universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

Language testing in program evaluation – Evaluating the learning outcomes of two language tracks in an Austrian “Gymnasium”

Verfasserin

Victoria Derntl

angestrebter akademischer Grad

Magistra der Philosophie (Mag.phil.)

Wien, Juni 2009

Studienkennzahl lt. Studienblatt:	A 190 344 333
Studienrichtung lt. Studienblatt:	Englisch
Betreuerin:	Ao. Univ.-Prof. Christian Dalton-Puffer

ACKNOWLEDGEMENTS

Language testing is a fascinating field within applied linguistics and I am grateful for the opportunity to look into this subject in my diploma thesis. Throughout the writing of my thesis there were many people who supported me and without whom I would not have been able to finish my thesis.

First of all, I want to thank Professor Tim McNamara. With a brilliant seminar he raised my interest in language testing and supported me in the planning stages of my paper. He also contributed to my paper with statistical matters and he was always willing to listen to my problems and questions.

I also owe a deep debt to my supervisor Ao. Univ.-Prof. Christiane Dalton-Puffer for her help during the writing of this paper. Her comments and suggestions guided me in the process of writing and I am grateful for her support. Our collaboration started before my diploma thesis in the course of the language testing project in an Austrian "Gymnasium", which is the basis of this paper. Prof. Dalton-Puffer was a great project leader, always available, well-organised and diplomatic in the contact with the school. Her clear-sighted way made this project possible in the first place.

For the collaborative work on the language testing project I thank all the members of the group: Marion Binder, Marlene Heider, Hannes Loch and Silvia Schweinberger. We can all be very proud of the project and I am grateful for the mutual effort we made. My gratitude also goes to the Austrian "Gymnasium" in which the tests were conducted: in particular to the headmaster, the parents' association and, of course, the test takers for their time and understanding.

With all my heart I would like to thank Christine Alt for her wonderful friendship and mental support. In the course of our studies she was always one step ahead, encouraging and challenging me to work for my goals.

Finally, but certainly not least, my heartfelt thanks go to my family and my close friends. My parents Heinz and Elfriede, my brother Christian and my sister Stephanie have always been my caring support; understanding, loving and helping me in all situations. My friends Sunny, Stefan, Kati and Karol have always brought out the best in me and I am grateful for their friendship.

ZUSAMMENFASSUNG

Diese Diplomarbeit behandelt einen Sprachentest, der von einem österreichischen Gymnasium in Auftrag gegeben wurde, und mit dem Ziel der Sprachlehrgangsevaluierung entwickelt wurde. In diesem Fall wollte der Elternverein der Schule die beiden Sprachenzweige in Bezug auf ihre Lernerfolge vergleichen. Zu diesem Zweck wurden zwei Tests in zwei verschiedenen Sprachen (Englisch und Französisch) entwickelt. Die vorliegende Arbeit beinhaltet eine umfassende Diskussion aller theoretischen Aspekte des Testens, die im Zusammenhang mit den tatsächlichen Prozessen der Entwicklung, Ausführung und Bewertung des zu analysierenden Sprachentests behandelt werden.

In Bezug auf den Vergleich der beiden Sprachenzweige, auf dem aber nicht das Hauptaugenmerk dieser Arbeit liegt, kann gesagt werden, dass die Tests relativ signifikante Ergebnisse zeigten. Während die TestteilnehmerInnen beider Sprachenzweige sehr ähnliche Resultate im Englischtest erzielten, überbot ein Sprachenzweig den anderen in ihren Ergebnissen beim Französischtest. Qualitative und statistische Auswertungen der Sprachentests selbst wurden durchgeführt um zu festzustellen, ob die entwickelten Tests reliabel genug sind um für Vergleichszwecke sinnvoll und aussagekräftig zu sein. Aufgrund der Ergebnisse der Analysen kann schlussgefolgert werden, dass, obwohl beide Tests noch Verbesserungspotential haben, sie ein Reliabilitätsniveau aufweisen, das für Vergleichszwecke zufriedenstellend hoch ist. Die Testauswertungen sind reliabel genug, um auf ihnen basierende Interpretationen zu ermöglichen.

ABSTRACT

This thesis presents a language test which was commissioned by an Austrian "Gymnasium" and which was developed with the purpose of program evaluation. In this particular case the parents' association of the school wanted to compare the two existing language tracks of the school with regard to the learning outcomes of the pupils. Therefore two tests in two different languages (English and French) were designed. The paper features an in-depth discussion of all components of testing theory connecting them to the actual stages in the development, trialling, administration and scoring of the test under consideration.

The results of the tests, which are not the main focus of the paper, show that the test takers from the two language tracks performed almost identically in the English test, whereas one language track exceeded the other one in the performance in the French test. A qualitative and statistical analysis of the tests themselves was carried out in order to ascertain whether the tests developed can be regarded as useful for the purpose of comparison. Results of this analysis demonstrate that although both tests still have potential improvement, they do show a satisfying level of reliability for the purpose of comparison. The results from the analysis of the usefulness of the tests show that the test scores are reliable enough in order to allow for interpretations which are based on them.

TABLE OF CONTENTS

1 INTRODUCTION

In the same way that evaluation takes place in various professional fields it should also be a core element of the implementation and adaptation of language programs, in particular in school settings. In Austria the secondary stage of education, i.e. “Oberstufe”, is differentiated into various school types with varying instructional contents, goals and focuses (cf. http://www.bmukk.gv.at/medienpool/17147/develop_edu_04_07.pdf). But even within one type of school there are differences due to a certain freedom of specialisation and focus on the part of school authorities. This means that schools may develop different tracks which may be unique in the Austrian school system. In particular with these specific tracks it can be very important to evaluate their learning outcomes in order to gather information on their usefulness.

This background is exactly the starting point of the research assignment which is the subject matter of this paper. The parents’ association of an Austrian “Gymnasium” with two different language tracks asked for an empirical and independent evaluation of the learning outcomes of these two tracks. They differ with regard to extent and duration of English and French classes; i.e. one track starts with English as the first foreign language, the other one begins with French. The research question to be answered was whether or not the pupils of the two language tracks possess identical language abilities in both languages by the time of the “Matura”, i.e. the school leaving exam.

In order to answer this research question two independent language tests (one in English and one in French) were developed and administered with pupils from both language tracks in their final year. Test results were then reported to the people who commissioned the test in order to support their decisions on possible future changes with regard to the language tracks. At the time of writing a standardised “Matura” is being implemented in Austrian schools in answer to the increasing desire for national and international comparisons of language abilities. What has been reality in other countries for some time will only be compulsory in Austria from 2013/14 onwards. In order to be in line with these developments in the direction of standardised tests, the language tests for the Austrian “Gymnasium” relate to the CEFR, which will also be described in this paper.

However, the focus of this paper will not be on a thorough description of the test results but on an analysis of the test itself. Generally speaking, tests can only be

considered to provide valuable information when they are designed carefully, taking into account certain rules and principles of language testing. Only a well designed and developed test will show a satisfyingly high level of reliability, or usefulness, which allows for reliable interpretations to be made on the basis of the test results. Therefore the theoretical considerations which form the basis of reliable tests will be discussed in detail in this paper as they were indispensable for the test development. Furthermore, a statistical and qualitative analysis of the tests will be presented by which their usefulness can be measured. The results from these analyses will show if the reliability of the tests is high enough to justify the interpretation of test scores.

2 LANGUAGE TESTING IN PROGRAM EVALUATION

Language program evaluation may take numerous different forms. As the title of this paper suggests language testing can be one form of evidence on the basis of which decisions concerning language programs are made. The purpose of this chapter is to introduce the role of language tests for evaluation. This background is important as the test which will be described later in this paper was developed in the context and with the purpose of language program evaluation. It has to be said, however, that language program evaluation is a very broad field within applied linguistics which has been subject-matter to debates and developments in the last decades and can therefore not be discussed comprehensively in the context of this paper. In order to proceed with the discussion of language program evaluation in this chapter two central terms have to be defined: *program* and *evaluation*.

Program refers to "a series of courses linked with some common goal or end product. A language education program generally consists of a slate of courses designed to prepare students for some language-related endeavour" (Lynch 1996: 2). There are different types of preparation which may range from one single course to an intensive preparation of students to use language in a second language culture. In order to include a wide range of possible preparations Lynch (1996) provides a very broad definition of *program* by using the term "to refer to any instructional sequence" (Lynch 1996: 2). The Austrian "Gymnasium" in which the language tests were conducted can be said to correspond to Lynch's definition of a language program. In this particular case, the language program is an extended and intensive form of preparation with at least seven years of instruction in a foreign

language. The courses are linked with the common long term goal of enabling the pupils to communicate in a second language culture. In a more short-term sense the common goal may also be defined as enabling the pupils to pass the “Matura”, i.e. the school leaving exam.

According to Davies et al. (1999: 56), *evaluation* is “[t]he systematic gathering of information in order to make a decision. Within a language education programme, evaluation may be carried out to provide information about the programme to stakeholders […] and to make decisions about the future of the programme”. For the term *language program evaluation* a different definition is provided by Davies et al. (1999: 106), which is “[t]he assigning of value to an abstract entity such as a language program (or curriculum), often by means of comparison to other programs or curricula”. In this connection it is useful to distinguish *evaluation* from related terms such as *assessment* or *testing*. Lynch (1996:2) decides to

> differentiate *evaluation* from these other terms primarily on the basis of its scope and purpose. That is, evaluation can make use of assessment instruments (including tests), but it is not limited to such forms of information gathering. […] Evaluation is defined […] as the systematic attempt to gather information in order to make judgments or decisions (Lynch 1996: 2).

