible to leave out single paragraphs if they do not contain essential information [These criteria are fulfilled as well.]
format	✓ include paragraph numbers (1,2,3, etc.) on the left hand side of the text

For closer discussion of the test specifications of the reading test see Derntl (2009: 70 ff.).

Summing up, test specifications, which are comprised by general descriptors, specific test descriptors, and item descriptors are not only useful for item writers but also for researchers who want to re-use a test in the context of a replication study, or who want to compare various test formats.

3.1.3.1.3. Trialling

The fourth stage of the test development process is concerned with the trialling. McNamara (2000: 23) proposes an appropriate trial population which resembles the actual test takers in age, proficiency, and background. Due to the fact that the present study is a replication study the original test population can be regarded as the trial population, as they corresponded fairly to the replication test population in the important characteristics such as age, proficiency, background.

The original test population were grade 12 test takers in the middle of grade 12, i.e. the test took place in February, while the replication test population were grade 11 test takers at the end of grade 11, i.e. they took the

test at the end of May. According to interviews with teachers 11th and 12th graders' development, the test takers' conditions can be considered as equal, since 12th grade pupils normally do not learn anything new between September and February. Rather do they try to perfect their already existing language skills with regard to the school leaving exam, i.e. the 'Matura', which normally takes place around May.

3.1.3.2. The reading text

Although the difference between end-grade 11 and mid-grade 12 students should not matter, a readability formula was used to determine the level of the newspaper article on tuition fees. In total, the passage consists of 846 words. To determine the readability of the text, the Flesch Reading Ease Formula, was calculated. The Flesch Reading Ease Formula takes into account the average sentence length, as well as the average number of syllables per word. Advantages of the Flesch Reading Ease Formula are that it is one of the most widely recognized readability indices, which are not only used by many US Government Agencies but also by numerous researchers, as for instance Kobayashi (2002). Moreover, the calculation can be done very easily online, at http://www.readabilityformulas.com/free-readability-formula-assessment.php (06 November 2009). A shortcoming, however, is that the formula had been created for texts of English native speakers and not for learners of English as a Second or a Foreign Language (http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.php, 06 November 2009).

According to the Flesch Reading Ease Formula, the newspaper article got 68.14 points. The output is usually a number ranging from 0 to 100. which can be interpreted as follows: scores between 90.0 and 100.00 are considered to be easily understood by an average 5th grade native speaker. Scores between 60.0 and 70.0 should be understood by an average 8th or 9th grade native speaker. Native college graduates, however, should not have any problems in understanding texts with scores between 0.0 and 30.0 (ibid). The present test scored 68.14 points and was used for grade 11 learners of English as a Foreign Language. Taking into account the difference between native

speakers of English, and students of English as a Foreign Language, the choice of this text for grade 11 students seems to be adequate.

Until now only one readability formula is useful for determining the ease of English texts for ESL or EFL readers. This formula is called the McAlpine EFLAW readability formula, and has been developed by Rachel McAlpine in 2004. This formula is based on a different concept: it takes into account two aspects that most commonly trouble foreign language learners of English: long sentences and a high proportion of mini words, i.e. short, common words consisting of one to three letters, which have many meanings as for instance for, of, by (http://www.webpagecontent.com/arc_archive/139/5/ , 06 November 2010). Consequently, the higher the score on McAlpine's formula, the more difficult the text is. Texts with a score from 1 to 20 are considered to be very easy to understand. Scores ranging from 21 to 25 are quite easy to understand. Higher scores from 26 to 29 are supposedly a bit difficult for ESL or EFL readers, while texts with a score higher than 30 are considered to confuse the reader due to the number of long sentences and the large amount of mini words (ibid). The present text scored 17.8 points on McAlpine's scale, accordingly it is supposed to be very easy to understand and should not create any confusions in the EFL reader. The reading text is to be found in appendix 7.2.

3.1.3. The Questionnaire

After having completed the reading test, students were asked to fill in a questionnaire on relevant personal information as well as on the reading test in general. The students were asked to state their mother tongue, the respective native languages of their parents, along with the languages spoken at home. This information was vital in order to find out about any native speakers or semi native speakers in the EFL classes. In the second part of the questionnaire students were asked to give their opinion on this reading comprehension task. Students also stated whether they thought that the newspaper article on tuition fees was interesting or not. Moreover, they were asked to assign a value reaching from 1 (very easy) to 5 (very difficult) on a 5-point Likert scale to each of the test formats and to justify their rating. This was done to ensure that the students had not just randomly assigned points on a 5-point Likert scale to each of the three test formats. Further questions on the questionnaire were related to

the length and the layout of the reading text. The questionnaire is to be found in the appendix 7.2.

While the first chapter of the procedures section was concerned with the materials of the reading test, the second chapter will focus on the setting and the participants.

3.2. Setting and participants

3.2.1. Description of the schools

This replication study was carried out in two Viennese “Gymnasien”. The choice of these two schools was based on their locations in two different districts as regards their social and economic nature. While, what will be henceforth referred to as school A, one of the oldest schools in Vienna which has a very good reputation, is located in a prestigious district, school B is in a not only socially but also economically less privileged part of Vienna where numerous immigrant families live. School A is one of the oldest high schools in Vienna and has a very good reputation.

School A offers its attendants both: traditional values as well as modern future outlooks, i.e. innovative school tracks. School B, however, does not share an equally long history. What is special about Gymnasium and Realgymnasium B is the offer of innovative high school tracks such as a serious sports track, a natural sciences track, a track with an emphasis either on arts or music, and the so called Audio Org, which focuses on information and communication technology. Up until now only one school track is widely known: the serious sports track which is unique to some states in Austria and has been and still is attended by popular Austrian athletes.

Another difference between the two schools is that in school A most pupils’ parents have an academic background and decide to send their children to a school that is not only innovative but also sticks to traditional and cultural values. In school B, however, according to an interview with the headmistress, the number of parents with an academic background is significantly smaller.

For this paper the most relevant difference between the two schools is related to the *Matura* forms they are running. The current *Matura* for modern languages tests two skills: the receptive skill of listening and writing, a

productive skill. Along with selected pilot schools, school A is running the school pilot project "listening neu", in which the listening comprehension part of the *Matura* is standardized. The school pilot project "listening neu" was created in view of the new standardized competence-oriented *Matura* assessing all four skills, which will be implemented in Austrian schools in 2013/2014 (Friedl Lucyshyn 2010: 7). The test formats featured in this standardized listening comprehension part are multiple choice, true-false, matching, grids, and cloze tests. Pupils attending school B, however, still take the traditional Matura with an unstandardized listening comprehension test, with summary writing and short answer question item formats[3]. Accordingly, and as argued by the teachers themselves, those from school A prepare their students for the Matura "listening neu" different to school B teachers who prepare their pupils for the traditional Matura. This difference in preparation might also be reflected in the students' familiarity with the three selected-response tasks of the reading test underlying this study. Although all teachers admitted that their students were familiar with the multiple choice, true-false-not given, and matching format, school A teachers stated that they practiced these formats more often with their students than school B teachers.

Furthermore, the teachers' attitudes towards the new standardized, competence-oriented Matura differed. While all teachers of the participating classes at school A were in favor of the new standardized *Matura*, at school B only one out of three teachers supported the latter. The other two teachers at school B openly expressed their negative views on standardization. This opinion was also reflected in their teaching. According to interviews, school A teachers averred that they focused in their teaching on all of the four skills, probably in view of the new standardized *Matura* which is due to be introduced in 2013/2014, while one teacher at school B admitted that she did not practice all four skills in class. More specifically, she said that she rarely provided her students with reading comprehension exercises. Research question number three will then explore if the different forms of Matura preparations influence

[3] Information on the different Matura forms was gained during teacher interviews, from information material provided by the teachers, as well as from a webpage: *www.durchstarten.at/sixcms/media.php/.../*ahsmaturaenglischvglaltneus5*.pdf,* 18 March 2010.

students' reading comprehension as assessed by multiple choice, true-false-not given, and matching tasks.

Having covered the differences between school A and B, the next section will focus on the different school tracks offered by the respective schools.

3.2.2. Description of the school tracks

Within the two different schools the pupils of three "Gymnasium" classes and of three "Realgymnasium" classes took the reading comprehension test. A "Gymnasium" is a humanistic branch of a high school where pupils are taught Latin as well as up to four modern languages (homepage school A, 4 Sept. 2009). While studying only two foreign languages, pupils attending a Realgymnasium branch receive thorough education either in natural sciences subjects such as descriptive geometry, biology, physics and chemistry, information and communication technology, music or sports (homepage school A; homepage school B 4 Sept. 2009). The two branches, i.e. Gymnasium and Realgymnasium, however, do not differ with regard to the amount of English lessons taught[4].

Within the two different schools and two different branches three different school tracks were tested. At school A one class of a "Gymnasium" branch and one class of a "Realgymnasium" branch, a so-called "notebook" class, focusing on information and communication technology, were tested. The class mentioned first is generally divided into two groups to ensure better and more effective language teaching.

At school B, however, three different tracks within the two branches were tested: two "Realgymnasium" classes and one "Gymnasium" class. The "Gymnasium" class is similar to the one at school A. The tracks of the Realgymnasium, however, are quite different: while one track focuses on natural sciences subjects, the other is a sports track. It has been designed for pupils involved in sports who want to complete their A-levels while training. Instead of 4 years of upper grade education they have to attend 5 years due to a reduction of their weekly school lessons. This reduction enables them to

[4] One subsidiary research question investigating the difference between the reading comprehension performance of Gymnasiums and Realgymnasiums pupils found that there was no significant difference. According to the limited scope of this paper this research question will be disregarded.

attend training sessions not only after but also before school. Due to the fact that the school can only accommodate a limited amount of pupils per year only prospective students who have national or even international career outlooks are admitted to attend this track. Consequently, the sports track is primarily not attended by pupils from the same district but by aspiring athletes from all over Austria, especially from Vienna and Lower Austria[5].

3.2.3. Description of participants

Test takers were grade 11 pupils. In total 49 girls and 48 boys took the reading test, though boys and girls were not equally distributed in each class. Due to the fact that different tracks were tested, some of which are unique in Austria, not all pupils came from Vienna but from all over Austria. Also, test takers differed widely as far as their own and their parents' nationalities, their mother tongues as well as further languages spoken at home were concerned.

Another difference between school A and school B pupils is the amount of time spent in English speaking countries. In school A all classes had already spent some time in English speaking countries in the context of school exchange programs. School B, however, had not taken part in any school exchange programs. Thus, only pupils attending the sports track had either been to English speaking countries or had had contact with English speaking people during Europe-wide or worldwide sports competitions or training camps. All pupils, regardless of the tracks they were attending, had had a similar amount of English lessons. The only differences with regard to English lessons is that sports track pupils are allowed to attend training camps and take part in competitions during the school year, which means that they are absent from school more often than their colleagues from other tracks. When pupils return from their training camps or competitions, however, they are offered additional tuition (homepage school B, 4 Sept. 2009).

Out of the 99 test takers, two could not be taken into consideration. With the help of the questionnaire[6] it turned out that one pupil was a semi-native speaker. Therefore he could not be compared to the other non native speakers.

[5] Information on school tracks was gained in interviews with the English teachers and from the respective school homepages.
[6] see appendix 7.2.

The second pupil whose results had to be dismissed, was a girl from Costa Rica, who spent a high school exchange year in this very class. In her country of origin English is presumably taught in a different way, something which remains unobservable to the researcher. Therefore her results were found to be not comparable to those of Austrian high school pupils.

3.2.4. Description of test administration

The reading test was carried out in school A and B at the end of May 2009. In school A all three classes were tested from 10 am until 11 am. In school B times when the reading test took place varied. While the sports track was tested from 10 am until 11 am, the other two classes of school B were tested between 8 am and 9 am, and between 1 pm and 2 pm. The instructions to the test were all given in German, for two reasons: firstly, to prevent misunderstandings altogether. Secondly, in the original study the instructions were also given in German. The test administrator had written down the instructions that should be given to the pupils so that all test takers took the test under the same conditions.

In the original study the students had 50 minutes to complete the reading test consisting of two different texts. Thus, in the present study students had half the time, i.e. 25 minutes, to read the newspaper article and to answer the questions in multiple choice, true-false-not given, and matching formats. Finally, it turned out that the students were unable to complete the reading test within 25 minutes. Therefore, they were given more time as this test had been constructed as a power and not as a speed test. Further, the interest of the researcher was not to find out whether the students can finish the test within the agreed time but to investigate whether testees perform differently on multiple choice, true-false-not given, and matching tasks. This condition required the students to complete all three reading comprehension tasks.

Test takers were allowed to use dictionaries, similarly to the original study. Answers related to the reading comprehension test were not answered by the test administrator. Further, care was taken to prevent testees from cheating. Generally, no problems occurred during the testing sessions. Subsequently to the reading comprehension test students were asked to fill in

the questionnaire, which is described in section 3.1.3. and which can be found in appendix 7.2.

3.3. Scoring

It goes without saying that the scoring procedure is an essential component of language testing since important decisions are made on the basis of test scores. Since this study was designed as a replication study the same methods used by the original project group had to be applied. With item formats such as multiple-choice, true-false-not given, and matching the scoring is normally easy and unproblematic. This was definitely not the case with multiple-choice items, though. The project group had scored the multiple-choice items in a rather "unconventional manner" (Derntl 2009: 80). Despite the fact that in the domain of language testing there is a general tendency to award one point for each correctly answered multiple-choice question, the project group did not apply this scoring method to their test. As in some of the multiple-choice items more than one answer was correct the original project group felt

> that the test takers should not only get the chance to get one point for the correct answer but that points should also be awarded to partially correct answers. Therefore the project group decided to apply a scoring method which could account for the range or [sic.] possible answers (Derntl 2009: 80 ff.).

Thus, they awarded half a point for each correctly ticked or not-ticked answer of the 4-options multiple choice items. A fully correctly answered question was consequently awarded with 2 points. Sticking to this method implies that students who did not tick any answer at all would at least get 1 point, in the case of 2 correct answers. If only 1 answer was correct students could even get 1.5 points,. This would lead to a distortion of the overall results with regard to this question format. Therefore, the project group decided to change their scoring procedure by marking test takers who did not tick any answer with a NA (no answer) and by giving them 0 points. Nonetheless, there remain some problems for the statistical analysis. For the calculations of the facility value only fully correct answers could be considered, which reduced the facility value of the multiple choice items to a certain extent.

For further studies, however, it would be advisable to assign 0 points for incorrect and partially correct answers and 1 point for fully correct answers. Derntl (2009: 81) argues that

> in order to make the test takers aware of the fact that some multiple-choice items may have more than one correct answer the number of correct answers could be indicated next to each multiple-choice item.

In the present study this very research question will be addressed in chapter 4. In contrast to the complex scoring procedure for multiple choice items, true-false-not given items, and matching items were easy to score since only one answer per item was correct. Thus, each correct answer was simply awarded with 2 points.

3.4. Variables

The construct of the present study is reading proficiency in EFL. A variable can be defined as "the observed or quantifiable representation of a construct" (Brown 1988:8). Generally, researchers distinguish between dependent and independent variables. The dependent variable, often referred to as "the variable of focus" or "the central variable" can be defined as the variable "on which other variables will act if there is any relationship" (Brown 1988: 10). Thus, the dependent variable in the present study is the testees' scores on three different types of reading comprehension tests: multiple choice, true-false-not given, and matching.

Independent variables are selected by the researcher "to determine their effect on or relationship with the dependent variable" (Brown 1988: 10). For the present study various independent variables were selected. Due to the fact that one research question is concerned with any possible gender differences in students' reading comprehension performance as measured by three different selected-response test formats, one of those independent variables is the gender of the test takers. Further independent variables are the different schools the testees are attending, i.e. school A and school B. On the reading test manipulations were made to find answers to the aforementioned research questions. These manipulations lead to further independent variables: students who had the number of correct answers indicated next to each multiple choice

item, and students who lacked that information, as well as students who had the publishing date of the newspaper article given and those who did not.

3.5. Data Analysis Procedures

The present section gives information about the data analysis procedures used to investigate the research questions (cf. *Results*). Chapter 4 is constituted by two main parts: one part is devoted to an analysis of test quality for objectively scored tests, while the second part focuses on the main and subsidiary research questions guiding this paper. Before the various test formats can be compared to each other it is important to analyze them separately according to the underlying criteria of objectively scored tests. Results of this analysis could provide important information on the quality of the different test tasks and on possible differences between them (cf. Hughes 2003: 225). Given the fact that each test format, i.e. multiple choice, true-false-not given, and matching, consists of a different number of items it is necessary to convert the total points that the students scored on each format into percentage points, so that comparisons between the students' results on the three different tasks can be made.

Section 4.1 explains the relevant criteria of test quality and calculates their values with the help of the statistical analysis software SPSS (Statistical Package for the Social Sciences) Version 17. While the facility values (definition see section 4.2.1.1) of the items were computed with the help of the descriptive statistics, the discrimination index (definition see section 4.2.1.2) was calculated with a Pearson correlation for interval scale data, which is an associational inferential statistics examining associations and relationships between two variables (Morgan 2004: 111). The concept of correlation as well as the assumptions, which were found to be met for the present data, are more fully explained in section 4.2. The facility values and discrimination indices for each test format, i.e. multiple choice, true-false-not given, and matching, were calculated manually by averaging the respective facility values and discrimination indices which were calculated with SPSS version 17. The reliability of the test was calculated with the help of Cronbach's alpha, which computes an inter-item consistency (Knoch 2009: 4). Furthermore, an analysis

that aimed at detecting items which particularly lower the reliability of the test was conducted. Both of these analyses were calculated with the help of SPSS version 17. This statistical software is also used for the exploration of the main and subsidiary research questions.