In a very similar way Brown (1996: 277) distinguishes between testing and evaluation by claiming that testing may be one source of information for evaluation but that it is not restricted to testing by itself. This means that the form of information can range from observation or interviews to pencil-and-paper tests (Lynch 1996: 2). Other sources of information may be “diaries, notes from meetings, [and] institutional records” (Brown 1996: 277). In the evaluation of the Austrian “Gymnasium” testing was used as the only source of information about the language program. Nevertheless, the term *evaluation* will be used in the course of this paper in order to account for the broad and variable concept. It can be argued that in the case of the Austrian “Gymnasium” decisions of stakeholders will not be based exclusively on the test results but may also rely on additional sources of information, such as interviews, questionnaires or grades from school leaving exams. The purpose of the evaluation corresponds to both definitions of Davies et al. stated above. Firstly, the tests were used to gather information about the program on the basis of which decisions about the future could be made. Secondly, they were used to assign value to a program in comparison to another one, i.e. in the case of the Austrian “Gymnasium” the two language tracks were compared to each other.

One main discussion in the field of language program evaluation has been "between advocates of positivistic, quantitative research methodology and advocates of naturalistic, qualitative research methodology" (Lynch 1996: 12f.), which is concerned with the basis for research and the form the evidence for evaluation takes. In the literature this debate has been called "the *qualitative-quantitative debate* (Reichardt and Cook 1979; Smith and Heshusius 1986; Howe 1988; Smith 1988), *paradigm wars* (Gage 1989), or *paradigm dialog* (Guba 1990)" (Lynch 1996: 9). An extended discussion of this topic can be read in Lynch (1996) but would go beyond the scope of this paper. One point to be made here is that this discussion reflects the concern about which form of evidence is acceptable in program evaluation. Apart from decisions about the form of evidence which will be used for program evaluation evaluators will also have to answer other questions which are all related to program evaluation: "What are the social and political factors that affect language learning and teaching? How can we best define and measure language abilities? What are the best research designs for our inquiry?" (Lynch 1996: 10). Due to the limited scope of this paper this chapter cannot answer all of the above stated questions. In the course of this paper questions two and three, in particular, will be discussed theoretically and with regard to the practical example of a language test. This chapter will serve as a basic overview of language tests' role in evaluation by describing two models of program evaluation. The first is the *context-adaptive* model (CAM), suggested by Lynch (1990), the second is the *systematic design of language curriculum*, which is an adaptation of the *systems approach* by Brown (1996).

The various forms of evaluation do not only differ from each other in form but also with regard to their purposes. Put differently, language programs are evaluated for different reasons, "be it motivated by an internal quest for program improvement or by an externally imposed requirement in order to justify program funding" (Lynch 1996: 2). Lynch (1996: 3) claims that evaluation has to be designed or adapted with regard to the specific education program which is being evaluated. He therefore suggested a model which fulfils this claim: the *context-adaptive model* (CAM) for language program evaluation. The main characteristic of the CAM is its adaptability: "it is meant to be a flexible, adaptable heuristic – a starting point for inquiry into language education programs that will constantly reshape and redefine itself, depending on the context of the program and the evaluation" (Lynch 1996: 3). By means of this model the crucial issues of language program evaluation will be outlined. The series of steps are shown in the following figure (cf. Figure 1).

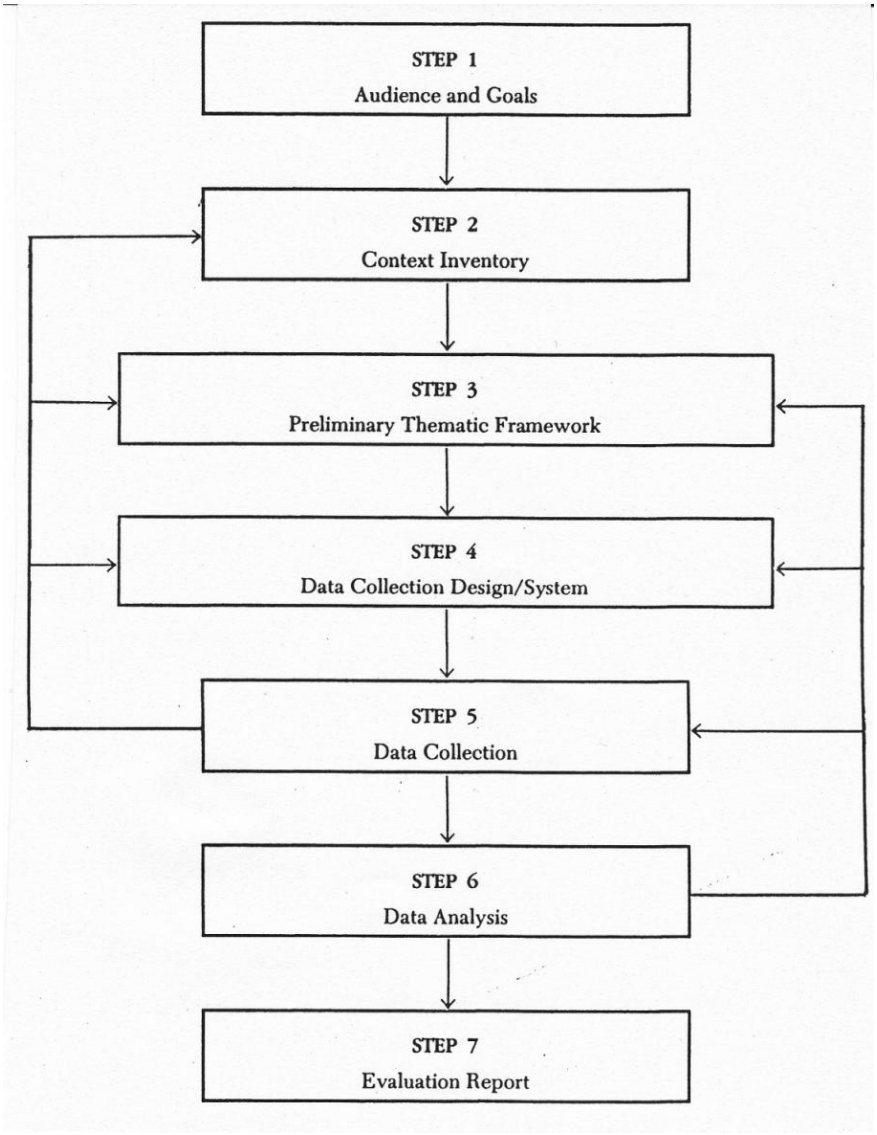

Figure 1: The Context-Adaptive Model for Program Evaluation (Lynch 1990a: 24)

The following description of the model is based on its discussion in Lynch (1990a: 24-39). The first step is to establish the audience and goals. For this step it is important to consider that the audience does not only include the people or organisations who commissioned the evaluation but also people who are more indirectly involved by having access to the evaluation report. The awareness of the range of the audience can be helpful to write an evaluation report which is relevant to all people involved. In relation to the audience the evaluation's goals should be identified. Depending on the audience these goals may vary and will also influence the form in which evidence will be gathered.

Once audience and goals have been identified the next step is a context inventory. This is the key stage in the CAM in which the characteristic aspects of the language teaching programs to be evaluated are collected. This context inventory includes 11 dimensions which will be listed subsequently. An extended discussion can be found in Lynch (1990a). The dimensions are

1. Availability of a *comparison group*
2. Availability of *reliable and valid measures* of language skills
3. Availability of various types of *evaluation expertise* (statistical analysis, naturalistic research)
4. The *timing* for the evaluation
5. The *selection process* of students into the program
6. The program *students*
7. The program *staff* (especially their availability, competence, and attitude toward the evaluation)
8. The *size and intensity* of the program (the number of students, classrooms, proficiency/course levels)
9. The *instructional materials and resources* available to the program
10. The *perspective and purpose* of the program
11. The *social and political climate* surrounding the program (Lynch 1990a: 27).

Although a description of each dimension is not possible within the scope of this paper one point to be made here is that the number and form of dimensions used to gather information on the program will always depend on the goals and the audience. Additionally, financial limitations may determine the extent of the context inventory step. The fact that financial issues are very likely to influence any step in the course of evaluation or testing will be discussed in section 5.6. Disregarding any limits to the context inventory it can be considered the “backbone of the evaluation” (Lynch 1990a: 31) on the basis of which decisions concerning data collection will be made.

The third step in the CAM is the development of a preliminary program-specific thematic framework. This framework is based on information from the context

inventory and includes important dimensions such as the students, the intensity and the purpose of the program as well as the status of the language. This information creates a picture of the program and its specific themes. Salient issues will determine the considerations of what is going to be tested and how the test is going to be carried out. The framework thus "provides a conceptualization of the program that will help guide the data collection and analysis phases of the evaluation" (Lynch 1990a: 32).

The next step, data collection design/system, makes use of information gathered during the first three steps in order to decide on the best possible form of data collection.

Following the design step, the fifth step involves data collection itself. Lynch (1990a) emphasises that during this step of the CAM revisions of the context inventory or the data collection system may be needed as new sources of data may present new important issues which have not been considered so far.