For the research questions guiding this paper (cf. chapter 1.3.) difference inferential statistics and associational inferential statistics will be conducted. While difference inferential statistics are used to compare groups, associational inferential statistics aim at examining associations or relationships between two variables. The main research questions will be examined with the help of difference inferential statistics, as the aim of these calculations is to compare two groups to each other and to find out if there are any differences in the performance of these groups. The statistical test used to compare students' reading comprehension performance on different assessment tasks is the t-test. A t-test is used to calculate differences between two groups. More specifically, the t-test compares the mean, the arithmetic average which is "the statistic of choice" (Morgan 2004: 45) for normally distributed data. This is done by calculating a so-called *t* score and by displaying the probability of the difference between the means. The significance level is set at $p<.05$, which is the most commonly used level in language studies (Brown 1988: 116). This significance level implies that the probability is 5% that the results have arisen by chance alone (ibid). In the present study two different types of t-test are used: independent samples and paired-samples t-tests. While the independent samples t-test is used to calculate the difference between two unrelated and independent groups (e.g. male and female test takers, school A and school B students) on an approximately normal dependent variable (e.g. on multiple choice items), the paired samples t-test is used when the two scores are related, i.e. to compare the students' performance on all three sets of scores on the multiple choice, true-false-not given and matching items (Morgan 2004: 136, 141).

Before the t-tests are conducted, the underlying assumptions have to be tested. For the independent samples t-test the assumptions are: equality of variances of the dependent variable, which is checked by SPSS automatically with the Levene test, normal distribution of the dependent variable, which is tested with the help of the explore command and if the explore command found

that a variable is normally distributed, this can be double-checked by displaying graphs with normal curves in SPSS. The third assumption for the independent samples t-test is that the data is independent, suggesting that the scores of one participant are not related to the scores of the others (Morgan 2004: 136). For the data under review, assumption three is met as the different students are neither matched nor related pairs and there is no reason to believe that one student's score might have influenced another student's. For the paired samples t-test only two assumptions have to be met. The first one is that the dependent variable is normally distributed in the two conditions, which was computed with the help of the SPSS explore command. The second assumption of the paired samples t-test, that the levels of the independent variable are paired or matched in some way, is met as well. One student's scores on the multiple choice, true-false-not given, and the matching part are related and not independent, thus they can be considered as repeated measures of the ability of one student (cf. Morgan 2004: 141).

In order to find out whether students' scores on the multiple choice, true-false-not given, and on the matching items are related associational inferential statistics will be conducted. The associational inferential statistics of choice is the Pearson correlation as the data under review is interval scale. The significance level is again set at $p<.05$. As previously mentioned, the concept of correlation as well as the assumptions, which were found to be met for the present data, are more fully explained in section 4.2. Furthermore, the frequencies command of SPSS Version 17 was used to count the students' ratings of the difficulty of the three selected-response formats on a 5-point Likert scale and for the calculation of the students' rating of their interest in the newspaper article on tuition fees on a 4-point Likert scale.

After having explained the data analysis procedures applied to the present study it is important to notice that statistical significance should not be confused with importance or practical significance (Morgan 2004: 89). If a researcher wishes to make judgments about the importance or practical significance of his findings it is not enough to state the statistical significance, but it is also essential to compute the effect size of the sample, which has to be calculated manually (ibid). The effect size can be defined as an estimate of the strength of the relationship or the magnitude of difference between two

variables (ibid). Effect size measures are generally divided into two types or families: the r family and the d family. While the r family expresses effect sizes in terms of strength of association, the d family centers on magnitude of difference between groups (Morgan 2004: 89). Consequently, if differences between the means of groups are calculated with the t-test, the d family of effect sizes has to be computed to be able to define the effect size. If however a correlation is calculated, the r family is used for the computation of the effect size. In the case of comparisons of means between two groups of subjects the effect size can be computed by subtracting the mean of the second group from the mean of the first group and dividing by the pooled standard deviation of both groups (Morgan 2004: 89). For the calculation of the effect size r, however, the correlation coefficient is already the r value, so no further calculations are needed. The interpretation of the effect size can then be read from the far left column in Cohen's guideline (cf. Table 5 below).

Cohen (1988: 79 f.) offers a guideline on how to interpret the manually calculated effect sizes (cf. Table 5) but at the same time he advises the reader to use it with caution due to its arbitrariness and recommends a context-specific interpretation of the magnitude differences (Cohen 1988: 79 f, In'nami, Koizumi 2009: 231).

Table 6 Interpretation of the strength of a relationship

General interpretation of the strength of a relationship (effect sizes) (Morgan 2004: 91)	**The d Family** **d**	**The r Family** **r**
Much larger than typical	≥1.00	≥.70
Large or larger than typical	.80	.50
Medium or typical	.50	.30
Small or smaller than typical	.20	.10

Taking all aspects into consideration (the results, the effect sizes, the implications of the results), the researcher has to determine the practical significance or as Brown (1988: 122) calls it "meaningfulness" of the results.

4. Results

Chapter 4.1 intends to explain substantial aspects of test quality which are important for the reader's understanding of chapter 4.3, which discusses all these aspects in order to compare the three selected-response test formats: multiple choice, true-false-not given, and matching. Chapter 4.2 presents how far the three selected-response formats correlate, as this is important for the interpretation of the comparison of the three formats.

4.1. Analysis of test quality

Before the various test formats can be compared to each other it is important to analyze them separately according to the underlying criteria of test quality for objectively scored tests. If the underlying test does not fulfill the criteria of test quality it is not suitable for a comparison, as only well constructed tests which are reliable and valid should be conducted and compared. Tests that turn out to be non-reliable cannot be used as on the basis of these test results no inferences on the ability tested can be made (Bachman & Palmer 1996: 20, Derntl 2009: 32). Non-valid tests should not be used either, as these tests do not measure what they are intended to measure (cf. Bachman & Palmer 1996: 21). Thus, items or tests as a whole that turn out to be non-reliable after piloting have to be either thoroughly revised or dismissed.

The analysis of test quality can be divided into 3 parts: descriptive statistics, item analysis, and reliability (Knoch 2009: 3). Descriptive statistics, which are calculated for the test as a whole, or in the case of a comparison of test formats for section totals, provide information about the test as a whole. Reliability again gives information about the test as a whole. The third part of test quality is constituted by the item analysis, which in contrast to the first two parts of test quality does not give information on the test as a whole, but about the individual items. The two essential components of item analysis are the facility value (FV), the discrimination index (DI), and in the case of multiple choice items a distractor analysis.

Classical test theory offers both advantages and disadvantages. One advantage certainly is that the test and individual test items are analyzed statistically and thus give information on the quality of the test as a whole, of

sections of the test, as well as on individual items. However, this advantage entails a disadvantage. The analysis of test quality is always a “sample-based descriptive statistic” (Bachman 2004: 139). This dependency of test statistics on a specific sample group of test takers implies that if the test is administered to a different group of testees, the statistical characteristics of the test might change completely. Another limitation of the classical test theory is that it cannot consider the level of ability of a particular test taker (Bachman 2004: 140), but fortunately item response theory (IRT) in its most popular form for language testers, ‘Rasch measurement’, can account for that. The Rasch analysis of the present test is to be found in section 4.1.2.5.

4.1.1. Descriptive statistics

Descriptive statistics, which are calculated only for the total test score, provide information about the overall performance of the test. Descriptive statistics can be subdivided into measures of central tendency or statistical average and measures of variability or dispersion (Morgan 2004: 45, 46; Henning 1987: 39, 40).

The three most commonly used measures to compute the central tendency of a frequency distribution are the mean, the median and the mode. The mean or arithmetic average is most commonly used, as it is “the statistic of choice” (Morgan 2004: 45) for normally distributed data. The mean can be easily calculated by summing up the individual scores of a distribution and by dividing them by the total number of scores in the distribution (Henning 1987: 39). In the case of a skewed frequency distribution, however, the median provides a better measure of central tendency than the mean. The median can be defined as “the numerical point in the distribution at which half of the obtained scores lie above and half below” (Henning 1987: 39). In contrast to the above mentioned measures of central tendency, the mode, also regarded as the most common category, provides the least precise information about central tendency. The mode is simply the most frequently occurring score (Henning 1987: 40).

While the measures of central tendency, as the name implies, compute the mid-point of a distribution, the measures of variability give information on the spread or dispersion of the scores (Morgan 2004: 46). The most common

used measures of variability are the range and the standard deviation. The range, which can be calculated by subtracting the lowest from the highest score, is the crudest measure of variability. The standard deviation, however, gives a clearer indication of the way the scores are distributed around the mean and is based on the deviation of each score from the mean of all scores (Hughes 2003: 156; Morgan 2004: 46). Due to the limited scope of this thesis measures of central tendency and variability will only be computed for section totals, so that a comparison between the three selected-response formats, multiple choice, true-false-not given, and matching is possible. This analysis is to be found in chapter 4.1.4.3.

4.1.2. Reliability

Reliability, another aspect of the analysis of test quality, can be defined as the "consistency of measurement of individuals by a test" (McNamara 2000: 136). A reliable test should be able to define levels of knowledge or ability among candidates consistently (ibid). Anastasi (1997) goes with her definition more into detail and defines three aspects of reliability:

> [t]he consistency of scores obtained by the same persons when they are reexamined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions (Anastasi 1997: 84).

It is apparent that both definitions focus on the term consistency. To put it differently, if the same group of students took the same test twice their results should be consistent on repeated measurement and should only differ incidentally (Henning 1987: 73). If tested on two different tests with similar test items the students should achieve rather equivalent scores. Furthermore, in a reliable test different examining conditions, as for instance a different setting, should not have an influence on the students' scores. Reliability can be understood as "a measure of accuracy, consistency, dependability, or fairness of scores resulting from administration of a particular examination" (Henning 1987: 74). With the help of a reliability coefficient the degree of reliability can be calculated. Reliability coefficients are to be found between the two extremes of 1 and 0. A reliability coefficient of 1 suggest a perfectly reliable test in which students would receive exactly the same score on repeated measurement. A

reliability coefficient of 0, however, would imply that the score that students get when doing a specific test for the first time would be no help at all in predicting their score if they took the same test again on a different day (Hughes 2003: 32).

In the domain of language testing the three most commonly used ways to estimate the reliability are the test-retest method, the split-half method as well as Cronbach's alpha. Due to the fact that in this very study only one test was carried out a test-retest method cannot be carried out. For practical purposes, Cronbach's alpha was preferred to the split-half method. Cronbach's alpha is a very commonly used measure of reliability in the research literature. It is used to estimate the internal consistency reliability of several items or scores which are added together to get a summary score (Morgan 2004: 122). According to Brown (1988: 99) the obvious advantage of internal-consistency estimates is that they can be calculated from a single form of a test administered only once. Alpha values range from 0 to +1.0. Morgan (2004: 122) argues that alpha should be "greater than .70 in order to provide good support for internal consistency reliability". But this of course depends on the purpose of the test. For a high stakes test for instance, a reliability of .9, an almost perfect reliability, could be demanded. In the present reading test Cronbach's alpha was .72, which points to a satisfying overall reliability of the test.

Additionally to the alpha calculation, an analysis was carried out in order to find out if there are any items which particularly lower the reliability of the reading test. If the overall reliability of a test rises after the deletion of an item, this hints at the fact that there could be a problem with this very item. In the present study, however, no problematic items, whose deletion would lead to an increase in the overall reliability, could be found.

4.1.3. Validity

This section aims at analyzing the reading test in terms of validity. To begin with the different types of validity are explained and statements about whether these types of validity are met in the present reading comprehension test will be given. Validity is another important characteristic of tests in general, and language tests in particular (Elamparo 2005: 11). Validity can be defined as the degree to which a test actually measures what it is intended to measure

(Bachman and Palmer 1996: 21, Kitao 1999: 11). Due to the limited scope of this paper only the most important aspects of validity can be briefly explained. For a closer account of validity see chapter 4 in Hughes (2003).

Validity is commonly divided into four different subtypes, although the number of subtypes differs among researchers (cf. Bachman & Palmer 1996, Hughes 2003, Kitao 1999, McNamara 2000). Generally speaking, there are four types of test validity: content, criterion-related, construct, and face validity (Hughes 2003: 26 ff.). Content validity determines whether a test measures the content it is intended to measure (Elamparo 2005: 11). An examination of content validity, which is also referred to as conceptual and non-statistical validity (Davies et al. 1999: 34) involves looking at the specifications, and in the case of an achievement test at the syllabus. According to Hughes (2003: 26)

> a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned.

Hughes (2003: 27) argues, that content validation should be carried out during test development, to ensure that the test specifications are reflected in the test content. Derntl (2009: 25) and Schweinberger (2009: 151), two members of the project team who developed the reading test, comment on the test development process and argue that the test possesses content validity. The present researcher also found that the test content of the reading comprehension test reflects the test specifications (cf. table 2).

The second type of validity, criterion-related validity, determines

> the degree to which results in the test agree with those provided by some independent and highly dependable assessment of the candidate's ability. This independent assessment is thus the criterion measure against which the test is validated (Hughes 2003: 27).

In contrast to content validity, criterion-related validity can be determined statistically by correlating the test with its criterion (Davies et al. 1999: 39). Criterion-related validity incorporates two types of validity: concurrent and predictive validity (ibid). While concurrent validity refers "to the degree to which a test correlates with other tests that test the same thing" (Kitao 1999: 13), predictive validity means the extent to which a test can predict future performance. For the present test it is impossible to determine its criterion

validity, as there is no similar test containing only three selected-response test formats at hand with which the present test could be correlated. This view is also shared by Schweinberger (2009: 151) who among others developed the present reading test. Furthermore, it is also too early to determine the predictive validity of the reading test. If teachers used similar reading tests subsequently to the present reading test, they could investigate the predictive validity of the reading comprehension test.

The third type of validity, construct validity, indicates the extent to which a test's content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned" (Hughes 2003: 26). Examples of these constructs are, for instance, reading ability or fluency of speaking. Similarly to the criterion-validity, construct validity can be statistically tested with the help of factor analysis or multi-trait multi-method analysis (Davies et al. 1999: 33). Due to the limited scope of this paper the construct validity of the reading test could not be statistically investigated. For a closer account of construct validity see Hughes (2003: 26 ff), Bachman & Palmer (1996: 21 f).

The fourth type of validity, face validity refers to

> the degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer, as for instance the candidate taking the test (Davies et al. 1999: 59).

Face validity can also be explained as the impression a test makes on the testees as to what skills are required (Kitao 1999: 12). As the testees are no experts in testing, face validity is often referred to as non-empirical judgment without any theoretical basis (Derntl 2009: 30). Nevertheless, it is useful to rely on the opinion of the test participants, because if they do not regard the test as valid they might not be motivated to take it. In the context of the present reading comprehension test, the testees agreed that the reading comprehension test possesses face validity, as it undoubtedly tests reading comprehension.

Summing up, the reading comprehension test fulfilled those aspects of validity that could be examined within the present study. For a closer examination of the aspects of validity of the present test, which was one part of the original test assessing all four skills, see Derntl (2009: 24 ff.) and Schweinberger (2009: 151 f.).

4.2. Item Analysis

4.2.1. Classical Item Analysis

Traditionally, the classical item analysis consists of three components: the item facility, also called facility value (FV) or facility index, the item discrimination, which is in the research literature most commonly referred to as discrimination index, and the analysis of the distractors of multiple-choice items. The aim of this analysis, which should always be employed and not only be carried out when there are problems with the reliability of the test, is to give information about the quality of the individual test items (Knoch 2009: 3). While the facility value (FV) measures the level of difficulty of an item, the discrimination index (DI) indicates "the extent to which the results of an individual item correlate with results from the whole test" (Alderson et al. 1995: 80) and "to what extent high-scorers on the test as a whole did better on that item than low-scorers" (Baker 1989: 51). In the case of multiple-choice items it is useful to examine the distribution between the correct answers and the incorrect alternatives, known as distractors, which in the research literature is referred to as distractor analysis (ibid).

To be able to judge the quality of the test, all three aspects comprising the analysis of test quality have to be considered. In the following subchapters the concepts behind the three aspects constituting test quality of objective tests will be outlined. Section 4.1.2.4 then discusses all three aspects in a comparative way and suggests which items have to be dismissed, as well as which item format contains the highest number of items which adversely affect the quality of the test.

4.2.1.1. Facility Value (FV)

The facility value, abbreviated as FV, measures the difficulty of an item and thus "helps us decide if test items are at the right level for the target group" (McNamara 2000: 60). To put it differently, the item facility expresses the proportion of students who answer an item correctly. Facility values are usually expressed on a scale from 0 to 1. Generally speaking, a facility value ranging from .33 to .67 is desirable (McNamara 2000: 61, Knoch 2009: 5). A higher value than .67 means that the item is too easy, as 67% of all the test takers

answer it correctly, whereas a FV lower than .33 signifies that the item is too difficult. According to Henning (1987: 49) “item difficulty is most appropriate when it approaches the mid-point of the difficulty range”, which is at .50, suggesting that 50% of the candidates answer the item correctly. FVs of 1, meaning that everybody got it right, and 0, implying that no one answered the item correctly, however, give no information about the difficulty of an item at all. Nonetheless, the desired FV depends largely on the purpose of a test. If a test aims at making distinctions between candidates’ performances, the items should be neither too easy nor too difficult, and a FV of around .50 would be ideal (McNamara 2000: 61) and *appropriate* (Henning 1987: 49). In the case of a high stakes test, however, a rather low facility value ought to be sought.