In the sixth step the data is analysed. Distinguishing between qualitative and quantitative analysis, Lynch (1990a) describes several forms which partly depend on the form of data collection, and in part also depend on available resources in terms of evaluators, time, etc.

The final step is the evaluation report which may take different forms with regard to formality and mode (written or oral) depending on the goals and audience as well as the context inventory. One crucial aspect during this stage of the CAM is the consideration of the political and social climate dimension which has already been referred to in step 2. "[D]ata and conclusions […] [may be] socially and politically sensitive […] [and] must be considered with extreme care" (Lynch 1990a: 39) and the evaluator must be aware of the fact that the intentions of the conclusions may be misinterpreted.

The context-adaptive model describes the potential procedure of program evaluation in a very flexible way as it does not confine the evaluation to a specific form. It thus allows for an adaptation according to different needs. In this paper it is impossible to discuss all forms of information gathering which evaluation may take. As this paper is concerned with the use of language testing for program evaluation this chapter intends to describe the place which language tests can take in the course of program evaluation. A thorough description of language test development will be given in a subsequent chapter (cf. chapter 9).

The potential of language tests for purposes of program evaluation can be seen in the following model taken from Brown (1996: 270). The model can be used for two purposes: either for the development of a program or for the improvement of an existing course. It is a simplified version of the widely accepted *systems approach*, which is "used in educational technology and curriculum design" (Brown 1996: 269). Brown's model (cf. Figure 2) is concerned with curriculum design and development and may at first sight appear to be very specific and therefore limited. For this paper, however, it is exactly this issue which is of particular importance as the language test which will be analysed was intended to be used with the purpose of program evaluation in view of curriculum development in mind.

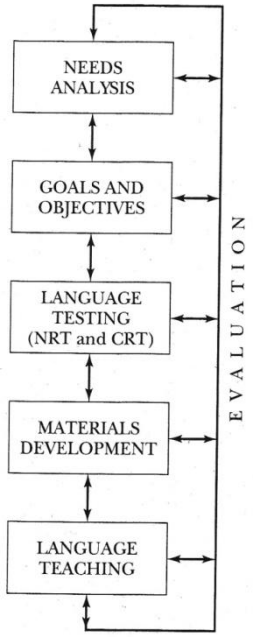

Figure 2: Systematic Design of Language Curriculum (Brown 1996: 271)

The figure (cf. Figure 2) indicates that language tests play an essential role in the course of language program evaluation. However, it also shows that they are not to be used on their own but that development and maintenance of programs is a circular and ongoing process (Brown 1996: 270). As Brown (1996: 269) puts it, "[a]lthough tests may be isolated for purposes of study, they should never be treated as though they are somehow divorced from the language teaching and learning processes that are going on in the same context".

The steps which are involved in this systems approach are "needs analysis, goals and objectives setting, testing […], materials development, teaching, and program

evaluation" (Brown 1996: 269). From the graphic representation it may be argued that program evaluation can be considered somewhat differently from the other steps involved as it seems to be the overall goal and every single step in the circle contributes to it. The following description of the systems approach is based on Brown (1996: 269-277) and is only intended as an overview as a close examination of the single steps would go beyond the scope of this paper. The step which is of major concern for this paper, language testing, will be discussed at length throughout this paper and will therefore be excluded from the discussion at present.

The first step – needs analysis – is concerned with the "identification and selection of the language forms that the target students are likely to require in actually using a particular language" (Brown 1996: 270). Brown suggests that with the above quoted definition two problems may arise. The first may result from the fact that learners or target students are not the only group of people involved in a language program. Teachers and administrators as well as people who are less directly involved such as later employers or institutions and in the broadest sense even whole societies have needs which should be taken into consideration in the course of the needs analysis. The second problem may arise from the solely linguistic focus of needs analysis. As students and their learning are not only influenced by linguistic factors but also by non-linguistic motivational and affective aspects any needs analysis should also gather information "about the students as people" (Brown 1996: 271). Therefore Brown provides a rather broad definition of needs analysis which seems to account for numerous aspects described above. According to him needs analysis is "the systematic collection and analysis of all relevant information that is necessary to satisfy the language learning needs of the students within the context of the particular institution(s)" (Brown 1996: 272). Under ideal conditions, without any financial burden and lack of trained personnel it would be highly eligible to fulfil all aspects suggested by this definition. It is to be doubted, however, if in reality needs analysis can always be carried out this way. Despite the question of feasibility his thoughts are worth taking into consideration in the course of the needs analysis.

In the second step of Brown's model goals and objectives of a program should be identified. By goals Brown (1996: 272) means "general statements of what must be accomplished in order to satisfy the students' needs. [...] Objectives, on the other hand, are statements of the exact content, knowledge, or skill that students must learn in order to achieve a given goal" (Brown 1996: 272). In the best case this second step is an outcome of the previous needs analysis and students' needs will be made to program goals.

The third step is concerned with language testing whereby tests should be developed on the basis of the program goals and objectives (Brown 1996: 273). The design, development or adaptation of a test is not always easy and its usefulness depends on various factors. As the following chapters will discuss at length selected aspects of test development, this step will not be explained further here.

The next step following language testing is materials development (Brown 1995: 274). Once needs analysis, identification of goals and objectives and finally, language tests have been carried out materials are designed in order to fit the students' needs. Of particular importance in this step are the test results gained in the previous step. The information from tests is very helpful to adopt and develop materials according to the defined goals and objectives of the course. However, all information from other components of the program, such as needs analysis and objectives will have effects on the development of materials as well.

For the final stage of language teaching only one aspect will be dealt with here. According to Brown (1996: 276) it can evoke a positive feeling on the side of the teachers to know that needs analysis has been carried out, goals and objectives have been identified and that they thus can base their teaching on an agreed framework of learning aims. For this positive feeling to be evoked it is a precondition that the above stated components have actually been carried out for the teachers rather than expecting teachers to do them on their own. "Needs, objectives, tests, and materials development should be group efforts, drawing on the collective expertise, time, talent, and energy, to do a more effective job of curriculum development than any one teacher could hope to do" (Brown 1996: 277).

This model offers a series of steps which may support the task of curriculum development or improvement and at the same time shows which role language testing can play in this process.

After this description of two different models of program evaluation, it is interesting at this point to discuss which steps were covered in the evaluation process of the Austrian "Gymnasium". It is apparent that several steps occur in both models. The first two steps from Brown, needs analysis and defining the goals correspond to the first two steps in Lynch's model, audience and goals and context inventory. Disregarding the different order of sequence, both models intend the same purpose with their first two steps. For the evaluation of the Austrian "Gymnasium" the main ideas of these two steps, i.e. defining the goals and examining the learning context,

were considered and covered. The definition of goals and objectives was based on the school curriculum and the level B2 of the Common European Framework of Reference (CEF/CEFR), which suggests certain skills and abilities at the time of school leaving exams[1]. This means that the testing of reading abilities and writing skills, etc. was designed according to a level of language ability which was to be expected from pupils in their final year. A context inventory, or needs analysis in the elaborate sense of Brown (1996) could not be carried out comprehensively. However, information on the context in which the tests were to be administered was gained by observations of language classes, by questionnaires about the learning and language background of the pupils and by gathering facts about the school, the language tracks and the general educational focus. Therefore it can be said that the two steps suggested in both models were covered.

In Lynch's model (1990a) the following two steps which include the development of a preliminary thematic framework and the designing of a data collection system are preparatory steps which precede the data collection. Only the second step was covered in the program evaluation of the "Gymnasium" as no explicit definition of a thematic framework took place. The form of data collection did not have to be discussed as it was required in the research assignment to use a language test as the source of information. However, the test format including item formats and the structure of the language tests had to be considered, which would correspond to the second of the preparatory phases. Brown's model does not suggest any preparatory stages but lists language testing as the next step. This is, of course, also present in Lynch's model in which it is called data collection. This step was covered by the test administration in the Austrian "Gymnasium". After this stage the two models differ significantly. Brown's focus is on curriculum development or improvement. Therefore the last two steps refer to the development of materials and to teaching. Lynch discusses the process of information gathering in more detail and suggests data analysis and reporting the results as the final two steps. In the context of the Austrian "Gymnasium" the evaluation process was structured according to Lynch's model. The data from the language tests were analysed and an evaluation report was provided for the people who commissioned the evaluation. Brown's focus on instructional implementations went beyond the scope of this particular language test and was not taken into consideration. Despite the extensive preparatory steps which precede the actual data collection in Lynch's model it can be said that this model of

[1] A detailed description of the CEFR can be found in chapter 6.

program evaluation provided a useful framework of evaluation guidelines. Due to its adaptive and more general structure it was relatively easily adapted to the specific needs of the program evaluation in the Austrian "Gymnasium".

The only step of both models which has not been discussed in this chapter is concerned with language tests. The reason for this is that the subsequent chapters will discuss language test development in detail. I will, however, begin with a brief examination of what tests cannot do by discussing limitations of tests.

3 MISCONCEPTIONS AND COMPETENCE IN LANGUAGE TESTING

"We believed that there was a model language test and a set of straightforward procedures – a recipe, if you will – that we could follow to create a test that would be the best one for our purposes and situations" (Bachman & Palmer 1996: 4). This belief, however, soon turned out to be unrealistic. Bachman and Palmer (1996: 6) speak from their own experience when they list misconceptions and unrealistic expectations language testers have about designing language tests. The first and most common misconception may be summarised under the belief that "language testers have some almost magical procedures and formulae for creating the 'best' test" (Bachman & Palmer 1996: 3). They, however, reject the idea of the existence of the one perfect test. The explanation for this will be seen in the following considerations of some problems that result from this misconception (cf. Table 1).

Misconceptions	**Resulting problems**
1 Believing that there is one 'best' test for any given situation.	1 Tests which are inappropriate for the test takers.
2 Misunderstanding the nature of language testing and language test development.	2 Tests which do not meet the specific needs of the test users.
3 Having unreasonable expectations about what language tests can do and what they should do.	3 Uninformed use of test or testing methods simply because they have become popular.
4 Placing blind faith in the technology of measurement.	4 Becoming frustrated when one is unable to find or develop the perfect test.
	5 Loss of faith in one's own capacity for developing tests appropriately, as well as feeling that only 'experts' can understand and do.
	6 Being placed in a situation of trying to defend the indefensible, since many students, as well as administrators, have unreasonable expectations.

Table 1: Misconceptions and Resulting Problems (Bachman & Palmer 1996: 7)

Assuming that there is one ‘best’ test which is used itself or used as a model for designing other tests, the replication of it may very likely result in the development of a test which turns out to be inappropriate in certain situations or for certain examinees (Bachman & Palmer 1996: 6). Misunderstanding of language test development and use may lead to uninformed and unqualified use of language tests. Unreasonable expectations and blind faith in the technology of measurement may very likely result in frustration and the loss of faith in one’s testing ability. The opinion that a test cannot be perfect is also expressed by Bachman (1990: 30) when he notes that “we know that our tests are not perfect indicators of the abilities we want to measure and that test results must always be interpreted with caution”. It is of particular importance to recognise the limitations of language tests and their restricted level of usefulness. The discussion of the properties of language tests with regard to their qualities in chapter 5 will show that tests are developed accepting compromises and accepting losses in one quality in favour of gains in another quality. One point to be made here is that language tests can never be perfect, and acceptable levels of reliability or validity are defined as a range of percentages, which reflects the notion that a language test which fulfils all criteria to 100 per cent is not realistic. It is not only the test development which cannot be perfect, but also the interpretation of test results. As language tests want to measure mental abilities it always has to be kept in mind that there are two limitations to this measurement which inevitably restrict the interpretations we make from test performance (Bachman 1990: 30). Bachman (1990) calls them limitations in specification, on the one hand, and limitations in observation and quantification, on the other hand.

The first limitation concerns the degree of specification of the language ability being tested. When a specific ability is tested this ability has to be defined on two levels (Bachman 1990: 31). On one level the ability has to be specified in order to contrast it other abilities and characteristics which may influence test performance. On the other level the relationship between the ability and the test score has to be identified, i.e. it has to be defined in what way this ability is scored and judged in the testing procedure. In theory this identification on both levels seems to be useful and necessary. Bachman (1990: 31), however, is aware that

> [i]n the face of the complexity of and the interrelationships among the factors that affect performance on language tests, we are forced to make certain

simplifying assumptions, or to *underspecify*, both in designing language tests, and in interpreting test scores.

This impossibility of taking into account all possible factors which may influence test performance resulting in an "indeterminacy in specifying what it is that our tests measure is a major consideration in both the development and use of language tests" (Bachman 1990: 32). This first limitation means that test developers have to be aware of the fact that interpretations of test scores will always be limited in their validity (Bachman 1990: 32).

The second limitation, which is called limitation in observation and quantification, applies on the level of interpretation of test performance (Bachman 1990: 32). It "derive[s] from the fact that all measures of mental ability are necessarily *indirect, incomplete, imprecise, subjective,* and *relative*" (Bachman 1990: 32). Measures are *indirect* as tests only indirectly indicate the underlying traits which are to be examined. They are *incomplete* as tests can only represent parts of a test taker's language ability. They are *imprecise* due to various characteristics of measurement scales and *subjective* due to decisions of individuals in test design, test items and test scoring. Measures have to be considered *relative* to an agreed norm as "the presence or absence of language abilities is impossible to define in an absolute sense" (Bachman 1990: 38). This is because neither of the ends of a spectrum from zero to perfect language competence can be identified (cf. Bachman 1990: 38f.).

After the description of common misconceptions and limitations of testing it seems to be useful to present strategies or competences which are necessary to avoid misconceptions and in particular the resulting problems. In order to be able to draw valid conclusions from test outcomes test developers have to be competent in test development and interpretation. Bachman and Palmer (1996: 7) suggest that the following competencies are involved in language testing.

- An understanding of the fundamental considerations what must be addressed at the start of any language testing effort, whether this involves the development of new tests or the selection of existing language tests;
- An understanding of the fundamental issues and concerns in the appropriate use of language tests;
- An understanding of the fundamental issues, approaches, and methods used in the measurement and evaluation;

- The ability to design, develop, evaluate and use language tests in ways that are appropriate for a given purpose, context, and group of test takers;
- The ability to critically read published research in language testing and information about published tests in order to make informed decisions.

The single steps cannot be discussed in detail here but selected issues will come up repeatedly in the discussions throughout the subsequent chapters of the paper. In particular, points three and four will be subject of the next chapter which will be concerned with the overall test usefulness.

4 TEST USEFULNESS

"The most important consideration in designing and developing a language test is the use for which it is intended, so that the most important quality of a test is its usefulness" (Bachman & Palmer 1996: 17). Having stated that test usefulness is the most important quality of a test there is, however, the remaining question of how it can be measured. It seems to be widely accepted that test usefulness comprises six qualities, namely reliability, validity, authenticity, interactiveness, impact and practicality, whose interrelationship can be illustrated as follows (cf. Figure 3).

Usefulness = Reliability + Construct validity +

Authenticity + Interactiveness + Impact + Practicality

Figure 3: Usefulness (Bachman and Palmer 1996: 18)

These six qualities will be discussed thoroughly in chapter 5. One important point to be made here is that they do not exist independently from each other but that they correspond, interrelate and complement each other. Therefore it is crucial

> that rather than emphasizing the tension among the different qualities, test developers need to recognize their complementarity. […] [T]est developers need to find an appropriate balance among these qualities, and that this will vary from one testing situation to another. This is because what constitutes an appropriate balance can be determined only by considering the different qualities in combination as they affect the overall usefulness of a particular test (Bachman & Palmer 1996: 18).

Whilst the existence of the six language test qualities seems to be generally accepted among linguists, Bachman and Palmer (1996: 18) add three principles

which emphasise the interrelated contribution of the qualities to the overall test usefulness.

Principle 1 It is the overall usefulness of the test that is to be maximized, rather than the individual qualities that affect usefulness.

Principle 2 The individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test.

Principle 3 Test usefulness and the appropriate balance among the different qualities cannot be prescribed in general, but must be determined for each specific testing situation.

Following these principles Bachman and Palmer (1996: 18) emphasise that "any given language test must be developed with a specific purpose, a particular group of test takers and a specific language use domain (i.e. situation or context in which the test taker will be using the language outside of the test itself) in mind".

The three considerations from the quote above will be discussed in the subsequent chapters with reference to the language test developed for the Austrian "Gymnasium". First, different test purposes will be outlined; the next section will deal with test takers and finally, the language use domain will be discussed.

4.1 TEST PURPOSE

Tests can be described and distinguished from each other in terms of their purpose or use (cf. Davies et al. 1999: 208). According to Davies (1990: 20) there are four main purposes of testing: "selection, feedback, evaluation and research". For the study reported here it is necessary to focus on the purpose of feedback. By this Davies means "*feedback* to the syllabus so that there is some of that external validity for language teaching which […] exists in language testing but not in language teaching" (Davies 1990: 20). This purpose of language testing is of particular importance in this paper as the study reported involved a language test developed for exactly this reason. Davies (1990) uses the term *test purpose* to refer to the four purposes above and distinguishes this concept from the related term *test*

use[2]. *Test use* then refers to different types of tests: McNamara (2000: 6), for instance, distinguishes between achievement and proficiency tests, while other sources define four types, namely the achievement test, the proficiency test, the aptitude test and the diagnostic test (*Dictionary of Language Testing* 2000*,* Davies 1990*)*. Although achievement and proficiency tests seem to be most common, all four test purposes will be described subsequently.

"An achievement test […] is an instrument designed to measure what a person has learned within or up to a given time" (Davies et al. 1999: 2). They are used in the context of instruction and should be applied supportive to teaching (McNamara 2000: 6). Achievement tests are normally carried out after a certain period of teaching, thus always relating to previous learning (Davies 1990: 20).

In contrast to achievement tests "*proficiency tests* look to the future situation of language use without necessarily any reference to the previous process of teaching" (McNamara 2000: 7). They rather examine language ability with regard to a specific real world purpose (Davies et al. 1990: 154) and thus try to predict future performance in the language (cf. Davies 1990: 20f.).

The third type, the *aptitude test*, also refers to future performance achievement, however, not relative to a specific purpose but with regard to language itself (Davies 1990: 21). This means that aptitude tests attempt to predict the general language learning ability of individuals (Davies et al. 1990: 10). Components of language aptitude are, for example, "phonetic coding ability […], grammatical sensitivity […], rote learning ability for new sound and meaning associations and inductive learning ability for language patterns" (Davies et al. 1990: 11).