A generally desirable effect is that items at the beginning of a test are easier so as to help students get accustomed to the testing situation, and to help them ease their tension. This item arrangement according to increased difficulty has a motivational effect on students and “will prevent them from getting bogged down by difficult items early in the test” (Gronlund 2003: 52), which would otherwise rather discourage students. However, if the aim of a test is to distinguish between the most able candidates, it would make sense to include some hard items with a low facility value at the end of a test.

4.2.1.2. Discrimination Index (DI)

The discrimination index also called item discrimination or “item-test correlation” (Hughes 2003: 160) “allows us to see if individual items are providing information about the candidates’ abilities consistent with that provided by the other items on a test” (McNamara 2000: 60). For the calculation of the discrimination index the performances on each item by different groups of test takers are compared: namely, those who have done well on the test altogether, and those who have done rather poorly (McNamara: 2000: 61). The underlying assumption of the discrimination index is that “the people who do best on the whole test should do best on any particular item” (Hughes 2003: 160). An item, which the weakest students get right, but the strongest students get wrong, has a poor item discrimination index. Thus, this item is clearly problematic and needs investigation, as the resulting “scores […] are misleading, and not reliable indicators of the underlying abilities of the candidate” (McNamara 2000:

61). The discrimination index is strongly related to the reliability of the test, which is "the overall capacity of […] a test […] to define levels of knowledge or ability among candidates consistently" (ibid). To put it differently, the more discriminating the items in a test, the more reliable the test is as a whole (Knoch 2009: 5).

The emerging results from a DI calculation can range from –1 to +1. A general rule of thumb is the higher the DI, the better the item discriminates between weak and strong students. DIs of +.3 and higher are generally accepted as okay, though it is important to look at the relative sizes of the indices. A DI of +.5 discriminates well, as "the high scoring students answer it better than the low scoring ones" (Alderson 1995: 82). The highest possible discrimination is +1, which suggests that all of the strong students get an answer correct, while none of the weak students do (ibid). A discrimination index of 0, however, implies that there is no discrimination between weak and strong students, as weak and strong students perform the same. The highest possible negative DI of -1 is achieved when all of the weak students get the item correct, while none of the strong students do, which puts the test result into question as the students' results on the items are confusing, non-interpretable and non-reliable. Baker (1989: 52) even gives an example of how a negative index of discrimination could emerge:

> This may happen if the item contains some 'trap' which more advanced learners fall into while the weaker ones, in blissful ignorance, avoid.

Knoch (2009: 5) and Alderson (1995: 82) argue that low discrimination indices of multiple choice items could be explained by the fact that one or more distractors are 'not working'. Non-working distractors, which can be defined as distractors that are chosen by very few or no candidates, as well as distractors that appeal too much either to all test-takers or just to the stronger students "make no contribution to test reliability" (Hughes 2003: 228). If distractors are not appealing at all this could imply that all students, regardless of their proficiency levels, choose the right answer which would in turn lead to a high FV but to a DI close to 0, implying that strong and weak students perform the same. However, if distractors are too appealing to all test takers, thus preventing them to tick the correct answer, the result would be not only a low FV but also a low DI or a DI of 0, which again means that both highly and low

proficient students perform the same. In the case of distractors that are not appealing to low scoring students but too appealing to high scoring students the result would be a negative DI, saying that weak students tend to perform better than strong students.

4.2.1.3. Analysis of the distractors

In the analysis of a multiple choice test we are not only interested in the frequency with which the correct option is chosen, i.e. the measure of item difficulty, but also in the frequency of the selection of incorrect, so-called distractor options (Henning 1987: 55). As previously mentioned, the most difficult part in designing a multiple choice item is the construction of well-functioning distractors. Thus, it is essential to analyze the distractors so as to be able to make judgments about their function. Interestingly, none of the language testing handbooks gives any suggestions on the percentage of test takers a well functioning distractor should attract, but on the other hand it would be a rather thorny issue to establish any benchmarks. When judging the function of a distractor it is indispensable to consider the relative discrimination indices as well as the facility values. Distractors that for instance attract 20% of the participants, though only the stronger ones, while none of the weaker students chooses them, do obviously not contribute to the discriminating function of an item, they rather challenge its reliability as weak students tend to perform better than stronger ones. If there is however a distractor, that is chosen by more students, who are all in the bottom-group while none of the top-group students choose it, this distractor could be labeled as functioning as it discriminates well between strong and weak test takers.

4.2.1.4. Discussion of all three components of classical item analysis

In this very section of the paper all three components of classical item analysis, the facility value, the discrimination index as well as in the case of multiple-choice items, a distractor analysis, will be discussed so that comparisons can be drawn from the three selected-response formats further on in this thesis. Due to the fact that a discussion of each individual item of the reading comprehension test would go beyond the scope of this paper, only items which are problematic in terms of their facility values, discrimination indices, and in the

case of multiple-choice items have non-working distractors, can be discussed in detail. These items are depicted in table 6. Tables indicating the FVs and DIs of all items are to be found in appendix 7.3. For the calculation of the FVs and the DIs only fully correct and no partially correct answers could be considered, which lead to a distortion of the FVs and DIs of the multiple choice items. Therefore, only items which are not slightly but truly beyond the acceptable ranges will be considered.

Table 7 Facility values and discrimination indices of problematic multiple choice (mc), true-false-not given (tfng) and matching (m) items.

	FV	**DI**
mctask1_5	**.20**	**-.04**
tfngtask2_2	.43	**.04**
tfngtask2_3	**.24**	**.17**
tfngtask2_5	**.24**	**.15**
tfngtask2_9	.35	**.13**
mtask3_8	**.16**	**.24**
mtask3_9	**.32**	**.26**
mtask3_10	**.24**	**.23**

According to this table only one multiple choice item, but four true-false-not given, and three matching items reached facility values and discrimination indices beyond the acceptable ranges. Multiple choice item number 5 appears to be a problematic item. The low and negative discrimination index of this item could hint at the fact that one or more distractors are not working. After a distractor analysis, which is depicted in table 7, this proved to be the case.

Table 8 Distractor analysis multiple-choice task 1 item 5

multiple-choice task 1 item 5	
chosen answer	**valid percent**
a*	24,45
b	17,55
c	5,95
d	49,25
no answer	3,1

While distractor d appealed to 49,25% of the students, the right answer a (marked with an asterisk) was only chosen by 24,45% of the test takers. If the 24,45% of test takers were the smartest students of the target group this would not matter too much, but the negative DI of -.04 suggests that it is not the top group students but rather the bottom group students who chose the correct answer a. Nevertheless, this DI is only slightly below a DI of 0 which does not discriminate at all between strong and weak students. Therefore, item 5 of the multiple-choice part of the test has to be either removed or revised as it questions the reliability of the item.

Item 2 of the true-false-not given has an acceptable facility value of .43, but it does not discriminate well between weak and strong students, as the bottom and the top group of test takers almost score the same (DI= .04). As this item produces non-reliable results it should be either revised or left out from the test. The same is true for item 9, which does not discriminate well between strong and weak students either, according to a DI of .13. Further items that are prone to problems according to their facility values and discrimination indices are items 3 and 5. Summing up, true-false-not given items 2, 3, 5 and 9 should be either rewritten or excluded from the test.

Considering the matching format, items 8,9 and 10 have to either be revised or rewritten as they are problematic in terms of their facility values and discrimination indices. More specifically, all three matching items turn out to be too difficult with FVs ranging from .16 to .32 and too little discriminating, as all three discrimination indices are with values from .23 to .26 below the acceptable DI of +.3.

Summing up, many items reached both, acceptable FVs and DIs (cf. appendix 7.3). Those items that have been pointed out in this analysis, though, should definitely be either thoroughly revised or eliminated from the test, as they would contribute to a non-consistent measurement. The test format which contains the highest number of items which need revision is the true-false-not given format, as here 44% of all items should be revised. In the matching, and the multiple choice format, however, only 23% and 14% of the items reached inacceptable facility values and discrimination indices. When looking at facility values and discrimination indices, it is yet important to acknowledge that the criterion for the inclusion or exclusion of items should not be entirely statistical

(Baker 1989: 54). Stakeholders as teachers might decide to retain certain items that that test an aspect of knowledge stakeholders need information about in the test, although they might have been labeled as too easy, too difficult, or as too little discriminating according to psychometric criteria.

4.2.2. Item response theory

In recent decades new methods of analysis, with many attractive aspects for people involved in testing, evolved. They are subsumed under the heading of item response theory (IRT), with the Rasch analysis being the most frequently used form in language testing (Hughes 2003: 228). Rasch analysis is based on the following assumptions: firstly, test items have a particular difficulty attached to them according to which they can be rank-ordered, and secondly, each test taker has a fixed level of ability (ibid). In contrast to classical test theory, which only takes into account the behavior of items, the Rasch model can predict both: the behavior of test takers and test items. The model accepts the fact that people's test performance will not be a perfect reflection of their actual ability, but "it does draw attention to test performance which is significantly different from what the model would predict" (Hughes 2003: 228). Thus, it identifies both test takers whose behavior does not fit the model, and test items that do not fit the model (ibid).

In the analysis solely misfitting items can be considered, as an analysis of misfitting test takers is only possible in cooperation with the latter, which was not possible within the scope of the present study. For a classroom teacher, however, it would be utterly important to focus not only on misfitting items but also on the test takers as these people cannot be assessed properly by the test which means that other, additional methods of assessment should be sought.

The index which indicates items that do not fit the model is known as the "fit". The fit index suggests how well or badly the items fit the Rasch model (Hughes 2003: 230). "The higher the positive value, the less well the item fits" (ibid). A misfitting item implies that there is an inconsistency in answers, i.e. while weaker test takers answer that item correctly, stronger test takers tend to answer it wrongly. Reasons for this misfit may be found either in the test takers (guessing, lack of concentration, tired, mood, cheating, etc.) or test items could have been badly designed which requires a revision of those misfitting items. In

the present study items with fit values ranging from .74 to 1.26 were considered as fitting[7], while items out of this range were considered as misfitting.

In the present reading test Rasch analysis was used for two reasons: firstly, to find out if there are any misfitting items, and secondly, to determine which selected-response format, i.e. multiple choice, true-false-not given, matching, the misfitting items belong to. In the reading test only two misfitting items could be detected. Both of these items were multiple choice formats. More specifically, items 1 and 5 of the multiple choice format were identified as misfitting, due to their high fit values of 1.28 and 1.44, respectively. The items can be seen in the appendix (cf. section 7.4). As previously mentioned, items with high fit values indicate that test takers' performance on these items is significantly different from what the model would predict. Test takers who are overall more able on the test tend to answer these items incorrectly, while less able test takers tend to answer them incorrectly. This inconsistency in answers is considered negative for the overall reliability of the test. At the same time, the reason for a high fit value could be attributable not to poor item construction but to guessing on part of the testees, which consequently leads to a deviation from their normal behavior.

Interestingly, within classical item analysis, which was explained in the previous section, item 1 of the multiple choice part proved to have a slightly too low discrimination index (DI= .23), but the facility value was within the acceptable range. Therefore, after the classical item analysis this item was not dismissed. After the application of Rasch analysis, however, item 1 of the multiple choice part was identified as misfitting due to its high fit value. This example highlights that for a thorough analysis it is not enough to rely on classical item analysis alone.

4.3. Correlation between the three selected-response formats

One of the main assumptions underlying this research study is that there exists a relationship between students' scores on the multiple choice, true-false-not given, and matching formats, as these formats assess the same underlying

[7] The fit range can be calculated by adding two times the standard deviation to the infit mean square (McNamara 1996:181). For a closer discussion of the fit range see McNamara (1996) chapter five.

ability: reading comprehension. The statistic that is used to test this assumption is the correlation. A correlation can be defined as the “relationship between two entities – constructs or variables – that can vary in terms of its strength and direction” (Bachman 2004: 84). It only makes sense to carry out a correlation if a researcher has a plausible reason to suppose that one variable will be related to another in a particular way, which is the case in the present study. The degree to which two constructs or variables are related can be estimated by calculating a correlation coefficient (Brown 1988: 96).

Correlation coefficients, “mathematical measures of similarity” (Hughes 2003: 28) take values ranging form 0 to +1.0 or -1.0. Positive coefficients indicate direct relationships, i.e. students who score high on the multiple choice part of the test also score high on the true-false-not given items. Negative correlation coefficients, however, suggest an inverse relationship, i.e. students who score high on the multiple choice part of the test score low on the true-false-not given tasks (Bachman 2004: 89). The higher the positive or negative value the stronger is the relationship between the two variables. A correlation coefficient of 0 suggests that there is no relationship between two sets of variables (Brown 1988: 97). Brown (ibid) offers some guidelines on how to interpret correlation coefficients. He considers correlation coefficients up to +/-.40 as representing weak correlations. Correlation coefficients ranging from +/-.80 to +/- 1.0, however, indicate strong ones. Cohen (1988: 79 ff.) interprets correlation coefficients differently according to the r family of effect size measures. While he regards correlation coefficients up to +/-.10 as representing small effects, correlation coefficients ranging to +/-.30 indicate medium or typical effect sizes. Coefficients amounting to +/-.50 are considered as representing large effect sizes, while coefficients larger than +/-.70 indicate effect sizes that are much larger than typical.

Due to the fact that the present data is based on an interval scale, the Pearson correlation is appropriate. In the case of ordinal data as ranks, however, the Spearman correlation would be preferred (Bachman 2004: 87). After the assumptions for the Pearson correlations (linear relationship, scores on one variable are normally distributed for each value of the other variable and

vice versa)[8] were found to be met, a correlation was calculated. Although all correlations were found to be significant at the alpha level of 0.01, the respective correlations are not very strong. Multiple choice items correlate according to Brown weakly, and according to Cohen (1988: 79 ff.) to a medium or typical extent with true-false-not given items, which is indicated by a correlation coefficient of r=.427. Another typical (Cohen) correlation is found between multiple choice and matching items, r=.411. The weakest correlation, according to Brown (1988: 79 ff.), exists between matching and true-false-not given items, r=.277, suggesting that students who score well on the matching part do not necessarily achieve good results on the true-false-not given part and vice versa. Following Cohen's guidelines, however, this correlation can be regarded to have a medium effect size.

Finally, the researcher decided to rely on Cohen's interpretation of the effect size r of the correlation coefficient, as these measures are subdivided into more levels than Brown's guidelines, who only mentions two benchmarks. Summing up, the three test formats, which do not only represent the same item type, i.e. selected-response format, but also assess the same underlying ability, reading comprehension, correlate only to a medium extent and not to a large extent. This question will be answered in the discussion part of this paper (chapter 5), which puts the findings in context and attempts to explain why the results turned out the way they did.

4.4. Comparison of multiple choice, true-false-not given, and matching formats

4.4.1. Overall comparison

After having established the underlying concepts that are important for the understanding of the analysis of test quality, the present chapter is dedicated to an overall comparison of the various test formats: multiple-choice, true-false-not given, and matching. This comparative analysis centers on descriptive statistics, which quantify the distribution by giving information about the central tendency and on the dispersion of scores, both of which are essential pieces of information when test formats should be compared. Furthermore, a discussion

[8] For a closer discussion of the assumptions and conditions for the Pearson correlation see Morgan (2004: 111).

of the facility values and discrimination indices of each of the three selected-response formats, which have been calculated manually by averaging the respective FVs and DIs, will be included. Additionally, a paired samples t-test will be conducted to find out if there are significant differences between the three selected-response test formats. A paired-samples t-test is the statistic of choice as in this case a within-subjects design was used, i.e. all students completed all three test formats (cf. Morgan 2004: 140). Furthermore, all three sets of scores (on the multiple choice, the true-false-not given, and the matching items) come from the same group of subjects, which means that the scores are not independent (cf. Brown 1988: 166). The performance of the test takers will be discussed for each test format separately, disregarding the different schools, school tracks, genders, and other distinctions that were made to find answers to the research hypotheses. It is in the subsequent chapters that the research questions and specific findings in relation to the research questions will be introduced and discussed.

Table 9 Descriptive statistics

		multiple choice percentagetotalt ask1 / 14	true-false-not given percentagetotalt ask2 / 18	matching percentagetotalt ask3 / 26
N	valid	97	97	97
	missing	0	0	0
Mean		,6760	,4204	,3775
Standard Error Mean		,01304	,01942	,02136
Median		,6786	,4444	,3846
Mode		,61	,44	,38
Standard Deviation		,12838	,19126	,21039
Range		,71	,78	,92
Minimum		,29	,00	,00
Maximum		1,00	,78	,92

Figure 1 Mean differences test formats

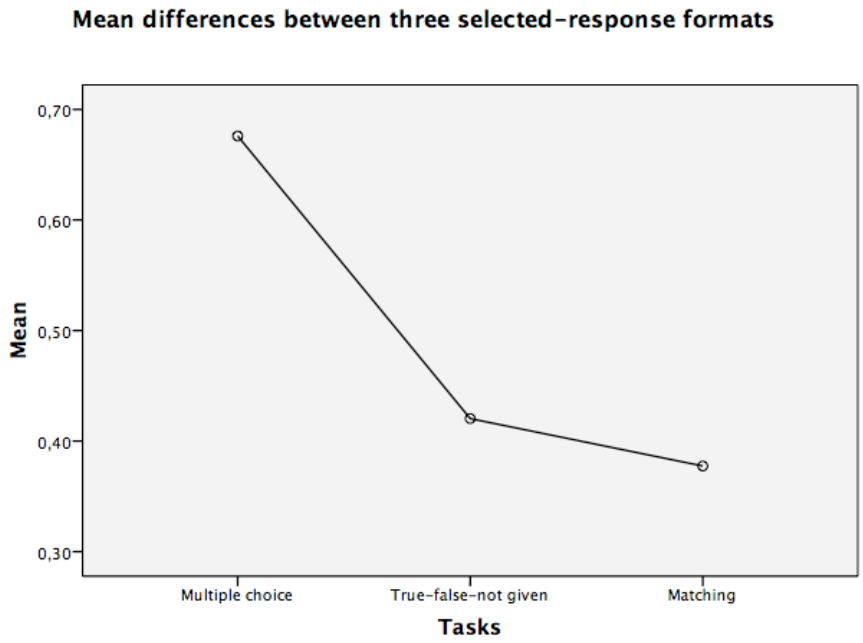


Table 8 and figure 1 illustrate that the highest mean score is achieved on the multiple-choice part of the test. While students get about 68% of all multiple-choice answers right, they can only answer 38% of the questions in the matching format correctly, which therefore accounts for the test format where the students achieve the lowest mean. On the true-false-not given format the test takers on average answer 42% of the questions correctly.

The multiple-choice format does not only account for the highest mean but also for the smallest standard deviation (SD) of 13%, which means that the students scores are not spread out too far on the score range. The matching format, again, is the format on which the students' scores are spread out furthest, which can be seen in a SD of 21%. On the true-false-not given format, students' scores are spread around 19% on each side of the mean. A high standard deviation implies that the students' scores are spread out a lot, which could indicate that the test discriminates strongly between weak and strong students (Baker 1989: 45). A calculation of the discrimination indices of each question format, proved that this is the case as the question format with the highest standard deviation also proved to have the highest DI (.38) as

compared to the multiple-choice (DI = .31) and true-false-not given format (DI= .27).

Table 10 Facility values and discrimination indices

Item format	Facility Value (FV)	Discrimination Index (DI)
multiple-choice	0.31	0.31
true-false-not given	0.42	0.27
matching	0.37	0.38
reading test total	0.44	0.33

When looking at the FVs of the multiple-choice task it is important to keep in mind that this FV might have been distorted as only totally and no partially correct answers could be incorporated in the calculation. Taking into consideration that students achieve the highest mean on the multiple choice items, the FV 0.31 cannot be taken too seriously as partially correct answers were not considered. Therefore, it can be argued that not the multiple choice but the matching format is the most difficult item format with a FV of .37, followed by the true-false-not given format which has a slightly higher FV of .42, signifying that 42% of the test takes answer the items of the true-false-not given format correctly. According to the respective discrimination indices, the matching format, with a DI of .38, discriminates better between bottom group and top group students than the multiple-choice (DI .31) and the true-false-not given format (DI .27).

When looking at the minimum and maximum scores of the respective formats, it is again apparent that the multiple-choice format is the easiest one, as not only the minimum score but also the maximum score is higher than on the other question formats. While in the true-false-not given and in the matching question format some test takers did not even get a single item correct, the minimum score that students achieve on the multiple-choice part of the test is 29% out of 14 points. This result might allude to the fact that students can get a multiple choice item correct by merely guessing, although a counter argument in this case might be that in the true-false-not given question format the guessing factor should not be underrated either. Another fact that highlights that the multiple-choice format is easier than the other techniques is that the

maximum score on this part of the reading test is 100%, i.e. 14 points out of 14, while on the true-false-not given and on the matching format it is 78% and 92%, respectively. These results again emphasize that the facility value of the multiple choice format, which is lower than the facility values of the other item formats (cf. table 9), cannot be taken too seriously as partially correct answers could not be included.

Additionally, an analysis of contrasts between pairs of means is computed to find out how big the differences between the students' results on the three different selected-response formats are. This is done with the help of a paired samples t-test of the null hypothesis that students' mean scores on the multiple-choice, true-false-not given, and matching format would be equal at the 0.5 level.

Table 11 Paired-samples t-test

Paired Samples Statistics

		mean (%)	N	Std. Deviation	Std. Error Mean
Pair 1	Multiple Choice	67.60	97	,12838	,01304
	True-False-Not Given	42.04	97	,19126	,01942
Pair 2	True-False-Not Given	42.04	97	,19126	,01942
	Matching	37.75	97	,21039	,02136
Pair 3	Multiple Choice	67.60	97	,12838	,01304
	Matching	37.75	97	,21039	,02136

Table 12 Significant contrasts between means on different test formats

contrasts	t-value	df	Sig. (p)[9]	effect size d
Multiple-choice/ True-False-Not given	14.049	96	.000*	1.625
True-False-Not given / Matching	1.746	96	.084	
Multiple-choice / Matching	14.980	96	.000*	1.76

Significant differences at the .05 level are found between the multiple-choice and the true-false-not given format, thus the null hypothesis of no difference between the multiple-choice and the true-false-not given question format can be rejected. As can be seen from table 10. students achieve significantly better results on the multiple-choice than on the true-false-not given items. The effect size of the difference between these two test formats is with a *d* of 1.625, following Cohen's (1988) guidelines, is much larger than typical.

Between the true-false-not given and the matching part of the reading test the differences are found to be not significant, thus implying that students achieve rather similar results on those parts of the test.

When comparing the students' results on the multiple-choice and on the matching question format, the paired samples t-test indicates that students on average achieve significantly better results on the multiple-choice than on the matching part of the test (see table 11). Thus, the null hypothesis of no difference between the multiple-choice and the matching question format can be rejected. The difference between these two formats is according to Cohen's (1988) guidelines with a *d* of 1.76 much larger than typical. Summing up, the biggest difference in question formats exists between the multiple-choice (students' mean: 67.6%) and the matching (37.7%) part of the test. This is also shown by the effect size d of 1.76, which is not only much larger than typical but also larger than the effect size of the difference between the multiple-choice and the true-false-not given format (1.62). The difference between the true-false-not given and the matching format, however, is not statistically significant,

[9] Significant p-values are marked with an asterisk.

suggesting that students' performance on these formats does not differ.

4.4.2. Comparison of the three selected-response formats according to students' proficiency level

After having discovered that students achieve very different results on the three different selected-response formats, a really interesting question to explore is whether the assessment format has different or similar effects on students of different proficiency levels. Do high-proficiency level students react less sensitively to different test formats than low-proficiency level students?

For the study under review the students only did one reading test, which thus served as the sole instance where information about the students' proficiency levels could be gathered. This test design does not allow regression analysis, but what can be found out is on which question format those students who achieved good and bad results on the reading test altogether scored best and worst. Therefore, the test takers were divided into two groups according to their scores on the reading test: subjects whose scores ranged from 0 to 29 points, i.e. who achieved less than 50% of the correct answers, were labeled as "low proficient", whereas test takers who achieved more than 29 points formed the "high proficiency group". Out of the 97 test takers, 43 were included in the high proficiency, and 54 in the low proficiency group. After the test takers had been ranked and the underlying assumptions of the t-test were found to be met, independent samples t-tests could be conducted.

Table 12 as well as graphs 2 and 3 show the mean scores of the high and low proficiency level students on the three assessment types.

Table 13 Comparison of mean scores by high and low proficiency students

	high proficiency level	**low proficiency level**
test format	**mean (%)**	**mean (%)**
multiple-choice	.7674	.6032
true-false-not given	.5452	.3210
matching	.5420	.2464

Figure 2 Mean differences high proficiency group

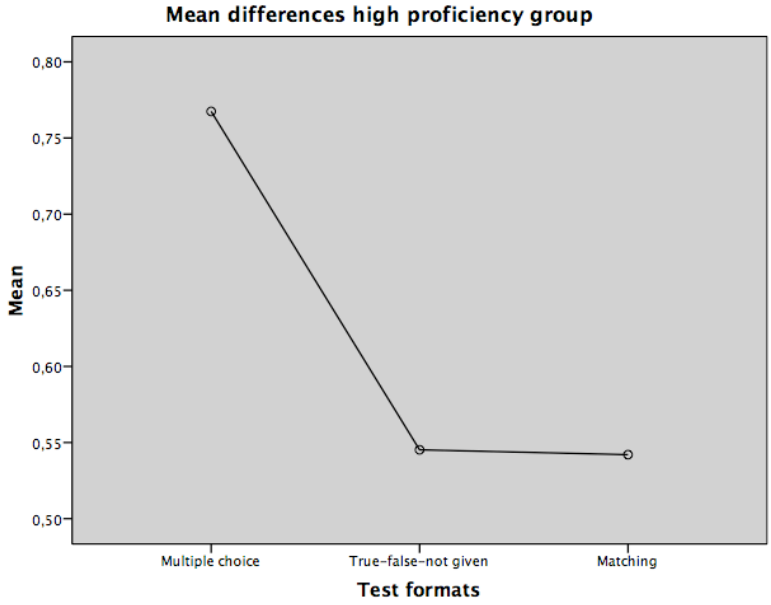
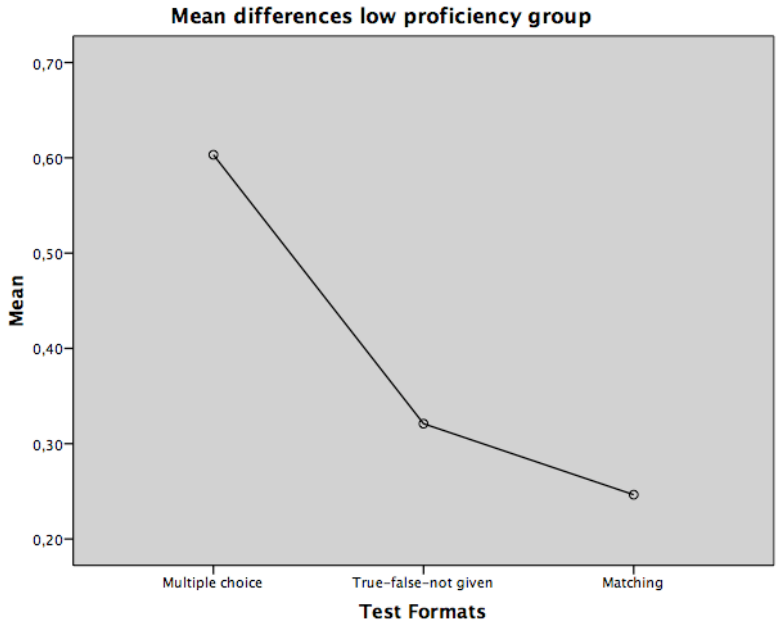


Figure 3 Mean differences low proficiency group



Figures 2 and 3 as well as table 11 highlight that the multiple choice format turns out to be by far the easiest test task, regardless of the test takers' proficiency levels. While high-scoring students achieved a mean of 77%, low-scoring students got 60% of the multiple-choice format correct. For the low-

scoring students, the matching task clearly qualified as the most difficult format, as they only got a mean of .25. High-scoring students, however, achieved rather similar means on the true-false-not given and the matching part of the test, accounting for .545 and .542, respectively.

Table 14 Contrasts between means on different test formats[10]

contrasts	t-values	df	Sig. (p)	effect size d
high proficiency mc / t-f-ng	9.617	42	.000*	2.00
high proficiency t-f-ng /matching	.093	42	.926	
high proficiency mc / matching	8.456	42	.000*	2.09
low proficiency mc / t-f-ng	10.591	53	.000*	2.00
low proficiency t-f-ng / matching	2.169	53	.035*	0.44
low proficiency mc / matching	13.489	53	.000*	2.7

An analysis of contrast, which was carried out with the help of a paired-samples t-test for each proficiency group separately, proved that in the low proficiency group the differences between all three formats were significant (cf. table 11). Nevertheless, the effect sizes differed. While the effect sizes of the differences between multiple choice and true-false-not given tasks, multiple choice and matching items were with d values of 2.00 and 2.7 according to Cohen's effect sizes (cf. Morgan 2004: 91) much larger than typical, the effect size of the difference between true-false-not given and matching items was small. In the high proficiency group, however, not all differences were statistically significant. As table 11 indicates, the difference between the true-false-not given and the matching format was not significant, thus implying that high-scoring students achieved rather similar results. Furthermore, disregarding the much larger than typical effect size of the difference between the multiple choice part and other item formats, this result could also suggest that highly proficient students are

[10] Significant p-values are marked with an asterisk.

less influenced by the test format than low-scoring students. A discussion of why the results might have turned out the way they did can be found in chapter 5.

4.4.3. Students' rating of the three selected-response formats

This chapter explores how the testees rate the difficulty of the multiple choice, true-false-not given, and matching items. To find an answer to this question students were asked to fill out a questionnaire subsequently to the reading test, where they were asked to rate the difficulty of each test format on a 5-point Likert scale ranging from 1 (very easy) to 5 (very difficult). Further, they were asked to justify their opinion so as to prevent that they just randomly selected a point on the 5-point Likert scale (cf. Appendix 7.2). The frequencies as well as the charts were calculated and produced with the help of the frequencies command of SPSS Version 17.

As can be seen from figure 4 the majority of the 97 testees, namely 45.4% considered the multiple choice format as "okay". An almost equal number of students, namely 23% and 24% argued that the multiple choice items were "easy" respectively "difficult". As can be seen from figure 4 below only few students rated the multiple choice items as "very easy" (1%) or "very difficult" (5.2%).

Figure 4 Students' ratings test formats

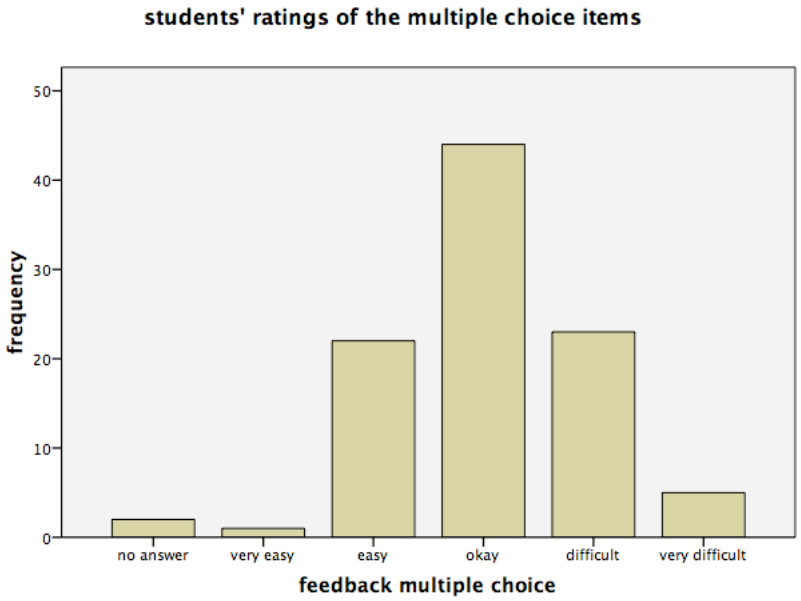


Figure 5 Students' ratings test formats

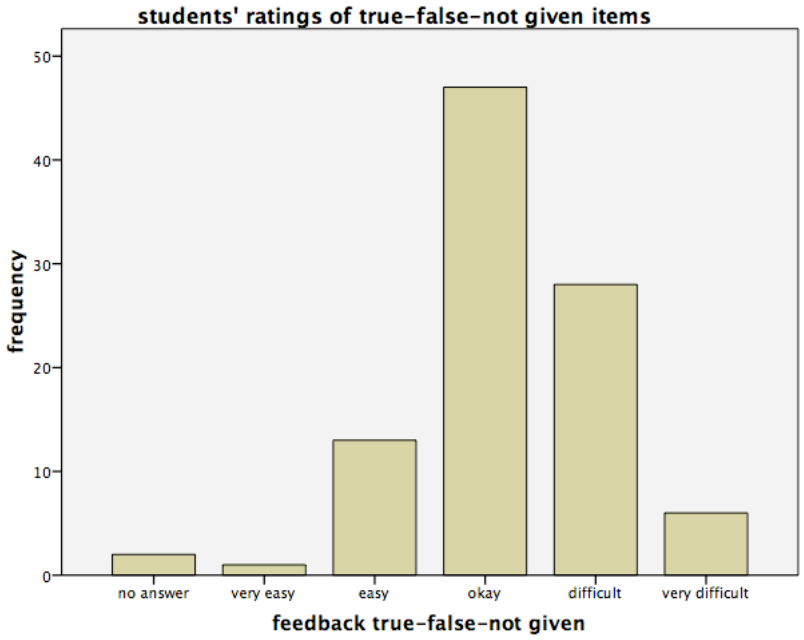
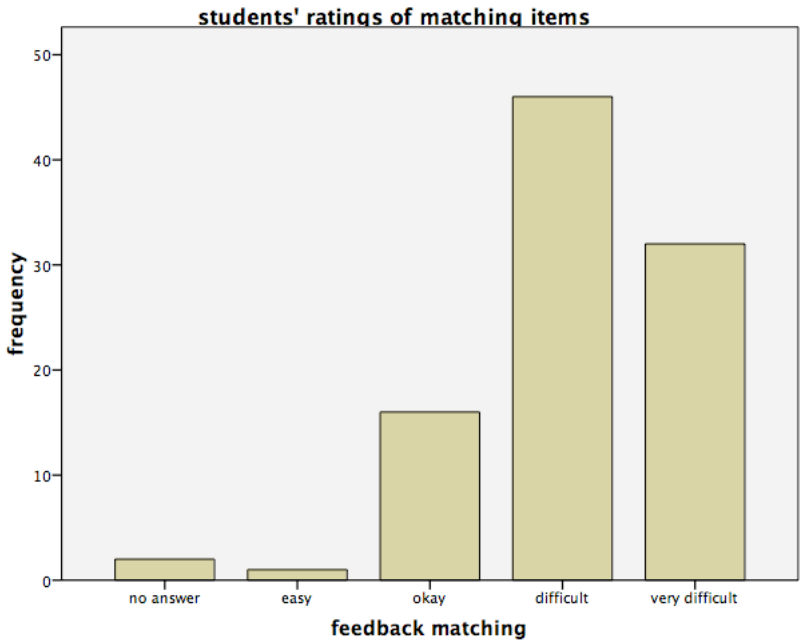


Figure 5 visualizes that the majority of the 97 testees, namely 48%, considered the true-false-not given format as “okay”, whereas 29% rated it as “difficult”. As compared to the multiple choice items, considerably fewer people agreed that the true-false-not given items are “easy”.