The fourth test purpose is the *diagnostic test*. This type is "[u]sed to identify test takers' strengths and weaknesses, by testing what they know or do not know in a language, or what skills they have or do not have" (Davies et al. 1999: 43). Davies (1990: 21) puts it more drastically when he states that "the interest in the diagnostic test is in failure, what has gone wrong, in order to develop remedies". In this sense it may better be seen as an elicitation than as a test and should be used after achievement tests or proficiency tests have taken place. Davies et al. (1990: 43) admit that few tests are designed for purely diagnostic purposes as their development is rather difficult and laborious.

---

[2] Apart from Davies there seems to be an agreement on the terminology as *test purpose* and *test use* are used interchangeably to refer to some or all types stated above.

The test developed for the evaluation of the Austrian "Gymnasium" can be considered to be an achievement test. Although the two parts of the test (English and French) were not used supportive to teaching, they related to previous learning and were carried out after a certain period of instruction (in this case, after eight years of instruction in the first foreign language and after seven years of instruction in the second foreign language). The two language tests did not relate to previous learning content-wise but rather on a more general level of language ability in terms of certain language skills which the pupils had learnt in the course of their instruction.

Once the test purpose has been identified the next step involves the consideration of the test takers for which a language test is developed.

4.2 TEST TAKERS

There are characteristics inherent in language users or test takers which may influence test performance and should therefore be considered in test development (Bachman & Palmer 1996: 61). The main characteristic obviously is language ability which is what is being tested. Other characteristics which may have effects on test performance are, for instance, "language background, age, sex, educational background, background knowledge, affective reactions to test taking, level of proficiency in the target language and familiarity with the test method" (Davies et al. 1999: 208). Bachman and Palmer (1996: 61) suggest very similar characteristics which they categorise according to three components of individuals' characteristics: "personal characteristics, topical knowledge, and affective schemata". Although the list of characteristics suggested by Davies et al. (1999) seems to be a very useful starting point for the consideration and choice of test takers the concept of Bachman and Palmer has the advantage of its clear categorisation. Therefore the following exposition will describe the three categories with regard to the test takers of the study presented in this paper. For the development of the test for the Austrian "Gymnasium" all characteristics of the three categories could – at least partly – be considered.

The first category of personal characteristics involves attributes which may affect test takers' performance on language tests without being part of their language ability (Bachman & Palmer 1996: 64). According to Cohen (1994) these attributes are: age, foreign language aptitude, socio-psychological factory, personality,

cognitive style, language use strategies, ethnolinguistic factors, and multilingual ability (Cohen 1994: 74 quoted in Bachman & Palmer 1996: 64). Bachman and Palmer (1996: 65) reduce this list to what they call "a starting place for describing the characteristics of test takers", comprising age, sex, nationality, resident status, native language, level and type of education and finally type and amount of preparation or prior experience with a given test. For the test development of the study presented Bachman and Palmer's list was very useful. Test takers were asked to fill in questionnaires concerning their age, sex, nationality, native language, other spoken languages, and the level and type of education with focus on foreign language instruction. For this particular language test the information from the questionnaires was not mainly used for a selection of the examinees as all pupils in their final year were tested in the skills of listening, reading and writing. For the speaking test, however, some pupils from both language tracks were selected for a group discussion. For this selection process, information on sex and native language was important as the choice of test takers was attempted to guarantee equal conditions for both language tracks. Besides, the information from the questionnaires was valuable for the analysis of the test results[3].

The second category by Bachman and Palmer (1996) refers to the topical knowledge of individual test takers. Topical knowledge may also be called knowledge-schemata or real-world knowledge (Bachman & Palmer 1996: 65). It provides the basic information which enables language users to communicate referring to the cultural and social context in which they live. For test development topical knowledge is important in so far as test tasks which demand certain topical knowledge are bound to be easier for those test takers who have this knowledge, and more difficult for those without. The topical knowledge of all test takers of the Austrian "Gymnasium" was not examined explicitly. However, with regard to the cultural and social context it can be claimed that all test takers shared very similar real-world knowledge as they all lived in Austria and were educated in the same school system.

The third component of Bachman and Palmer's categorisation relates to affective schemata. They "can be thought of as the affective or emotional correlates of topical knowledge" (Bachman & Palmer 1996: 65) and determine how test takers react to

[3] The questionnaires showed, for instance, that the pupils from the French-track had spent more time in French-speaking countries which may have affected their listening and speaking performances. It was also interesting to see if native speakers (of French or English) were among the test takers and in how far their presence improved the test results of the language track they attended.

certain tasks depending on previous experiences in similar situations. Affective schemata may influence test takers' completion of tasks negatively or positively. If, for example, an emotionally charged topic is to be dealt with test takers' use of their language knowledge may be limited due to the affective responses to the topic (Bachman & Palmer 1996: 66). However, affective schemata can also be positively influential on test performance. Tasks with "controversial topics may stimulate some individuals to perform at a high level, precisely because they feel strongly about the topic" (Bachman & Palmer 1996: 66). This means that affective schemata as personal characteristic which may influence test performance is not completely bad or good. Emotionally charged topics should not be totally avoided in test design. The effects they may have on results should simply be considered and test developers should be aware of their potential influence. For the test development in the Austrian "Gymnasium" this category was taken into consideration for the choice of topics. Emotionally charged ethical or religious topics were not used in any part of the test (neither in the listening or reading parts, nor in the writing or speaking parts). Controversial topics were used for the writing and the speaking tests in order to stimulate test takers' responses[4].

Both Bachman (1990: 114) and Brown (1996: 190) add 'test-wiseness' as a further personal characteristic which may influence test performance. It relates to the test takers' familiarity with certain test methods. Experience with tests should in principle have positive effects on the test-taking process as test takers' performance is not negatively affected by unfamiliarity with the test format (Bachman 1990: 114). This consideration could, unfortunately, not be taken into account for the test development for the Austrian "Gymnasium". The choice of test formats, as described in section 9.2, was made with regard to considerations of reliability, validity and practicality and with regard to standardised testing procedures. Although familiarity with test formats very likely affects test performance this aspect was not taken into consideration due to practical restrictions[5].

In addition to test purpose and test takers, the third aspect to identify for language test development is the specific language use domain.

[4] As to be seen in the appendix (cf. section 15.1), one topic of the writing test was the introduction of tuition fees; the topic of the speaking task was the implementation of a smoking ban in Austria.
[5] Familiarity of test takers with test formats will be discussed further in section 5.2.4.

4.3 LANGUAGE USE DOMAIN AND LANGUAGE USE TASK

In language testing the main aim should not be a mere measurement of language ability which is completely separate and independent of any particular language domain. It can also not be the purpose of a language test to make general inferences about "all language use domains" (Bachman & Palmer 1996: 44) but language testing rather aims at making inferences about examinees' language ability in a specific target language use domain. This is, however, often ignored in the public discourse of testing.

Bachman and Palmer (1996: 44) define "a *target language use domain* as a set of specific language use tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to generalize". They refer to two general target language use (TLU) domains (Bachman & Palmer 1996: 44). The first type comprises 'real-life' domains with language use for communicative purposes and is referred to as *real-life domain*. The second type includes language use in situations of teaching and learning and is called *language instruction domain*. Although Bachman and Palmer (1996: 60) distinguish between these two domains they are aware of the fact that they are not necessarily separate as also the teaching and learning language has in recent years adopted a communicative dimension. In a similar manner McNamara (2000: 25) claims that "the test domain is typically defined in one of two ways. It can be defined operationally, as a set of practical, real-world tasks". This first description can be compared to the real-life domain in Bachman and Palmer's terminology. "Alternatively, the domain can be defined in terms of a more abstract construct, for example, in terms of a theory of the components of knowledge and ability that underlie performance in the domain" (McNamara 2000: 25). This second description may be seen as the equivalent to what Bachman and Palmer called language instruction domain.

Once the TLU domain of a particular test has been defined the tasks should be designed within the defined domain as they may have effects on test performance in general. Considerable amounts of research have been done in order to find out about the effects which test tasks have on test results (for example, Bachman and Palmer 1982, Shohamy 1984, Alderson and Urquhart 1985). It seems to be generally agreed that the inferences which are drawn from tests are not only about language ability but also about the effect of test tasks. Bachman and Palmer (1996: 46) conclude that "since we cannot totally eliminate the effects of task

characteristics, we must learn to understand them and to control them so as to insure that the test we use will have the qualities we desire and are appropriate for the uses for which they are intended".

In a similar manner, Bachman (1990: 111) states that

> characteristics, or 'facets' of test methods constitute the 'how' of language testing, and are of particular importance for designing, developing, and using language tests, since it is these over which we potentially have some control.

These test method facets consist of five categories: "(1) the testing environment; (2) the test rubric; (3) the nature of the input the test taker receives; (4) the nature of the expected response to that input, and (5) the relationship between input and response" (Bachman 1990: 118). These categories can be controlled by the test developers and are therefore of great importance. For the development of test tasks with the desired qualities it is essential to define test task characteristics' impact on language use and test performance (Bachman & Palmer 1996: 46). According to Hymes (1972 quoted in Bachman & Palmer 1996: 112) language use is mainly determined by contextual components such as the interlocutors, the setting and the subject matter. In the same way it can be said that characteristics of test methods or test tasks determine language use and can therefore be considered as contextual components in the sense of Hymes (Bachman 1990: 112). One ambition for test developers should be to design test tasks so that they resemble the characteristics of language use contexts outside the language testing context as much as possible.