Figure 6 Students' ratings test formats



The matching format is considered to be the most difficult item type as 80% of the test takers admitted that this format is “difficult” or even “very difficult”. More specifically, 47% of the testees rated the matching tasks as difficult, whereas 33% of the participants even agreed that this format is very difficult (cf. figure 6). Only 16% thought that the matching tasks were “okay” and only 1% of the test participants considered matching items as “easy”. Compared to the multiple choice and true-false-not given tasks, which 1% of the students regarded as “very easy”, no one rated the matching format as “easy”.

Taking the above mentioned aspects into consideration, students considered the multiple choice format as the easiest format, as 69% of the test takers rated it from “very easy” to “okay”. The true-false-not given items qualified as the second easiest format, as 63% of the participants ranged it from “very easy” to “okay”. The matching format, however, was considered the most difficult format as 80% of the testees rated it as “difficult” or even “very difficult”.

4.5. Specific findings in relation to research questions

In line with the theoretical part of the thesis the main research questions are concerned with whether and how the different test formats affect students’ reading comprehension performance. Further research questions are concerned with reader attributes affecting test performance: gender, and

interest. It is up to additional research questions to investigate whether differences between the two schools and their different preparations for the Matura, i.e. school leaving examination, are reflected in the students' performance on the three selected-response formats. Section 4.4.3 explores if an indication of the number of correct statements next to each multiple choice question improves students' performance.

4.5.1. Are reader attributes affecting students' performance on the three test formats?

4.5.1.1. Gender

In the scientific literature the findings on whether and how male and female test takers' performance on reading comprehension tests differs are rather diverse. More specifically, the only fixed-response format research literature has focused on so far is the multiple choice format. Other fixed-response formats such as true-false-not given and matching seem to have been neglected (cf. chapter 2.2.4.2). The majority of studies found out that boys outperform girls on multiple choice tasks (Spiel 2008, Barboza 1993, Gering 1993, Hardcastle 2001). Contrary to most of the literature on gender differences in multiple choice reading tests, Freeman (2008) and Phakiti (2003) found that male and female testees' score did not significantly differ. Still other studies conducted by Wen & Johnson (1997, as referred to in Phakiti 2003) and Chavez (2001, as referred to in Phakiti 2004) concluded that females outperform males on reading comprehension tests regardless of topics. Due to a variety of different study designs, disciplines and cultural differences regarding the role of men and women in society, as well as what are considered as appropriate male and female activities divergent findings across different studies are barely surprising.

The aim of this research question is thus to shed some light on differences in male and female testees' performance on three different selected response formats: multiple choice, true-false-not given, and matching. For a comparison between means, again difference inferential statistics with the help of an independent samples t-test were conducted, after the assumptions for the t-test, i.e. normal distribution, equality of variances, independent data, were found to be met. The null hypotheses can be formulated as follows: The means

of male and female test takers on the multiple choice, true-false-not given, and the matching format are the same at p<0.05.

Table 15 Comparison of male and female test takers' performance

Variable	Mean	SD	t	df	p[11]
multiple-choice task			-1.563	95	.121
female	.6560	.12127			
male	.6964	.13343			
true-false-not given task			-2.087	95	.040*
female	.3810	.18840			
male	.4606	.18760			
matching task			-.107	95	.915
female	.3752	.25349			
male	.3798	.15741			

p<0.05

Table 12 illustrates that male test takers achieved significantly better results than females on the true-false-not given format, (p=.040), while on the multiple-choice (p=.121) and on the matching-format (p=.915) boys' and girls' results did not differ significantly, which can be seen from the respective p-values in table 12. An inspection of the means of the two groups indicates that the average girls' score on the true-false-not given part of the test is with 38% out of 18 points significantly lower than the mean of the boys, which is 46% out of 18 points. The effect size *d* is .42, which is, according to Cohen (1988: 79), a small to medium size effect.

Summing up, the null hypothesis of no difference between boys and girls results can be accepted for the multiple-choice and the matching test format. On the true-false-not given part of the test, though, the null hypothesis of no differences has to be rejected due to t (95)= -2.087, p=.040. On the one hand, these results do not fit the findings of most research studies (Barboza 1993, Geering 1993, Hardcastle 2001, Spiel 2008) which showed that the multiple-choice format significantly favors male test takers. On the other hand, however, the results of the present study confirm the results of recent studies by Freeman

[11] Significant p-values are marked with an asterisk.

(2008) and Phakiti (2003) that male and female testees' scores on the multiple choice reading comprehension test do not significantly differ.

4.5.1.2. Interest

The aim of the present research question is to explore if interest, a reader attribute, affects students' performance on the three selected-response formats. As previously outlined in chapter 2.2.4.2. a large number of research studies suggests that interest plays a vital role and influences students' reading comprehension positively (Asher & Markell 1974, Belloni & Jongsma 1978, Le Loup 1993, Oakhill & Petrides 2007). Accordingly, the present study wants to find out whether students achieve better results when the reading text is more interesting, although this can be accomplished only to a very little extent according to the underlying conditions of this test being a replicated test.

The topic of the reading comprehension text on tuition fees can be considered a neutral topic. The idea for the above mentioned research question emerged from the fact that the newspaper article for the reading test about tuition fees at universities was published in the year of 2003. In September 2008, however, the Austrian parliament passed an amendment to the university law regarding tuition fees saying that students who are able to complete the degree program within the minimum time plus two additional semesters do not have to pay any tuition fees (student point – information on financial matters – tuition fees: http://studieren.univie.ac.at/index.php?id=657, 23 October 2009). If the reading test had taken place before the end of 2008 the topic might possibly have been relevant and interesting for grade 11 test takers. This reading test, however, was carried out in May 2009 and thus the researcher assumed that a newspaper article published roughly 6 years before, covering a circumstance that has been almost abolished, would not be of great interest to the target group. Therefore, on the odd-numbered test papers of test takers the publishing date was left out while the other group of students with even-numbered test papers had the publishing year 2003 given. The researcher is well aware that it is rather arguable whether firstly, such a little modification makes a difference at all and secondly, whether the omission of the publishing date can be directly related to an increase in the pupils' interest in the newspaper article. But research may not grow and unfold if new interesting aspects are not tested.

Due to the fact that again a comparison between the means of two independent variables (test takers who have the publishing date of the newspaper article given, and test takers who have no publishing date given) was desired, an independent samples t-test was carried out after the underlying assumptions were found to be met. Table 13 shows the results of the independent samples t-test with the null hypothesis that the means of the two groups of students (those who do have and those who do not have the publishing date given) would be equal at α<.05.

Table 16 Comparison of students who had the publishing date (not) given

Variable	Mean	SD	t	df	p
multiple-choice task			.593	95	.554
publ date given	.6682	.13463			
publ date not given	.6837	.12286			
true-false-not given task			.660	95	.511
publ date given	.4074	.18096			
publ date not given	.4331	.20190			
matching task			.040	95	.968
publ date given	.3766	.23334			
publ date not given	.3783	.18766			
all three tasks total			.429	95	.669
publ date given	.4565	.15010			
publ date not given	.4690	.13698			

Table 13 illustrates that the differences between students who had the publishing date of the reading text given and those who did not are statistically not significant. Due to the fact that *t* is neither significant for the multiple-choice part of the test (p=.554), nor for the true-false-not given part (p=.511) nor the matching part (p=.968) it can be concluded that there is no difference between test takers who had the publishing date given and those who had not. Thus, the null hypothesis of no difference between the two groups has to be accepted. As opposed to the studies by Asher & Markell 1974, Belloni & Jongsma 1978, Le

Loup 1993, Oakhill & Petrides 2007 in the present study students' interest in the topic, as defined by placing the reading test chronologically, does not influence their performance positively.

To find out whether students who have the publishing year 2003 given consider the newspaper article on a no longer relevant topic to be less interesting than students who are not informed about the publishing year, students were asked to rate their interest in the topic on a four point Likert scale in the questionnaire. Out of the 98 test takers, 47 had the publishing date given, while 51 testees missed that piece of information. Figures 8 and 9 below illustrate that in both groups 45% of the testees rated the newspaper article on tuition fees as interesting, while 38% of the students who had the publishing year given and 39% of those without the indication of the publishing year evaluated the newspaper article as rather uninteresting.

Figure 7 How interesting is the reading text when the publishing date is given?

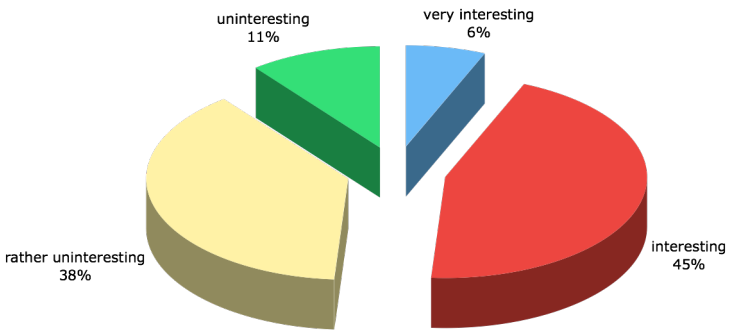
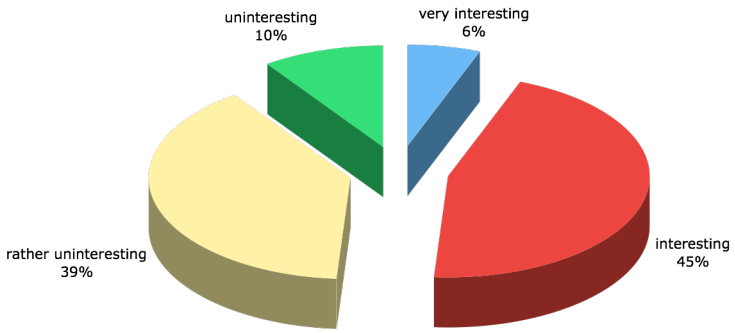


Figure 8 How interesting is the reading text when the publishing date is left out?



The above figures show that the indication of the omission of the publishing year does not influence the testees in their rating of the text as interesting or uninteresting. In short, in this very case it does not matter whether the publishing year of the reading text is omitted or not. Furthermore, this result might also indicate that the omission of the publishing year does not contribute to an increased interest in topic that is no longer relevant to the students.

4.5.2. School A and school B compared on the three selected-response formats

The aim of this research question is to explore whether differences between school A and school B are reflected in the respective students' performance on multiple choice, true-false-not given, and matching items. As previously explained in chapter 3.2.1. school A and school B differ among other aspects in the *Matura* forms they are administering. While school A is running the school pilot project "listening neu", school B pupils still take the traditional *Matura* with an unstandardized listening comprehension part. In consequence of the different item formats featured in the *Matura* forms, teachers admitted in interviews that they prepared their students differently. While school A teachers engage their students in numerous different item formats and in exercises focusing on all four skills, school B teachers do not use a wide range of item formats and do not regularly include all four skills in their lessons. On the basis of short teacher interviews it might be supposed that teachers of school A are

more in favor of the new standardized *Matura,* which will be implemented in 2013/2014 (Friedl Lucyshyn 2010: 7), than school B teachers, and that this opinion is reflected in how they engage students in all four skills and not just those skills featured in the current *Matura*. According to these interviews, school A teachers practiced reading comprehension exercises more frequently with their pupils than school B teachers. Nevertheless, teachers and pupils from both schools stressed that the three selected-response formats used in the present study: multiple choice, true-false-not given, and matching were familiar to them. All teachers agreed that within the three question formats their students were most familiar with the multiple-choice format, followed by the true-false-not given and the matching format.

Given that school A pupils practiced reading comprehension exercises more often than school B pupils, it might be assumed that school A pupils outperform school B pupils on each of the question formats. Thus, a one-tailed mean comparison with the help of an independent-samples t-test will be carried out, so as to find out whether the null hypothesis that the means of school A and school B pupils are equal on the three question formats can be rejected at the .05 alpha level. To test whether school A and school B pupils are equally affected by the test format, paired-samples t-tests will be conducted for each group separately[12].

[12] The underlying assumptions of independent-samples and paired-samples t-tests were found to be met.

Figure 9 Mc tasks / school A and B

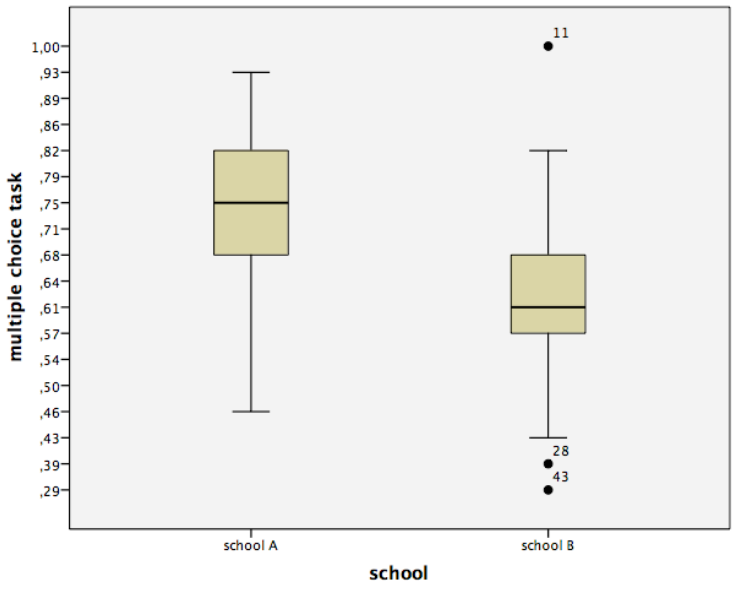


Figure 10 T-f-ng tasks / school A and B

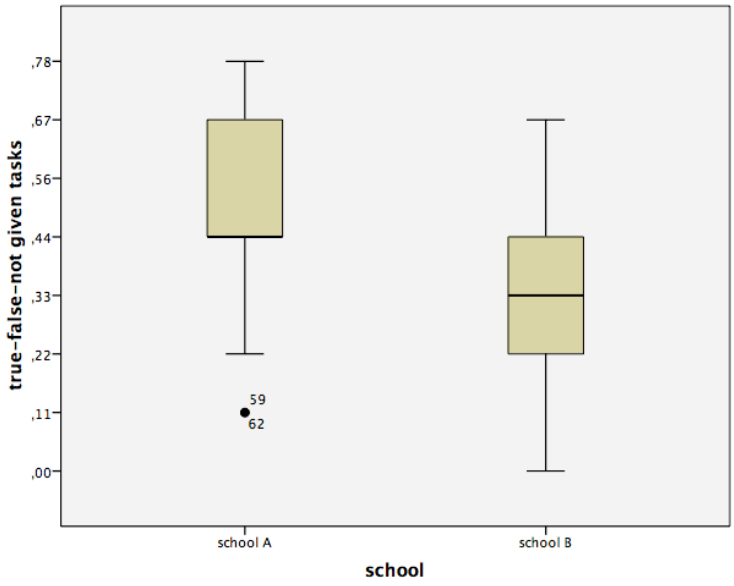
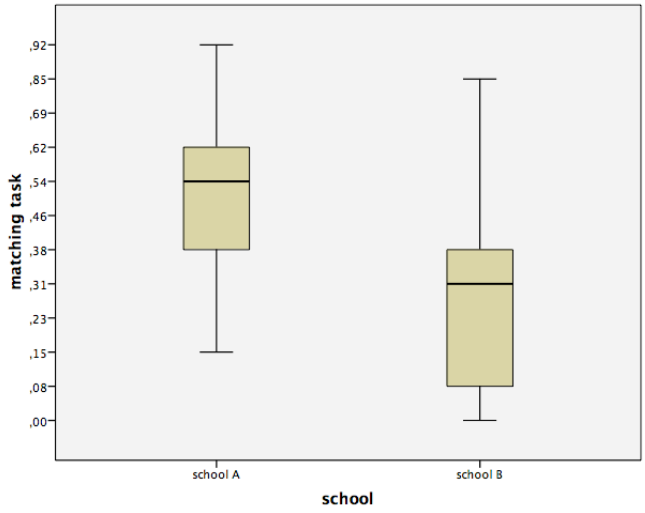


Figure 11 Matching tasks / school A and B



Figures 10 to 12 provide information on the mean values (thick line), on the ranges (from the beginning to the end of the vertical line) as well as on outliers (dots accompanied by the number of the test taker). Figures 10. 11, and 12 highlight that school A pupils outperformed school B pupils on all three question formats. The difference of school A and school B pupils' means on the multiple choice part is statistically significant (t(95)= -4.707; p=.000). Analysis of the two group means indicates that the average multiple-choice score for school A students (73% of 14 points) is significantly higher than the score (62% of 14 points) for school B students. In addition to being significant, the difference between school A and school B students' scores on the multiple-choice question format is meaningful because the effect size d is large (d= .92).

Considering the true-false-not given part of the test, the difference between school A and school B pupils' scores on this part of the test is statistically significant (t (95)= -3.949, p=.000). Thus, school A pupils (49% of 18 points) outperform school B pupils (35% of 18 points) on the true-false-not given format. The effect size d is .78, according to Cohen (1988: 79) a large effect.