Bachman and Palmer (1996: 45) summarise the importance of the adaptation of tasks according to the language use context, i.e. target use domain as follows:

> A language test can be thought of as a procedure for eliciting instances of language use from which inferences can be made about an individual's language ability. It therefore follows that in order for such inferences to be made, a language test should consist of language use tasks.

Bachman and Palmer (1996: 44) define a *language use task* as "an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation". In this way it constitutes the elemental components of language use. Vice versa, language use can be seen as a series of interrelated language use tasks. In order for test tasks to allow for any inferences to be made from them they must be designed in a way that they resemble the TLU tasks with regard to their distinguishing features (Bachman & Palmer 1996: 45).

Bachman (1990: 112) emphasises the importance of the correspondence between tasks and language use context by referring to the concept of 'authenticity'.

> The particular way test method facets are designed and controlled, and the correspondence between these and the features of language use contexts will have a direct effect on the 'authenticity' of the test and test tasks. In general, one would expect that the closer the correspondence between the characteristics of the test method and the essential features of language use contexts, the more 'authentic' the test task will be for test takers (Bachman 1990: 112).

The correspondence between test tasks and the target language use domain as indicated in the quote above is of great importance for test development and will therefore be considered further in the section on authenticity (cf. section 5.3).

As test tasks may vary significantly even within one type of test tasks a complete framework of test task description cannot and is also not necessary to be provided here. The intention of this section was to emphasise the relation between test tasks and TLU tasks. This relation partly affects the qualities of language tests, in particular, the authenticity of test tasks and the validity of inferences which are made on the basis of test performance (Bachman & Palmer 1996: 43).

5 SIX QUALITIES OF LANGUAGE TESTS

This chapter is concerned with the six qualities of language tests, which have been referred to in chapter 4. These are validity, reliability, authenticity, practicality, impact and interactiveness. The comparison of different scholars' views on the qualities and, in particular, the extent to which they are discussed shows that although the six qualities are widely accepted, different linguists put different emphasis on them.

Bachman and Palmer (1996: 19) take the point of view that test usefulness has to be considered in a systemic way, "considering tests as part of a larger societal or educational context". This means that tests are not the only components in an educational context and have to be considered among other components, "such as teaching materials and learning activities" (Bachman & Palmer 1996: 19). The difference between tests and other components in an educational context is their purpose. While a test can be said to mainly aim at measuring, other components have a rather promoting purpose. With regard to their different purposes the six qualities of tests are therefore categorised as follows by Bachman and Palmer

(1996: 19): four principles are shared by tests and other components of instructional programs while two qualities are mainly critical for tests as they justify their usefulness and validity. The four qualities shared by all components of instructional programs are: authenticity, interactiveness, impact as well as practicality. The two qualities which are of particular importance for tests as they "provide the major justification for using test scores – numbers – as a basis for making inferences or decisions" (Bachman & Palmer 1996: 19) are validity and reliability. It can be seen from their emphasis that reliability and validity seem to be of greater importance and will therefore also be discussed more extensively in this paper.

5.1 (CONSTRUCT) VALIDITY

Hughes (2003: 26) defines validity as follows: "A test is said to be valid if it measures accurately what it is intended to measure". Bachman and Palmer (1996: 21) provide a more elaborate definition by stating that "validity pertains to the meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores". Justification of the interpretation depends on the evidence that the scores indicate the abilities which are to be tested. For this evidence to be made the test construct has to be defined. Construct in this sense means "the specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task" (Bachman & Palmer 1996: 21). Examples of these constructs are 'reading ability', 'fluency in speaking' and 'control of grammar' (Hughes 200: 26). Following the definitions of *validity* and *construct* Bachman and Palmer (1996: 21) use the term *construct validity* to describe the extent to which interpretations of test scores reflect the construct which is to be measured. In a similar manner, Davies et al. (1999: 33) define "[t]he construct validity of a language test [...] [as] an indication of how representative it is of an underlying theory of language learning". Bachman (1990: 236) states that the traditional classification of validity into content, criterion and construct validity has been abandoned in favour of a concept of validity which comprises all three aspects. Hughes (2003: 26) supports the view that construct validity is used as a term to describe the general concept of validity. Also Messick (1988: 13 quoted in Bachman 1990: 236) similarly describes validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores".

It seems obvious that the assertion that a test has validity is not enough on its own. Validity has to be confirmed by empirical evidence which may take different forms (Hughes 2003: 26). Following Messick's concept of validity as an integrated evaluative judgement, none of the following evidences should be considered sufficient in itself, but only the collective consideration of all types mentioned above can demonstrate validity (Bachman 1990: 237).

5.1.1 CONTENT VALIDITY

Content validity is the "conceptual or non-statistical validity based on a systematic analysis of the test content to determine whether it includes an adequate sample of the target domain to be measured" (Davies et al. 1999: 34). Hughes (2003: 26) considers a test to have content validity if the language skills and structures which are to be tested are represented in the test content. What the relevant structures for each test are has to be defined and specified depending on the purpose of the test. The relevant structures are written down in test specifications which should be made at an early stage of test construction. The extent of content validity can be judged on the basis of a comparison of test content and test specifications (Hughes 2003: 27). Hughes emphasises the importance of content validity by giving two reasons. First, content validity affects the extent of construct validity: the greater the former, the more likely the latter as content validity can ensure that a test measures what it intends to measure. When main areas which were identified in the specifications are not represented or under-represented in the test, it means that the relevant content which was defined in advance is not represented in the test. Therefore this test may be unlikely to be accurate. The second reason is that an inaccurate test as described above is very likely to have a harmful backwash effect[6]. This means that areas which are not tested in the test are likely to be ignored in teaching and learning. In order to avoid harmful washback effects test specifications should be written at an early stage and content validation, i.e. an analysis of the content validity of a test, should be carried out before the test is being used (Hughes 2003: 27).

[6] The concept of backwash will be discussed in section 5.5.

5.1.2 CRITERION-RELATED VALIDITY

Hughes (2003: 26) uses the term criterion-related validity to refer to another form of evidence of a test's construct validity. Bachman (1990: 248) uses a slightly different terminology and calls this evidence *criterion relatedness* or *criterion validity*. As all three terms refer to the same concept they will be used interchangeably in this paper. Criterion-related validity

> demonstrates a relationship between test scores and some criterion […] [which] may be level of ability as defined by group membership, individuals' performance on another test of the ability in question, or their relative success in performing some task that involves this ability (Bachman 1990: 248).

Hughes (2003: 27) defines criterion-related validity similarly by saying that it

> relates to 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).

Davies et al. (1999: 39) state that "[c]riterion-related validity (which incorporates concurrent and predictive validities) of a new test is established statistically (using correlation) in terms of the closeness of a test to its criterion". As claimed in this quote there are two kinds of criterion-related validity: *concurrent validity* and *predictive validity*. The former refers to cases in which the test administration occurs at about the same time as the criterion behaviour; the latter means cases in which one wants to predict future behaviour (Bachman 1990: 248).

Concurrent validity

Concurrent validity "is concerned with the relationship between what is measured by a test (usually a newly developed test) and another existing criterion measure" (Davies et al. 1999: 30). According to Bachman (1990: 248), information of this kind can typically take two forms:

(1) "examining differences in test performance among groups of individuals at different levels of language ability"

(2) "examining correlations among various measures of a given ability"

Hughes (2003: 27) illustrates the concept of concurrent validity with the following example which refers back to form (2) described above: An oral exam of 45 minutes is valid to test what has been specified before. Due to reasons of practicality there are only 10 minutes available. A sample of test-takers should be evaluated during the 45 minutes session; the same sample should then take the 10 minutes test. The scores on both tests are then correlated: If the results of the two versions show a relatively high level of correspondence the shorter test may be regarded as valid. Both versions are then said to have concurrent validity (Davies et al. 1999: 30). On the contrary, if the comparison reveals a low level of agreement the validity of the shorter version cannot be said to be satisfactory (Hughes 2003: 28). The comparison of sets of scores is done in standard procedures, which generate the 'correlation coefficient' (or 'validity coefficient' when validity is measured), which is the mathematical measure of similarity (Hughes 2003: 28). The procedure shall not be of concern here, but it is important that the level of agreement, the so called coefficient can vary from 0 to 1, 1 being the perfect agreement. The level of agreement which is regarded acceptable depends on the purpose of the test and may vary. Having illustrated the coefficient with the example above, it has to be said that the length of a test is not a necessarily significant criterion for a higher level of concurrent validation (Hughes 2003: 29).

One serious problem with concurrent validity is that it presupposes that the criterion behaviour "can be validly interpreted as an indicator of the ability in question" (Bachman 1990: 249). This does not necessarily have to be the case. Strictly speaking, the criterion behaviour would have to be examined according to its validity itself, in order to avoid the correlation between the criterion behaviour and the test to be the only justification of both their validity (Bachman 1990: 249).