On the matching part of the test again, school A students achieve significantly higher means than school B students (t (95)= -6.523, p=.000). The difference between school A and school B test takers, however, is biggest on the matching format (cf. Figure 12, Table 14). While school A test takers on

average achieve 12.9 points on this part of the test, school B students only get 6,9 points. The fact that the difference between the means on this part of the test is bigger than on the other item formats is also confirmed by the effect size d which is with 1.41 according to Cohen's guidelines (1988: 79) not only a very large but also a much larger than typical effect size.

As school A and school B students' results on all three test formats differ significantly, the question arises if school A and school B students are equally affected by the test format. To tackle this question paired-samples t-tests were conducted for each school separately. The outcomes can be seen in table 14.

Table 17 Comparison of school A and school B pupils' performance

	school A	**school B**
test format	**mean (%)**	**mean (%)**
multiple-choice	.733	.622
true-false-not given	.494	.3511
matching	.498	.2646

Table 14 highlights that while school B students achieve the lowest mean in the matching format, school A students score worst on the true-false-not given format, although the difference between school A testees' scores on the true-false-not given and the matching test could have arisen due to chance as this difference proved to be statistically non significant (cf. table 15 below). The multiple choice format, however, qualifies in both schools as the easiest item format. Further, it can be seen that school A pupils almost perform the same on the true-false-not given and the matching part. At school B, however, the mean differences between all three selected-response formats are statistically significant, which is visualized in table 15 below.

Table 18 Contrasts between means on different test formats[13]

contrasts	t-values	df	Sig. (p)	effect size d
school A mc / t-f-ng	10.360	46	.000*	1.71
school A t-f-ng /matching	-.104	46	.918	
school A mc / matching	8.312	46	.000*	1.64
school B mc / t-f-ng	9.719	49	.000*	1.69
school B t-f-ng / matching	2.460	49	.017*	0.47
school B mc / matching	13.992	49	.000*	2.38

The mean differences between multiple choice and true-false-not given and matching tasks are significant in school A and B and the effect sizes d, which can be seen in the far-right column are with values ranging from 1.64 to 2.38 much larger than typical. While in school A students perform the same on the true-false-not given and the matching part of the reading comprehension test, school B attendants achieve statistically significantly better results on the true-false-not given part, with d = 0.47 indicating a medium or typical effect size.

Taking all aspects into consideration, it can be concluded that school A test takers outperform school B test takers on all three question formats at p=.000. This result puts into question whether the different preparation methods for the *Matura* can be held accountable alone for the large differences between school A and school B. Further, this research question proved that higher scoring students, in this case pupils from school A, are less sensitive to the testing method than weaker students.

[13] Significant p-values are marked with an asterisk.

4.5.3. Number of correct answers (not) given in the multiple choice test

The present research question aims at investigating the case of multiple choice items in which more than one answer is correct: namely if an indication of the number of correct answers next to each item improves the students' scores. This research question appears to be especially relevant to teachers, or more generally to people involved in language testing who use multiple choice tests as a means of assessment.

To tackle this question each class was divided into two groups, which was done in a rather simple but straightforward way. Students who got even-numbered test papers had the number of correct statements given in brackets next to each multiple-choice item, while students with odd-numbered test papers had not. This method made it possible to divide each class into two groups. In the course of the explanation of the test procedure, the test administrator deliberately avoided to mention this distinction and its sense so as not to encourage students to look at their neighbors' papers to catch a glimpse of the number of correct statements.

A one-tailed mean comparison was made between those students who had the number of correct statements given and those who missed this piece of information. This was done by using the independent-samples t-test, of the null hypothesis that the means of these two groups of test takers would be equal at $\alpha<.05$. All assumptions underlying the t-test (normal distribution, independent data, equality of variances) were found to be met.

Table 19 Comparison of students who had the nocs[14] (not) given next to each mutiple choice item

Variable	Mean	SD	t	df	p
multiple choice items			-.593	95	.554
nocs given	9.3542	1.88487			
nocs not given	9.5714	1.71998			

Table 16 shows that based on a comparison of the means, students who had the number of correct statements indicated next to each multiple-choice

[14] NOCS = number of correct statements

question did not achieve better results on this question format than students who did not have this information. Due to the fact that the *t* is not statistically significant (p=.554), it can be concluded that there is no difference between the results of these two groups on the multiple-choice part of the reading test. Therefore, the null hypothesis of no difference between students who do and do not have the number of correct statements given has to be accepted.

This was the test that had to be carried out first, because its result determines whether the differences between test takers with odd- and even-numbered test papers prove to be statistically significant, which would imply that they have to be treated as separate groups in each of the independent variables, i.e. schools, tracks, genders, and further tests.

5. Discussion

5.1. Summary of the results

The aim of the present study was to investigate the influence of three different selected-response methods, i.e. multiple choice, true-false-not given, and matching, on the testees' reading comprehension performance. The subjects were 97 grade 11 students of two different Viennese *Gymnasien*, i.e. a type of secondary school preparing students for higher education at university. While the present section briefly summarizes the results, the following section (5.2.) is dedicated to discussing them and suggesting implications. From an examination of the results (chapter 4) the following findings emerged.

1. The three selected-response formats multiple choice, true-false-not given, and matching, which all assess the same underlying ability, reading comprehension, correlate only to a medium extent.
2. Students' reading comprehension as assessed by multiple choice items is significantly higher than their performance on true-false-not given and matching items. Students' performance on true-false-not given, and matching items, however, does not differ significantly.
3. Highly-proficient students achieve their best results on the multiple choice items. Their performance on true-false-not given and matching items does not differ significantly. Low-level students, however, achieve significantly different results on each of the test formats. They achieve their best results on the multiple choice items, while they score worst on the matching format.
4. The majority of students regard the multiple choice and the true-false-not given formats as "okay", while they consider the matching tasks to be "difficult".
5. Male and female test takers' performance as assessed by multiple choice and matching items does not differ significantly. Male testees, however, are significantly favored by the true-false-not given format.
6. Within the present study interest, as previously specified, does not influence students' reading comprehension performance positively.

7. School A students' reading comprehension performance on each of the three selected-response formats is significantly better than school B students' performance.
8. An indication of the number of correct statements next to each multiple choice item does not influence students' performance significantly.

5.2. Discussion

For a discussion of the findings relevant research studies as well as the opinion of the present researcher will be taken into consideration. Furthermore, suggestions of further research will be included.

5.2.1. Discussion of the criteria of test quality

According to the criteria of analysis of test quality, the matching format is the most satisfying format as here only 23% of the items should be revised due to their facility values and discrimination indices. The true-false-not given format, however, qualified as the least satisfying format as here 44% of the test items need revision. In the multiple choice format 28% of the items should be revised. The reason for the revision of these items is that they are too difficult and too little discriminating. If the test was an achievement test it would be possible to retain too difficult items in the test if the stakeholder needed information about these items. The present test, however, is a proficiency test and as in this case the language tester does not know about the students' learning history it is more important that the items fulfill the criteria of test quality. Thus, items which are too difficult and inappropriately discriminating should be revised as they question the reliability of the test. The overall reliability of the test with a Cronbach alpha coefficient of .72 is at a satisfying level.

5.2.2. Discussion of the correlation of the three selected response formats

The assumption that all three-selected response formats correlate highly because they are testing the same underlying ability, reading comprehension, was not completely met. According to Cohen's (1988) benchmarks the multiple choice, true-false-not given, and matching items only correlate to a medium or typical extent. Brown (1988: 97), however, who uses different benchmarks

regards the correlations of the three selected-response formats as weak correlations. The two formats that show the smallest correlation among the three fixed-response formats are the true-false-not given and the matching items (r= 28). Here, the question arises why the test formats only correlate weakly (Brown) or to a medium extent (Cohen).

One of the reasons might be found in the construction of the test. While researchers such as Shohamy (1984), Wolf (1991, 1993), and Elinor (1997) (cf. section 2.2.4.) developed different but stem-equivalent reading test formats which are all assessing the same part of the text, Bensoussan's (1984) and the present study were designed to compare test formats of an already existing test, which has not been specifically developed for the purpose of a comparison. Consequently, the multiple choice items do not assess the same part of the text as the true-false-not given and the matching items. The different parts of text the items are assessing might differ in their degree of difficulty, which might have led to the medium size correlation coefficients of the three formats. This argument is supported by the students' opinion on the difficulty of the different assessment formats. While the majority of testees rate the multiple choice and the true-false-not given formats as "okay", they consider the matching format as "difficult".

Another interpretation of the weak to medium size correlations might be that the three selected-response formats do not assess the same abilities in the reader. While multiple choice and matching items require comprehension and selection on part of the students (cf. Shohamy 1984: 157), true-false-not given items seem to involve comprehension, the ability to understand the concept of something not given, something inexistent, and selection. This view is supported by Alderson (2000: 222) who asserts that the category "not given" can lead to considerable confusion among test takers.

Recently, the *Bundesinstitut bifie Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens* (cf. Friedl Lucyshyn 2010: 10) removed the true-false-not given format from the new standardized listening comprehension test, which is part of the new standardized competence-oriented *Matura.* To date there is no official document in which the *bifie* states the exact reasons for the removal of this test format. Researchers like Alderson (2000), Alderson et al. (1995), and Hughes (2003), however, mention negative

aspects of the true-false-not given format, which along with unpublished research commissioned by the *bifie* might have led the latter to this decision.

Alderson et al. (1995: 51) argue that true-false-not given items are very demanding when used in listening comprehension tests, especially if students can listen to the text just once. The high difficulty level of this item format in listening comprehension tasks might be explained by the fact that listening is a very spontaneous activity in which the category "not given" can lead to considerable confusion especially if the students cannot listen to the passages again. While true-false-not given items in listening comprehension tests are considered highly demanding, students might find them easier in reading comprehension tests provided that students are able to carefully look at the passages covered by the respective true-false-not given items again. But still, it could be supposed that the concept of something not given creates considerable confusion on part of the reader (cf. Alderson 2000: 221). Whatever the reason, if true-false-not given items are used in reading comprehension tests, further research on the underlying concept of the true-false-not given format, its advantages and disadvantages, as well as on the comparability of this format with other selected-response formats is indispensable.

5.2.3. Discussion of the main research questions

The central research question guiding the present study, i.e. whether and how students' performance on three different selected-response formats assessing reading comprehension differs, led to the following results. Students performed best on the multiple choice format, the format they reportedly were most familiar with. Their scores on the true-false-not given and the matching format were significantly worse. Students' familiarity with the multiple choice format was significant, and the effect sizes were with d values of 1.63 (difference between multiple choice and true-false-not given items) and 1.76 (difference between multiple choice and matching format) much larger than typical (Cohen 1988: 79). The difference between testees' results on the true-false-not given and the matching part, however, was non-significant. One reason why testees performed best on the multiple choice format might be that this is the test format they are most familiar or almost "all too familiar" with. According to teacher interviews students are confronted with multiple choice items not only in English

lessons but increasingly also in subjects other than English. Concerning true-false-not given and matching tasks, teachers reported that students were more familiar with the true-false-not given than with the matching format, which they referred to as the "most modern" fixed-response format. Interestingly, the familiarity factor did not play a role in the students' performance on matching and true-false-not given items, as their performance on these formats did not differ significantly. After having discovered that the students' performance on the multiple choice and the other two fixed-response formats significantly differed, the question arises whether high proficiency students react less sensitively to different test formats than low proficiency students.

A study by Shohamy (1984) investigating the influence of assessment format, text type, and language of the text (L1 or EFL) on students' reading comprehension performance revealed that low-level students were considerably more sensitive to the testing method (multiple choice and open-ended items). High proficiency students, however, appeared to be hardly affected by this variable. In the present study partly similar results emerged. High-level students' results on the multiple choice part were significantly better than their scores on the true-false-not given and matching formats, which is highlighted by very large effect sizes of 2.00 and 2.09. Their performance on true-false-not given and matching formats, however, did not significantly differ. While the results of the high proficiency group are in line with the overall results, low-level students' performance differs. In contrast to the high-proficiency group, low-level students performed significantly better on the true-false-not given than on the matching format, which is exemplified by an effect size d of 0.44, a slightly smaller than medium effect size. In line with the overall results and those of the high-proficiency group, low proficiency students also achieved significantly better scores on the multiple choice than on the other selected-response formats. These findings indicate that low-level students react more sensitively to true-false-not given and matching items, while for highly proficient students the difference between true-false-not given and matching formats is not reflected in their performance. Interestingly, however, all students performed significantly best on the multiple choice format, which could be explained by the testees' "overfamiliarity" with this test format.

The reason why for high-level students as well as for the overall group of testees, the difference between their performance on the matching and true-false-not given assessment tasks was not significantly different, could be that these students are able to manipulate language well, regardless of the test format, especially because of their higher level of proficiency. In contrast, low-level students seem to be more influenced by the different formats and they also appear to be affected by their familiarity with the test formats, meaning that they achieve the worst result on the format with which they are least familiar.

Given that the performance of the overall group of test takers as well as of the high-proficiency students on the true-false-not given and the matching tasks does not differ significantly, the question arises as to whether testees consider these two formats to be of similar difficulty. The majority of test takers rated the multiple choice and the true-false-not given format on a 5-point Likert scale ranging from 1 (very easy) to 5 (very difficult) as “okay”. The matching format, however, was regarded as “difficult” or even “very difficult” by 80% of the test takers. These results seem to suggest that students have trouble in judging correctly their performance on true-false-not given items. While they regard the true-false-not given format as “okay” they actually achieve the same results as on the “difficult” matching format. In other words, it is the true-false-not given format that leads to a discrepancy between actual performance and self-evaluation. Thus it could be inferred that the true-false-not given format is indeed the “trickiest” format. A further reason for the test participants’ misjudgment of the true-false-not given format could be that they are not able to judge the difficulty that arises from the concept of something not given, something inexistent. As previously mentioned, this interpretation is supported by Alderson (2000: 222) who argues that the “not given” option creates “considerable confusion” on part of the students, “especially with items intending to test the ability to infer meaning”. Combining the findings with the literature (cf. Alderson 2000), it might be presumed that the concept of something not given is indeed very difficult for students. In order to determine whether true-false-not given items are too demanding for students, and thus might interfere with their reading comprehension abilities, further research seems necessary.

5.2.4. Discussion of RQ 1

The following discussion centers on differences in male and female test takers' performance on the three reading comprehension test formats. Quite contrary to most research studies which either reported that male test takers significantly outperformed females on multiple choice tests (Barboza 1993, Geering 1993, Hardcastle 2001, Spiel 2008) or commented on a superiority of female participants on the multiple choice format (Chavez 2001, as referred to in Phakiti 2003), the present study showed that male and female test takers' reading comprehension performance as assessed by multiple choice and matching formats does not differ significantly. The result of the present study corresponds to findings by Freeman (2007) who reported that male and female test takers' performances on a reading comprehension test did not differ, which he claims is contrary to most of the literature found on gender differences. The true-false-not given format, however, significantly favors male test takers. The effect size of the difference between male and female performance on the true-false-not given task indicates with a d of .42 a small to medium effect. One reason why male test takers outperform female testees on the true-false-not given format in the present study might be that the male learners were able to understand the concept of something not given somehow better than the girls.

Another possible explanation, quite frequently mentioned in the context of multiple-choice items but due to the given set of answers also true for other fixed-response formats, might be that boys rather than girls are more willing to guess, an argument listed in numerous studies (Barboza 1993, Geering 1993, Freeman 2997, Hardcastle 1991, Spiel 2009). Additionally, male testees might be at an advantage here by their employing of a so-called "eyes down" approach, which is said to be better suited for selecting a correct answer out of a set of options, while girls might be inhibited by seeing the relative "rightness/correctness" of each option (Stobart, Elwood and Quinlan as referred to in Geering 1993). Female test takers' inhibition of recognizing the relative "rightness/correctness" of each option could be extremely crucial in the case of true-false-not given items, where a further difficulty is related to the understanding of the concept of something not given, something non-existent. Nevertheless, additional research on male and female test takers' performance differences on true-false-not given items seem to be warranted. It would be

really interesting to take stem-equivalent multiple choice and true-false-not given items, which are assessing the same part of a reading text, and explore whether male and female test takers' performance differs significantly.

5.2.5. Discussion of RQ 2

An examination of the second reader attribute, interest, could only be explored to a small and probably rather limited extent within the present study which is reflected in its results. No significant difference was found between the performance of test takers who had the publishing year 2003 of the newspaper article on tuition fees given and those who missed that piece of information. Further, no difference between the students' rating of their interest in the topic of the newspaper article on a 4-point Likert scale ranging from 1 (uninteresting) to 4 (very interesting) could be found between the two groups. Thus, the present study cannot be grouped with a large number of research studies (Asher & Markell 1974, Belloni & Jongsma 1978, Le Loup 1993, Oakhill & Petrides 2007) which revealed that interest influences students' reading comprehension performance positively. The relation between an increased interest and an improved performance might be the following: a high level of interest is said to be interrelated with a motivation to understand a text covering an interesting topic, which in turn triggers cognitive processes central to reading comprehension (Oakhill and Petrides 2007: 232).

An explanation why in the present study students were not positively affected by the article without the publishing year, which was supposed to be more interesting than the article with the publishing year 2003, could be that the omission of the publishing year is not directly related to an increased interest in a topic that is not that relevant to the students anyway. An additional explanation why the omission of the publishing year did not make a difference could be that students did not read attentively enough. Therefore, it might be possible that those students who had the publishing year 2003 given, overlooked the date and did not even notice that the newspaper article had been published six years ago and was no longer relevant.