Predictive validity

The second form of content-related validity is called predictive validity or predictive utility (Bachman 1990: 250). "This concerns the degree to which a test can predict candidates' future performance" (Hughes 2003: 29). The criterion measure can then be the outcome, e.g. failing or passing an exam, or predictions of students' grades in a language course (Brown 1996: 248). If course grades and test scores were then to be compared, the correlation coefficient between them would indicate how good the predictive validity of the test is. In many cases, however, the correspondence between the prediction and the actual performance is rather low (in particular when

failing or passing an exam is the criterion measure). Then the expected validity coefficient will be around .4, i.e. 20 per cent concurrence (Hughes 2003: 30). This low level of agreement between the predicted and the actual performance is partly due to various factors other than linguistic ability which contribute to an outcome (cf. Hughes 2003: 29f.). Bachman (1990: 250) discusses some major problems of predictive utility. His main concern is that if it is examined alone the question of what is being tested can to a large extent be ignored. This means that this ignorance of what is being measured is often justified by the relatively satisfying level of prediction (Bachman 1990: 251). To illustrate this with an example: it is common practice to use multiple-choice items to test grammar in order to place test takers into writing programs. Obviously, multiple choice items cannot thoroughly measure one's ability in writing to a satisfying extent. It is not that they test a completely different ability, but rather that they measure "only limited aspects of the criterion ability" (Bachman 1990: 251). This is the reason why mismatches in predicting placement for individuals are relatively frequent. In most cases, however, wrong placement can quite easily be corrected. Thus the practice of using multiple-choice items to test writing abilities is justified with the satisfyingly high level of correct placement prediction and by another economical reason. "In situations where predictive utility is the primary consideration, there is a tendency to simplify, to reduce the number of measures we use to the smallest set, or to the single measure that provides the greatest accuracy of prediction" (Bachman 1990: 251). This happens for reasons of practicality and economy (Practicality will be discussed in section 5.6).

Summarising the most important restrictions to predictive utility, it can be said that there are two main problems which should be considered: it seems impossible

1) "to identify and measure all the abilities and factors that are relevant to the criterion" (Bachman 1990: 252)

2) "to specify not only whether the predictors are related to each other and to the predicted behavior, but also the strength and type of these relationships" (Bachman 1990: 252)

The second restriction refers to individual development as described above, i.e. the question of if and how motivation, aptitude and language ability interact and influence each other as well as the predicted behaviour (Bachman 1990: 252). These insufficiencies which have to be faced in prediction in general are

problematic when predictive validity is used to define the overall validity of a test. If considered separately, both concurrent and predictive validity are insufficient to describe the validity of a test for reasons which have been outlined in the last two sections. They are, however, components which –combined with other methods to define validity – should be taken into consideration in order to identify the overall validity of a test.

All three validities – content, concurrent and predictive validity – play a role and have to play a role in the development of the test. They "provide evidence for its overall, or construct validity. However, they are not the only source of evidence" (Hughes 2003: 30). This means that – given the assumption that a test fulfils the requirements of content validity and criterion-related validity – one can still not be sure if the items used in the test are actually measuring what they are supposed to measure, i.e. if the test has overall construct validity (Hughes 2003: 31). One example of a construct would be the ability to guess the meaning of unknown vocabulary from the context. Only by means of research it can be investigated whether this specific ability exists, whether it can be measured and whether a given test does measure this distinct ability. Without this evidence it cannot be said that the part of a test which attempts to measure certain abilities has construct validity. The next step will lead to the question if the whole test has construct validity when there is no confirming evidence from research. There are at least two ways of obtaining this evidence by trying to find out what test takers actually do when they answer specific items. One is called *think aloud*, the other one is *retrospection* (Hughes 2003: 32).

In the former method examinees speak out what they think while answering test items. In the latter the test takers try to recollect the thoughts they had while answering the item. The drawback to think aloud is that the voicing of thoughts may change the response which would have been given without thinking aloud. The disadvantage of retrospection is that test takers may forget or misremember thoughts. Despite these weaknesses these two methods and such research in general can give insights into how items work and what they actually test.

5.1.3 FACE VALIDITY

The term face validity is used when a test appears to measure what it is said to measure (Hughes 2003: 33). Despite the fact that it is no scientific concept it can be

relatively important. A test which is supposed to measure pronunciation but does not require test takers to speak may be considered to lack face validity. It may result in the test not being used or if it is used it may mean that test takers' performance on the test does not reflect their abilities. It is interesting to observe that face validity is discussed by Hughes (2003) in his recent book. Years earlier, the term and the concept of face validity were heavily criticised by different researchers (for example: Cattell 1964, Cronbach 1984 and Stevenson 1981, 1982, 1985). Also Bachman (1990: 287) has on different occasions pointed out the problematic issue of judging tests by their 'test appeal'. Nevertheless, the term, as well as the concept, continues to be used in the field of language testing. In the study reported in this paper it was interesting to observe that the people who commissioned the evaluation explicitly asked for a language test as means of evaluation. It seems that tests by themselves have a very good reputation and thus have face validity. If the tasks and the contents of the test do appear to measure the relevant skills and structures, clients may very likely be satisfied with the specific test. Thus, face validity seems to be particularly important to people who are no experts in the field of language testing. Having said this, face validity is often regarded as a non-empirical judgement without any theoretical basis and will therefore not be discussed further in this paper.

5.1.4 HOW TO MAKE TESTS MORE VALID

In the field of high stakes tests test designers are obliged to carry out a "full validation exercise before the test becomes operational" (Hughes 2003: 33). In the case of tests created by teachers or students Hughes (2003: 33) gives four pieces of advice:

1. Explicit specifications should be written before designing the test. A representative sample of the content of these should be included in the test.
2. Direct testing should be applied whenever feasible.
3. The scoring should be valid and relate directly to what is being measured.
4. The test must be as reliable as possible. Without reliability a test cannot be valid.

The first point has already been discussed in section 5.1.1 and will therefore not be repeated here.

The second point refers to direct testing. This means that a particular language ability should be tested with a test format which appropriately represents this ability: for instance, writing skills should be tested by means of a written composition, and not by multiple-choice items which require test takers to choose the correct tense or grammar form of a sentence.

The third point of Hughes' advice relates to validity in scoring. Validity is not restricted to the development of tests. Also the scoring of tests has to be valid as invalid scoring may easily invalidate carefully and validly designed test items. In order to achieve valid scoring, only one ability should be measured with one item. When more than one ability is measured the measurement becomes less accurate (Hughes 2003: 33). To explain this with an example, in the case of short-answer questions the answers should be scored according to their contents and should not judge grammar or spelling mistakes. This would be another ability. If it is scored, the test takers should be instructed accordingly.

As indicated in the last point, reliability is of great importance in connection with validity as a test cannot be valid without reliability. The next section will therefore deal with reliability and the interrelationship between these two language test qualities.

5.2 RELIABILITY

Reliability can be "defined as consistency of measurement" (Bachman & Palmer 1996: 19) This consistency or agreement relates to three aspects all of which are covered by Anastasi's (1997) definition: According to her reliability is "[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).

This means that firstly, the scores obtained by the same group of individuals on the same test on different occasions should not differ other than incidentally (Bachman & Palmer 1996: 20). Secondly, the scores obtained by the same group of test takers on two different tests with equivalent test tasks should not differ significantly either. Thirdly, the scores obtained by the same group of individuals who take the same test under different examining conditions, i.e. for example, a different setting, should not differ due to the variable conditions. Reliability thus "indicates the extent to

which individual differences in test scores are attributable to 'true' differences in the characteristics under consideration and the extent to which they are attributable to chance errors" (Anastasi 1997: 84). Chance error or measurement error "is a term that describes the variance in scores on a test that is not directly related to the purpose of the test" (Brown 1996: 188).

If scores in the three situations described above do differ significantly – i.e. are inconsistent – the test or tests will be said to reflect the abilities at test unreliably (Bachman & Palmer 1996: 20). The purpose and advantage of reliability is to find out if diverging test scores reflect the individuals' different language abilities or if the test itself and surrounding circumstances are the reasons for varying test scores. Thus "measures of test reliability make it possible to estimate what proportion of the total variance of test scores is *error variance*" (Anastasi 1997: 84).

For test developers reliability is an essential quality of language tests. If test scores are not reliable they cannot be used to draw any inferences on the ability which is to be measured by a test (Bachman & Palmer 1996: 20). It is important, however, to know that the elimination of all inconsistencies is not entirely possible but that it is under the control of test developers to minimize the effects of inconsistencies. Those potential sources of variance have been described at length by, for example, Thorndike (1951), Lord and Novick (1968), Cronbach et al. (1970), Stanley (1971), and Feldt and Brennan (1989). Brown (1996: 189) summarises their discussions in a list of possible sources of inconsistencies: variance due to environment, due to administration procedures, attributable to examinees, due to scoring procedures and attributable to the test and test items. Anastasi (1997: 85) suggests that the maintenance of uniform testing conditions can reduce error variance and improve test reliability. These uniform testing conditions concern, for instance, the testing environment and time limits. For the evaluation of the Austrian "Gymnasium" two language tests in English and French were developed which were administered on two different days within the same week. Therefore it was particularly important to provide similar conditions in order to reduce the effects of inconsistencies. With regard to Brown's list of inconsistencies above it can be said that all sources of variance were tried to be avoided. Concerning the environment conditions were attempted to be as similar as possible for both administrations. Therefore the tests were administered in the same room of the school. The administration procedures were basically identical: i.e. the order of the test parts and their time limits were the same for both tests. They were also administered at the same time in order to avoid variances of the examinees' disposition due to time of day. Variance due to test

takers was not a problematic issue for the administration of these tests as the examinees were the same for both tests. However, the number of test takers differed unexpectedly as fewer examinees took part in the test of the second day, which was the French test. Therefore the reduced number of test takers may have affected the overall results in the comparison between the English and the French test. The scoring procedures were the same for both language tests. They were agreed upon in advance and defined in the test specifications. For the scoring of the written compositions two judges were used in both languages. The last point in Brown's list above concerns the tests themselves and their tasks. The tests in both languages were designed to be very similar. Therefore the design and layout of the test papers as well as the form of the written instructions were almost the same for both language tests. The length of the tests, the sources of texts and the tasks used were similar. For instance, both tests included multiple-choice items, short-answer questions and true/false/not given items.