Furthermore, the fact that students had to rate their interest in the reading test after having completed the latter, could have blurred their true interest and this fact could be responsible for the similarities related to the two

groups' rating of interest of the reading comprehension text. If students read a very interesting text which is hard to read and where the reading comprehension tasks are difficult to answer, this could make them confound interest with difficulty. Contrarily, if students read a less interesting but easily readable text and can answer the reading comprehension tasks to their satisfaction, this could enhance their interest in the text and the topic as such.

Given all these reasons, different research designs seem to be vital for a closer examination of the influence of interest on students' reading comprehension performance. Researchers like Asher & Markell 1974, Belloni & Jongsma 1978, Le Loup 1993 employ research designs, which appear to be better suited for an investigation of the influence of interest on the students' performance. In their studies students have to select from among a pool of different topics the most and least interesting ones prior to the reading comprehension test. The actual reading test then features one high and one low interest topic. Thus, their research design, which really takes into account the students' interests, appears to be better suited.

5.2.6. Discussion of RQ 3

An additional research question was concerned with the investigation of possible differences between school A and school B students' reading comprehension test performance. Are differences between school A and school B reflected in the respective student' performance? Results demonstrate that school A test takers outperform school B test takers on all three test formats at p=.000. The largest difference between school A and school B students' performance is to be found on the matching format, which is indicated by a d of 1.4, an effect size which is much larger than typical. The second largest difference between school A and school B students' reading comprehension performance emerged on the multiple choice test (d=.92). The difference between school A and school B testees' performance on the true-false-not given and matching tasks is with a d of .78 still large, but the smallest of the three test formats.

The reason for the significant and large difference between school A and school B students' reading comprehension performance could be related to the different *Matura* forms the schools are currently running, i.e. school A is

administering the school pilot project standardized “listening neu”, while within the traditional *Matura* carried out at school B, the listening comprehension test is unstandardized. These two distinct listening comprehension tests feature different test formats. While the standardized listening comprehension “listening neu” consists of multiple choice, true-false, matching, cloze tests and grids, the formats featured in the listening comprehension tests of the traditional Matura are summary writing and short answer questions. Given the different test formats of the *Matura*, teachers might prepare their students differently. While school A teachers increasingly practice the assessment formats featured in the standardized listening comprehension test “listening neu”, school B teachers do not focus that much on a range of different test formats. Thus, it is little surprising that school A students perform better than school B students as they are more familiar with the assessment formats featured in the present reading comprehension test. Moreover, this difference in performance could also be attributed to the respective teaching styles of school A and school B teachers. While all three teachers at school A said during the teacher interviews that they frequently practiced reading comprehension exercises with selected and constructed-response formats, only one teacher at school B admitted the frequent practice of reading comprehension exercises.

Thus, it can be concluded that school A students outperform school B students because their teachers prepare them better, on the one hand for reading comprehension exercises in general, and on the other hand on a number of different test formats similarly to the ones featured in the present reading comprehension test. The reason why the largest difference between school A and school B pupils’ performance is to be found on the matching part, the format the students are least familiar with, might also be related to the fact that while matching items are featured in the standardized listening comprehension test “listening neu”, school B students are hardly confronted with that test format.

Moreover, the significant difference between school A and school B pupils’ reading comprehension performance could be related to the pupils’ time spent in English speaking countries. While according to teacher interviews, school A pupils had already spent some time in English speaking countries in the context of school exchange programs, school B pupils had not taken part in

any of these programs. These stays in English speaking countries could have motivated school A students to increasingly engage in extracurricular reading activities, which might have probably led to their better reading skills.

A further interpretation of the fact that school A students significantly outperform school B students on all three test formats could also be related to the different sociocultural backgrounds of the students. While according to teacher interviews most school A pupils' parents have an academic background, a large number of school B pupils' parents do not hold a university degree. Although this interpretation is highly speculative, the better performance of school A pupils might possibly be related to the fact that parents holding a university degree rather encourage their children to engage in reading activities even outside of school. Furthermore parents' academic background might have an influence on their children's attitudes towards studying, and reading in particular. The present study can only speculate on this relation between parents' background and students' reading comprehension performance. Nevertheless, it would definitely be interesting to explore this question further by carrying out additional reading comprehension tests and by introducing questionnaires in which the parents mention their profession as well as whether or not they encourage their children to study hard and engage in extracurricular reading activities. On the basis of such a study more reliable interpretations about the relation of parents' background and students' reading comprehension performance could be made.

Within this research question one further investigation was conducted. Each school's individual performance on the three assessment tasks was calculated. Results showed that while school B students' performance on each of the three fixed-response formats differed significantly, school A students' performance on true-false-not given and matching formats did not differ significantly. Taking into consideration the students' overfamiliarity with the multiple choice format, it can be concluded that school A students, who outscored school B students on all three fixed-response formats and can thus be regarded as the high-proficiency group, react less sensitively to matching and true-false-not given formats than school B students, in this case referred to as the low-proficiency group. This result is also in line with the results of all high and low proficiency test takers, disregarding the different schools (cf. section

4.3.2). The reason why high-proficiency students react less sensitively to different selected-response formats they are not overfamiliar with might be that their higher level of proficiency enables them to manipulate language better. Therefore, it does not make a huge difference to them whether they have to answer matching or true-false-not given items. The results of the present study do not correspond to findings by Wolf (1991, 1993), who found that high level students and low level students were equally affected by different test formats.

Shohamy (1984), however, comments on study outcomes that are in line with the findings of the present study. In her research project investigating the influence of testing method (multiple choice and open-ended tasks), text, and language of the text (L1 or FL) on the students' reading comprehension performance, she divided students into three groups according to their proficiency levels following their results on the control part of the test: low, middle, and high. Shohamy found that in her test low-level students were most sensitive to the test method, while high-proficiency students were hardly affected by this variable.

5.2.7. Discussion of RQ 4

The last research question investigated is whether in the case of multiple choice items which have more than one correct answer, an indication of the number of correct answers next to each item improves the students' performance. Precisely because no study up to date has investigated this problem, the results of the present research work cannot be compared. Results revealed that the indication of the number of correct answers does not make a significant difference. This result was quite surprising as normally pupils who know how many of the 4 options are correct would be supposed to achieve better results than students who just know that one or more than one answers are correct.

Owing to the research design, that one group of each class had the number of correct statements given, while the second group missed that piece of information, the test administrator was forced to not comment on this difference. Otherwise, this could have encouraged students' cheating. Probably this fact contributed to the result of a non-existent difference in the performance of students who had the number of correct statements given and those who missed that piece of information.

A further reason for the similar performance of students who had the number of correct statements given and those who had not could be that possibly some students did not read the instructions and multiple choice questions, in particular, attentively enough. Therefore they possibly neither noticed in the instructions to the multiple choice items nor in the indications next to each multiple choice item, that the number of correct statements was given. A counterargument, however, could be that according to the order of the reading comprehension test the multiple choice format was the very first format, which implies that students normally are most concentrated while doing the very first task. Owing to the fact, that students' concentration is highest at the beginning of the test, the argument that they did not notice the indication of the number of correct statements appears to be rather vague and is doubted by the present researcher. Thus, the fact that the students did not notice the indication of the number of correct statements next to each item appears to be related rather to the fact that students do not read instructions attentively enough, than to the order of tasks.

For a closer investigation of whether an indication of the number of correct statements in multiple choice reading comprehension items makes a significant difference, the number of subjects should be extended substantially, so that sound interpretations on basis of these results can be made. Furthermore, students should be separated by separation walls so that cheating can be prevented altogether. Additionally, future researchers could split the informants into two groups and inform only the one group of students who will have an indication of the number of correct statements given about the fact so that they know how many answers to each multiple choice item are correct.

6. Conclusion

This replication study is situated in the applied linguistics area of language testing, and reading comprehension testing in particular. The present study investigated the influence of three selected-response test formats: multiple choice, true-false-not given, and matching on the testees' reading comprehension performance in EFL. Additionally, the extent to which high and low-proficiency students were equally influenced was examined. Further, gender differences in reading comprehension performance were tried to pinpoint. Additionally, the variable of interest was explored. A manipulation was made concerning the multiple choice items: does the number of correct statements given influence students' performance positively? Moreover, any differences between school A and school B pupils' performance were attempted to discover. One authentic newspaper article on tuition fees, taken from a British quality newspaper, was used to test the participants' reading comprehension performance. The order of the tasks: multiple choice, true-false-not given remained constant. In total 97 grade 11 students from two according to their social and economic nature highly different *Gymnasien*, i.e. type of secondary school preparing students for higher education, took the reading comprehension test. An equal amount of males and females participated in this study.

Literature investigating the effect of test formats on the testees' reading comprehension performance has so far solely focused on differences between selected and constructed-response formats. Thus, the present study was the first one in comparing three selected-response formats. What most studies (Shohamy 1984, In'nami, Koizumi 2009) found was that selected-response formats, i.e. multiple choice, were easier than constructed-response formats, i.e. short answer questions. While Shohamy (1984) averred that high-proficiency students were less influenced by different test formats, Wolf (1991, 1993) claimed that even for high-proficiency students the test format made a difference.

The results that have emerged from the present study point to differences in students' reading comprehension scores which are indeed attributable to different testing methods as well as to more or less intense

preparations on the part of the teachers for these assessment tasks. Out of the three selected-response formats some methods turned out to be more difficult than others and some had a greater effect on students of low-level proficiency and on female test takers. Altogether, the results point to a high degree of the testees' familiarity with the multiple choice format, which is also reflected in their scores. While the performance of the overall group of test takers, as well as of the high-proficiency group on the true-false-not given and matching tasks did not differ significantly, low-level students achieved significantly worse results on the matching than on the true-false-not given format. This result questions a link between low level students' familiarity with the test format and their performance. For high level students' and the overall group of test takers, however, their familiarity with the test formats was not reflected in their performance.

Taking all aspects into consideration, the true-false-not given format turned out to be the most problematic reading comprehension assessment format for the following reasons. Firstly, according to the criteria of test quality a high number of true-false-not given items need revision as they are too difficult and insufficiently discriminating for the present sample of test takers. Secondly, this format significantly favored male testees. Generally speaking, testing methods which significantly disadvantage one gender should be omitted as they would contribute to unfair and gender biased assessment. Thirdly, the concept of something not given, something inexistent seemed to have led to considerable confusion among students. While the majority of testees rated the true-false-not given format as "okay" they achieved results similar to the ones on the "difficult" matching format.

Further, this argument is supported by the outcome of the correlation between the three selected-response test formats, which are all assessing the same underlying ability: reading comprehension. The formats that showed the smallest correlation were the true-false-not given and the matching format. Therefore, the question arises whether multiple choice and matching formats are testing comprehension and selection, while true-false-not given formats require not only comprehension and selection, but also the ability to understand the concept of something not given, something inexistent on the part of the students.

The results of this research call for further investigations of the differences between the three selected-response formats on a bigger group of subjects. Despite its limited scope this study has interesting implications for conducting more academic research on questions like: Are true-false-not given items indeed assessing a different construct than multiple-choice and matching items? Do students' results on the matching and multiple choice formats only differ according to their high degree of familiarity with the latter? In order to answer these questions different but stem-equivalent selected-response formats, which are all assessing the same part of a text, should be constructed and tested on a substantial number of students of different proficiency levels.

7. References

Alderson, J. Charles. 2000. *Assessing reading.* Cambridge: Cambridge University Press.

Alderson, J. Charles; Clapham, Caroline; Wall, Dianne. 1995*. Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Alderson, J .Charles; Urquhart, A.H. 1984. *Reading in a foreign language.* London: Longman.

Alexander, Patricia A., et al. 1994. "The influence of topic knowledge, domain knowledge, and interest on the comprehension of scientific exposition". *Learning and Individual Differences* 6 (4), 379-397.

Anastasi, Anne; Urbina, Susana. 1997. *Psychological Testing*. (7th edition). Upper Saddle River, New Jersey: Prentice-Hall.

Asher, Steven R ; Markell, Richard A. 1974. "Sex differences in comprehension of high- and low-interest reading material". *Journal of Educational Psychology*, 66 (5), 680-687.

Bachman, Lyle F. 2004. *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press.

Bachman, Lyle F.; Palmer, Adrian S. 1996. *Language testing in practice.* Oxford: Oxford University Press.

Baker, David. 1989. *Language Testing: A Critical Survey and a Practical Guide.* London: Edward Arnold.

Barboza, Helen Christina. 1999. "Comparing the achievement of eighth-grade boys and girls on norm referenced and performance tests in language arts, reading and mathematics". PhD thesis, University of Rhode Island and Rhode Island College.

Belloni, Loretta Frances; Jongsma, Eugene A. 1978. "The Effect of interest on reading comprehension on low-achieving students". *Journal of Reading* 22 (2), 106-109.

Bensoussan, Marsha. 1984. "A comparison of cloze and multiple- choice reading comprehension tests of English as a Foreign Language" *Language Testing*, Jun 1984 (1), 101-104.

Bernhardt, Elizabeth B. 1991. *Reading development in a second language.* Norwood, New Jersey: Ablex.

Blerkom, Malcolm L. 2009. *Measurement and statistics for teachers.* New York, London: Routhledge.

Brandtmeier, Cindy. 2005. “Effect of reader’s knowledge, text type, and test type on L1 and L2 reading comprehension in Spanish“. *Modern Language Journal* 89, 37-53.

Brown, James Dean. 1988. *Understanding research in second language learning: a teacher’s guide to statistics and research design.* Cambridge: Cambridge University Press.

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

Carr, Nathan Thomas. 2003. “An investigation into the structure of text characteristics and reader abilities in a test of second language reading“. PhD thesis, University of California, Los Angeles.

Carroll, J.B. 1961. “Fundamental considerations in testing for English language proficiency of foreign students.” *Testing*, 31-40.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Davies, Alan. 1990. *Principles of language testing.* Oxford: Blackwell.

Davies, Alan; et al. 1999. *Dictionary of language testing.* Cambridge: Cambridge University Press.

Derntl, Victoria. 2009. “Language testing in program evaluation”. MA thesis, University of Vienna.

Elamparo, Robyn L. 2005. “Multiple choice vs. constructed response: Does it really matter*?”* MA. thesis, California State University, Fullerton.

Elinor, S.-H. 1997. “Reading native and foreign language texts and tests: The case of Arabic and Hebrew native speakers reading L1 and English FL texts and tests“. Paper presented at the Language Testing Symposium, Ramat-Gan. Israel.(ERIC Document Reproduction Service No. ED 412746).

Freeman, James A. 2007. “An examination of the relationship between test scores, gender, ethnicity, attendance, and graduation”. PhD thesis, The Ohio State University.

Friedl-Lucyshyn, Gabriele. 2010. “Die standardisierte Reifeprüfung”. Presentation given at FDZ Vienna University, Vienna, 23.03.2010.

Geering, Margo. 1993. “Gender differences in Multiple Choice Assessment“. Submitted as part requirement for the degree of Master of Education. University of Canberra.

Gronlund, Norman E. 2003. *Assessment and Student Achievement*. (7th edition). Boston: AB Longman.

Hardcastle, Joanna G. 1991. “Gender differences and the effect of test format on student performance in mathematics”. MA thesis, Kean University.

Heaton, J.B. 1988. *Writing English Language Tests.* New York: Longman.

Hedge, Tricia. 2000. *Teaching and learning in the language classroom.* Oxford: Oxford University Press.

Henning, Grant. 1987. *A Guide to Language Testing.* Cambridge, Massachusetts: Newbury House Publishers.

Hughes, Arthur. 2003. *Testing for Language Testing.* Cambridge: Cambridge University Press.

In’nami, Yo; Koizumi, Rie. 2009. *“*A meta meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats*”. Language Testing* 2009 26 (2), 219-244

Kitao, Kathleen S.; Kitao,Kenji. 1999. *Essentials of English Language Testing.* Tokyo, Eichosha Co Ltd.

Kobayashi, Miyoko. 2002. “Method effects on reading comprehension test performance: text organization and response format.“ *Language Testing* 19 (2), p. 193-220.

Knoch, Ute. 2009. lecture notes “ Analysis of test quality“. The University of Melbourne, Sept- Dez 2009.

Lado, Robert. 1961. *Language Testing.* London: Longman.

LeLoup, Jean Willis.1993. ”The effect of interest level in selected text topics on second language reading comprehension“. Unpublished PhD thesis, The Ohio State University.

McNamara, Timothy F. 2000. *Language Testing.* Oxford: Oxford University Press.

McNamara, Timothy F. 2009. lecture notes “Assessing receptive skills (1): Reading”. Power point presentation held at the seminar “Language Testing”, The University of Melbourne, Sept-Dez 2009.

Morgan, George A., et al. 2004. *SPSS for Introductory Statistics : Use and Interpretation*. (Second Edition). Mahwah, New Jersey : Lawrence Erlbaum.

Oakhill, Jane V; Petrides, Alison. 2007. “Sex differences in the effects of interest on boys' and girls' reading comprehension.” British Journal of Psychology 98 (2), 223-235.

Oosterhof, Albert. 2009. *Developing and Using Classroom Assessment.* Upper Saddle River, New Jersey: Merrill.

Phakiti, Aek. 2003. “A Closer Look at Gender and Strategy Use in L2 reading“. *Language Learning* 53 (4), 649-702.

Popham, W. James. 2002. *Classroom Assessment: What teachers need to know.* (3rd edition). Boston: Allyn & Bacon.

Robison, Terry J. 1989. “Open-ended vs multiple-choice response formats: does format affect responses?”. MA thesis, University of Melbourne.

Schweinberger, Silvia. 2009. *Sprachen testen: Hören und Lesen.* Wien: Präsens Verlag.

Shehadeh, Adanan Muhammed. 1997. “The effect of test type on reading comprehension in English as a Foreign Language: The case of recall protocol and multiple choice“. PhD thesis, The Ohio State University.

Shohamy, Elana. 1984. “Does the testing method make a difference? The case of reading comprehension“. *Language Testing* 1(2), 147-170.

Spiel, Christiane. 2008. “Auswahlverfahren: Ziele, Anforderungen, Bespiele, Empfehlungen”. Presentation given at: Workshop “An der Schnittstelle zwischen Schule und Hochschule: Kompetenz, Eignung und Begabung auf dem Prüfstand” der Österreichischen Forschungsgemeinschaft, Mauerbach, 6 – 7 June 2008. http://www.oefg.at/text/veranstaltungen/schnittstelle/Beitrag_Spiel.pdf (14 January 2010).

Urquhart, A.H.; Weir, C.J. 1998. *Reading in a second language: process, product and practice.* Essex: Pearson Education Limited

Van Blerkom, Malcolm L. 2009. *Measurement and Statistics for Teachers.* New York and London: Routhledge.

Wolf, Darlene Faye. 1991. “The effects of task, language of assessment, and target language experience on foreign language learners performance on reading comprehension tests“. PhD thesis, University of Illinois at Urbana-Champaign.

Wolf, Darlene F. 1993. “A Comparison of assessment tasks used to measure FL reading comprehension“. *The Modern Language Journal* 77 (4), 473-489.

Yazdanpanah , Khatereh. 2007. “The effect of background knowledge and

reading comprehension test items on male and female performance“. *The Reading Matrix* 7 (2), 64-80.

Internet resources

Fulcher, Glenn; 2009. “What is language testing? A competition for winter 2009/10”. http://209.85.229.132/search?q=cache:I68_T92HnxAJ:languagetesting.info/whatis/lt.html+what+is+language+testing&cd=5&hl=de&ct=clnk&gl=at&lr=lang_de&client=firefox-a (5 February 2009).

McAlpine, Rachel. 2004. “Mc Alpine EFLAW readability formula”. http://www.webpagecontent.com/arc_archive/139/5/. (06 November 2010).

Universität Wien, 2009. “Student point – information on financial matters – tuition fees”. http://studieren.univie.ac.at/index.php?id=657 (23 October 2009).

“Free Readability Formulas Assessment: Free Readability Assessment Test”. http://www.readabilityformulas.com/free-readability-formula-assessment.php. (06 November 2009).

“Pearson Longman Exams”. http://www.pearsonlongman.de/main/main.asp?page=exams/bookdetails&ProductID=131710. (9 February 2010).

“SAT – The most widely used college admission exam”. http://sat.collegeboard.com/home. (21 February 2010).

“The Flesch reading ease readability formula”. http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.php. (06 November 2009).

“TOEFL testing”. http://www.testpreppractice.net/TOEFL/toefl-testing-1.aspx. (9 Februar 2010).

8. Appendix

8.1. Reading comprehension test

Reading Test	
Tuition fees gain allure in cash-hit European campuses Luke Harding in Berlin Monday October 13 *The Guardian*	
1	[...] Like many universities in Europe, Humboldt - founded in 1810 by the statesman Wilhelm von Humboldt - is overcrowded and under-funded. With German universities in crisis, and no help forthcoming from the government, vice-chancellors in Germany are now contemplating the previously unthinkable: tuition fees.
2	"We've got two choices. One of them is for Germany to become merely average. The other is for us to really invest in education and research," said Jürgen Mylnek, the Humboldt's president. "If the public sector isn't able to give us the money we need alternatives. In the mid to long-term there is no way round tuition fees."
3	[...] Five years after Britain introduced tuition fees, the rest of Europe is following suit. Holland, Austria, Italy, Spain and Portugal have all recently introduced tuition fees ranging from €600 to €1,450 a year. France has modest fees too; while in Germany a law that prevents them from being charged is now being challenged.
4	It is only in relatively affluent Scandinavian countries like Sweden, Finland and Denmark that the principle of free education has not received a battering. In Sweden there are no tuition fees. Nor is there any prospect of introducing them. [...] The funding system, which dates from the early 1990s, is generous.
5	The right and left in Sweden agree that tuition fees are a bad idea. "There may be some good arguments for having such a system but it is not on the agenda," Henrik von Sydow, a conservative MP said. "We

	don't want to have a system where students have to pay for higher education." [...] But in struggling Euro-zone countries like Germany - and to a lesser extent Holland and France - tuition fees are now on the agenda.
6	[...] The prospect of tuition fees has caused dismay among students, many of whom already work to make ends meet. Student union president Thomas Sieron said that fees would be a disaster.
7	The extra money would not be invested in universities; instead Berlin and other federal regions would simply cut higher education budgets even more, he said. Tuition fees would also deter students from poor backgrounds from going to university - an argument that student unions in Britain have deployed to little effect.
8	[...] Supporters of tuition fees, meanwhile, argue that fees would not only generate extra revenue for the hard-pressed higher education sector, but they might also encourage students to take their studies more seriously. In France anyone with a baccalaureat - France's A-levels - can in theory attend the university course of his or her choosing. This democratic, if impractical, principal creates serious problems of overcrowding and de-motivation, and of course a sky-high dropout rate.
9	[...] In Italy, dropout rates are also high. Anyone who obtains the secondary school certificate - the Italian equivalent of A-levels - has a right to go to university. But only 30% of Italian students graduate. "Italians have developed a habit of 'parking' themselves in universities while they make up their minds what to do with their lives," said Franco Pavoncello, a political scientist at John Cabot University in Rome.
10	[...] Most German students do not graduate until the age of 26. There are no fees, and little financial or institutional pressure for them to sit final exams; as a result, middle-aged students are commonplace. Sitting in Humboldt's student canteen, Jana Wendering - a 22-year-old law student - said she was in favour of tuition fees, provided hard-up students could get scholarships. "Fees might be an incentive for people

	to work a bit harder," she mused, over a plate of goulash.
11	[...] In Britain, of course, the argument has already moved on, with the education secretary, Charles Clarke, proposing top-up fees, which would see tuition fees rise from £1,025 a year to as much as £3,000. British fees are already the highest in Europe, followed by Holland, which charges its students €1,445 (£960) a year. But student funding in Holland is fairly generous: all Dutch students are entitled to a loan of €2,640 (£1,760) a year, which automatically becomes a gift or a grant if they subsequently meet certain minimum academic criteria, which most do.
12	The fear, among students in European countries where education is free, is that once tuition fees are introduced the cost of education will increase. University presidents admit tuition fees of, say, €1,000 a year will not be enough in the long run.
13	"The figure is too low. You can't fund a world-class university on €1,000," said Dieter Lenzen, the president of Berlin's Freie Universität, which was founded in 1948 in the American sector of Berlin. "How are we expected to compete with American universities which charge up to $28,000 a year? Colombia University has recently spent $145m on multi-media computers."
14	With fees now a reality across much of the EU, Britain does appear, for once, to be leading in Europe. But many students believe this is a dismal trend. "Just because Europe is moving in a certain direction doesn't mean this is the right direction," Colin Töck, of Germany's national union of students, lamented. [...]

TASK 1: Multiple Choice Questions

Tick the correct option(s) in the following Multiple Choice Tasks. Bear in mind that more than one statement will be correct in some tasks. If more than one statement is correct, the number of correct statements is given in brackets.

1. According to the text, Germany
   a. changed the laws in order to introduce fees.
   b. has made attempts to change the laws in order to introduce fees.
   c. has not made moves on changing the laws in order to introduce fees.
   d. successfully challenged the laws in order to introduce fees.

2. According to the text, in Sweden there are no fees (2)
   a. although they have been on the agenda.
   b. as the government does not want to charge students
   c. due to disagreement among the political parties.
   d. due to financial support from the government.

3. According to the text, fees in Germany would
   a. be invested in education funds.
   b. facilitate university access for disadvantaged students.
   c. force students to work.
   d. make students take their courses seriously.

4. According to the text, consequences of free university access are (2)
   a. high numbers of dropouts.
   b. many older students.
   c. motivation but overcrowding.
   d. too many graduates.

5. According to a student, scholarships should be given to students who
   a. are poorer.
   b. have best results.
   c. work during their studies.
   d. work hardest.

6. According to the text, Britain discusses
   a. fees of almost £2,000.
   b. fees of £1,025.
   c. increasing fees by almost £2,000.
   d. increasing fees by £3,000.

7. According to the text, Dutch students (2)
   a. get a loan which most have to pay back.
   b. get a loan which some may keep.
   c. pay the second highest fees in Europe.
   d. may keep the loan if they fulfil certain requirements.

TASK 2: True – False – Not given

Read through the statements 1-9. Are they "true" or "false"? If there is not enough information to answer, choose "not given".

1	The German government has long contemplated tuition fees.	□ True □ False □ Not given
2	If Germany wants to invest in education one possibility would be money from the state.	□ True □ False □ Not given
3	France has lower fees compared to other European countries.	□ True □ False □ Not given
4	Universities in Scandinavia are supported with enough money, which makes fees unnecessary.	□ True □ False □ Not given
5	In France students with better grades in their baccalaureat are privileged in their choice of university courses.	□ True □ False □ Not given
6	About two thirds of Italian students do not finish their studies.	□ True □ False □ Not given
7	In Germany scholarships are planned to be given to hard-working students.	□ True □ False □ Not given
8	The British education secretary has introduced fees of £3,000 a year.	□ True □ False □ Not given
9	European students who already have to pay fees are afraid of increasing expenses of education.	□ True □ False □ Not given

TASK 3: Matching

The following statements are summaries of the single paragraphs. Match the most appropriate statement to each of the paragraphs by indicating the letter of the statement next to the number of the paragraph in the grid. There are more statements than paragraphs but match only one statement to each paragraph! One example has already been done for you.

paragraph	statement	paragraph	statement
1		8	
2		9	
3		10	
4		11	
5		12	
6		13	
7		14	C

A	America as model for tuition fees
B	Amounts of tuition fees
C	Are tuition fees a step in the right direction?
D	Depressed about tuition fees
E	Drastic dropout rates
F	Anxious prospect of increasing tuition fees
G	Negative preview of tuition fees
H	No need for tuition fees
I	No wish for tuition fees
J	Tuition fees against dropout rates
K	Tuition fees against lazy students
L	Tuition fees against long-time studies
M	Tuition fees against old students
N	Tuition fees against overcrowded and under-funded universities
O	Tuition fees as a means of competition
P	Britain as forerunner
Q	Tuition fees too low to compete with other countries
R	Tuition fees to spend on universities

8.2. Questionnaire

BITTE ERST AUSFÜLLEN WENN DU MIT DEM LESETEST FERTIG BIST!

Feedback zum Lesetest

Um deine Meinung zum Lesetest zu erfahren, bitten ich dich, die folgenden Fragen zu behandeln. **Kreuze bitte an, welche Antwort am ehesten für dich zutrifft!**

Persönliche Informationen:

Geschlecht: ☐ männlich ☐ weiblich

Nationalität: _____

Muttersprache: _____

Muttersprache der Mutter: _____

Muttersprache des Vaters: _____

Sprachen die zu Hause gesprochen werden: _____

1. Wie gut findest du diese Art von Lesetest?

sehr gut	gut	weniger gut	schlecht

2. Wie interessant war der Text?

sehr interessant	eher interessant	eher uninteressant	uninteressant

3. Bitte ordne die drei Antwortformate (Multiple Choice Questions, True-False-Not Given, Matching) nach ihrer Schwierigkeit

sehr schwer	schwer	okay	leicht	sehr leicht

5. Bitte begründe kurz warum die drei Antwortformate (Multiple Choice Questions, True-False-Not Given, Matching) für dich sehr schwer, schwer, okay, leicht oder sehr leicht waren.

_____

_____

_____

_____

_____

_____

4. Wie findest du die Länge des Textes?

zu lang	lang	okay	kurz	sehr kurz

5. Wie war die optische Gestaltung des Textes / Layout / Übersichtlichkeit?

sehr gut	gut	weniger gut	schlecht

Weitere Kommentare:

8.3. Facility values and discrimination indices of all items

	FV	DI
mctask1_1	.46	.23
mctask1_2	.22	.42
mctask1_3	.50	.28
mctask1_4	.29	.33
mctask1_5	.20	-.04
mctask1_6	.25	.42
mctask1_7	.27	.56
tfngtask2_1	.45	.36
tfngtask2_2	.43	.04
tfngtask2_3	.24	.17
tfngtask2_4	.53	.49
tfngtask2_5	.24	.15
tfngtask2_6	.72	.32
tfngtask2_7	.38	.33
tfngtask2_8	.44	.45
tfngtask2_9	.35	.13
mtask3_1	.49	.35
mtask3_2	.24	.32
mtask3_3	.65	.54
mtask3_4	.66	.57
mtask3_5	.57	.50
mtask3_6	.27	.30
mtask3_7	.23	.48
mtask3_8	.16	.24
mtask3_9	.32	.26
mtask3_10	.24	.23
mtask3_11	.41	.45
mtask3_12	.32	.49
mtask3_13	.35	.33

8.4. Rasch analysis

```
Item    1: item 1                              Infit MNSQ = 1.28
                                                    Disc = 0.35

Categories          0 [0]      1 [1]      2 [2]      3 [3]      4[4]
missing

Count                   3          2         44          3         46
0
Percent (%)         3.1       2.0      44.9       3.1      46.9
Pt-Biserial      -0.13     -0.14     -0.24      0.04      0.31
Mean Ability      -0.60      -0.71      -0.31      -0.03       0.07
NA
StDev Ability      0.61       0.67       0.54       0.34       0.62
NA

Step Labels                  1         2         3         4

Thresholds                -2.06     -1.87     -0.15     -0.06
Error                      0.66      0.63      0.35      0.35
.....................................................................
................................................


Item    5: item 5                              Infit MNSQ = 1.44
                                               Disc = 0.08

Categories          0 [0]      1 [1]      2 [2]      3 [3]      4[4]
missing

Count                   2          5         66          6         19
0
Percent (%)         2.0       5.1      67.3       6.1      19.4
Pt-Biserial      -0.16      0.12     -0.11      0.12      0.04
Mean Ability      -0.78       0.16      -0.18       0.14      -0.09
NA
StDev Ability      0.19       0.40       0.61       0.57       0.65
NA

Step Labels                  1         2         3         4

Thresholds                -2.50     -1.99      0.57      0.77
Error                      0.73      0.64      0.40      0.39
.....................................................................................................
```

Curriculum Vitae

Persönliche Daten

Vorname:	Susanne Elisabeth
Zuname:	Hinterlehner
Geburtsdatum:	18.11.1985
Geburtsort:	Linz
Staatsangehörigkeit:	Österreich
Eltern	Dr. Hansjörg und Monika Hinterlehner
Familienstand:	ledig

Schulbildung

1992-1996	Volksschule Dr. Ernst Koref in Linz
1994-2001	Violinunterricht an der Musikschule der Stadt Linz
1995-2004	Klavierunterricht an der Musikschule der Stadt Linz
1996-2000	Akademisches Gymnasium Linz
2001-2004	Violinunterricht an der Anton Bruckner Privatuniversität
2000-2004	ORG der Diözese Linz unter besonderer Berücksichtigung der musikalischen Ausbildung: Matura

Studium

2004-2010	Universität Wien: Lehramtsstudium Englisch und Spanisch
09/2009-12/2009	Forschungsaufenthalt im Rahmen der Diplomarbeit an der University of Melbourne (Australien) bei Prof. Tim McNamara
09/2008-10/2008	Studienaufenthalt in Kanada im Rahmen einer interdisziplinären Lehrveranstaltung der Universität Wien
02/2008-07/2008	Erasmus Auslandsaufenthalt an der Universidade de Santiago de Compostela (Spanien)

Auslandsaufenthalte zu Studienzwecke

07/2005-08/2005	Academia La Rambla 85, Barcelona: Zertifikat C1 (mit Auszeichnung)
07/2006-08/2006	Churchill House, School of English Language, Ramsgate: advanced certificate (with honour)
07/2007-08/2007	Taronja School Valencia: Zertifikat C1+
02/2008-07/2008	Erasmus Auslandsaufenthalt an der Universidade de Santiago de Compostela
07/2008-08/2008	Au pair Aufenthalt in Sanxenxo (Spanien)
09/2008-10/2008	Studienaufenthalt in Kanada im Rahmen einer interdisziplinären Lehrveranstaltung des Literaturwissenschaftlichen Departments am Institut für Anglistik und Amerikanistik der Universität Wien

09/2009 – 12/2009	Forschungsaufenthalt im Rahmen der Diplomarbeit an der University of Melbourne bei Prof. Tim McNamara

Weitere Qualifikationen

Zertifikat English for Specific Purposes (ESP)

Relevante berufliche Erfahrungen

seit 2003	Englisch Nachhilfe
03/2006–07/2006	Pädagogisches Praktikum UF Englisch am Musikgymnasium Wien (bei Mag. Karl Eigenbauer)
10/2007–01/2008	Fachspezifisches Praktikum UF Englisch am Bundesgymnasium Wien 9, Wasagasse (bei Mag. Birgit Strasser)
10/2008–01/2009	Fachspezifisches Praktikum UF Spanisch am Bundesgymnasium und Bundesrealgymnasium GRG 19, Billrothstraße (bei Mag. Ingrid Hofbauer)
12/2008	Unterrichtsforschungs-Praktikum UF Englisch an der Praxishauptschule der KPH Wien/Strebersdorf (bei Mag. Monika Greiner)
seit 03/2009	Nachhilfetätigkeit (Englisch und Spanisch) im Institut Schülerhilfe Wien,20
seit 07/2009	Delegation Manager, Student Ambassador Programs People to People (USA) and PDM (Europe)