While the discussion so far has concerned qualitative aspects of reliability the following section will describe the statistical analysis of reliability.

5.2.1 RELIABILITY COEFFICIENT

The reliability of a test can be expressed in numbers in the form of what is called the *reliability coefficient* (Hughes 2003: 38). "Reliability coefficients, or estimates […] can be interpreted as the percent of systematic, or consistent, or reliable variance in the scores on a test" (Brown 1996: 193). Similar to validity coefficients they allow a comparison of reliabilities of different tests, whereby the perfect reliability coefficient is 1 and the lowest 0. A coefficient of 0 means that the results of a test taker may in no way be of help when predicting the very same test taker's results on the next day. A coefficient of 1, on the other extreme, means that a test taker will have the same scores on the same test, independent of when the test is carried out. Between these two extremes the reliability coefficient of real tests can be found. There have been suggestions of how high the reliability of certain tests should be. Lado (1961 quoted in Hughes 2003: 39), for example, suggested a coefficient between .90 to .99 for good vocabulary, structure and reading tests. Auditory comprehension tests are to be found in the .80 to .89 range. The coefficient of oral productions may be between .70 and .79. McNamara (2000: 62) suggests a reliability of .9 or better on "comprehension tests, or on tests of grammar or vocabulary". This means that a reliability coefficient of .85 may be seen as very high in an oral production test

whereas it may be considered low for a reading test. These considerations reflect the “difficulty in achieving reliability in the testing of the different abilities” (Hughes 2003: 39). Besides, the coefficient will also depend on other considerations. The more important the decisions which are made on basis of the test results, the greater reliability must be demanded (Hughes 2003: 39). Besides, the purpose of the test is decisive for the acceptable degree of reliability. Lienert and Raatz (1998: 14), for instance, consider a reliability coefficient from .5 to .7 high enough for the purpose of comparisons.

There are several possibilities to calculate reliability coefficients. One method is the so called *test-retest* method, in which the same test is administered twice to a group of people (Hughes 2003, Bachman 1990). The drawback is fairly obvious. Leaving aside the assumption that test takers are unlikely to be motivated to take the same test twice, there are two more problems. If the second administration of the test is too soon after the first, test takers may still remember items and therefore a high coefficient will be likely. This source of inconsistency is called *practice effect* by Bachman (1990: 182). If the second administration, however, is too long after the first one, learning or forgetting is very likely to have taken place in the meantime, which will result in a lower reliability coefficient (Hughes 2003: 39). Bachman (1990: 182) calls this second source of inconsistency *learning* (or *unlearning*). This dilemma can be summarised in the following way.

> [P]roviding a relatively *long* time between test administrations will tend to minimize the practice effects, while providing greater opportunity for differences in learning. Giving the test twice with *little* time between, on the other hand, minimizes the effects of learning, but may enhance practice effects (Bachman 1990: 182).

Bachman (1990: 182) therefore concludes that one cannot define a specific amount of time between two test administrations to be appropriate in all testing situations. This problem can be reduced by using two different forms of test items in the two administrations. For this it has to be assumed that the alternate forms are equivalent by being on the same difficulty level (Bachman 1990: 183). This alternate form procedure necessarily involves giving one form first and a second form second. As the order of forms may influence the performance on the two administrations, the effect may be minimized by what is called the ‘counterbalanced’ design, i.e. by dividing the group of test takers into two groups. The first gets form A in the first administration, the second group gets form B in the first administration. The procedure in the second administration will then be vice versa.

Hughes (2003: 40), however, admits that this *alternate forms method* is not always easily done as there are often simply no alternate forms available. He suggests that there is a rather simple solution which gives good estimates of alternate forms coefficients with only one administration of the test, most commonly done in the form of the *split half method*, which provides us with a *coefficient of internal consistency*. Thereby a test is split in two halves with equivalent items. They are scored separately and used to calculate the reliability coefficient as if there had been two test administrations. “This method is rather like the alternate form method, except that the two ‘forms’ are only half the length” (Hughes 2003: 40). An extended discussion of internal consistency and the spilt-half method can be found in Bachman (1990: 172ff.).

In addition to these methods there is a statistical model, called Rasch analysis, with which the reliability value of tests can be estimated. As this model was used for the statistical analysis of the reliability coefficient of the language tests developed for the Austrian “Gymnasium” it will be described subsequently.

Rasch analysis belongs to the item response theory, which has been used in language testing in recent years (Hughes 2003: 228). Rasch analysis is based on the assumptions that firstly, test items can be ranked according to their level of difficulty and that secondly, test takers have a certain level of language ability. The model then uses the test results to predict certain behaviour of test takers and test items. A detailed description of how the model works and how conclusions are made cannot be included here. The most important aspect which is crucial for the analysis in section 12.2.2 is that the model is very sensitive to departures from the predicted model. It “draw[s] attention to test performance that is *significantly different* from what the model would predict. It identifies test takers whose behaviour does not fit the model, and it identifies items that do not fit the model” (Hughes 2003: 229). The analysis in section 12.2.2 will focus exclusively on the items, not on test takers as an analysis of test takers who do not fit the model would only be feasible in collaboration with these test takers, which is not possible at this point of the study. Furthermore, it is rather the characteristics and the reliability of the tasks which are of interest for this analysis as the tasks are those elements which can be controlled and changed by the test developers. The index which is of importance with regard to the indication of items which do not fit the model is the “fit”. In the “fit” column the values indicate how well or how badly the items fit the model (Hughes 2003: 30). “The higher the positive value, the *less well* the item fits” (Hughes 2003: 230). If the fit value of an item is relatively high this means an inconsistency in

answers, for example that the item is answered correctly by weaker test takers and answered wrongly by better test takers. The reason for this misfit may be due to the test takers' behaviour (guessing, lack of concentration, etc.), in which case the item may be perfectly acceptable. If the reason, however, is not based on test takers' behaviour this means that the test item itself is not well designed and has to be revised (Hughes 2003: 231).

5.2.2 THE STANDARD ERROR OF MEASUREMENT AND THE TRUE SCORE

The reliability coefficient allows for a comparison of reliabilities of tests. It does, however, not account for what is called the *true score* of an individual test taker. The true score means the average score which a person gets in repeated administrations of a test (Hughes 2003: 40). For this score, a person would have to take the same test indefinitely often, without any affect on the test score by the fact that s/he has taken the test before, and by changing language ability. The average score from these numerous test administrations would then be assumed to be the best representation of a person's true score and true language abilities. This average score is for obvious reasons not exactly definable. It is, however, possible to define the probability of a candidate's true score to lie within a certain number of points from the actual score obtained on the test (Hughes 2003: 41). The true score is estimated using the actual score from the test and adjusting it within one standard error of measurement (Davies et al. 1999: 214f.). This calculation depends on the *standard error of measurement* of the particular test which "is based on the reliability coefficient and a measure of the spread of all the scores on the test ([…] the greater the reliability coefficient, the smaller will be the standard error of measurement)" (Hughes 2003: 41). The calculation of the standard error of measurement is based on test trials and indicates the probability of the accuracy of a score by defining a probable range around it (Davies et al 1999: 186). A statistical and mathematical guide to calculations on the true score and measurement error, such as Bachman (1990: chapter 6) or Woods, Fletcher and Hughes (1986) provide, cannot be included here. Important for the purpose of this paper are only the conclusions which can be drawn on the basis of the standard error of measurement.

According to Hughes (2003: 41) it is about 68 per cent certain that a candidate's true score lies within one standard error of measurement of the score actually obtained on the test (i.e. if a candidate scores 56 points and the standard error of

measurement of this test is 5, the candidate's true score will – with a certainty of 68 per cent – lie in the range of 51-61).

It is about 95 per cent certain that – under the same conditions as above – the true score will lie within two standard errors of measurement; i.e. between 46 and 66.

With a certainty of 99.7 per cent it can be said that the true score lies within the range of three standard errors of measurement, which in this case would be between 41 and 71.

These statistical statements are based on the fact that scores of tests which are taken indefinitely often would follow the so called normal distribution, from which conclusions on the percentage of scores which would fall within one standard error of measurement etc. can be drawn (Hughes 2003: 52). Leaving aside the statistical rationale, it is important to recognise what the standard error of measurement tells about a test and the scores obtained by an individual test taker and how this informs decisions which are made on the basis of a test (Hughes 2003: 41). If the reliability coefficient of a test is rather low the standard error of measurement is rather high. For all decisions that are made on the basis of test scores the standard error of measurement should therefore be considered. This is particularly the case when the actual score of a candidate would lead to a negative decision about their future, "the standard error of measurement, [however], indicates that their true score is quite likely to be equal to or above the score that would lead to a positive decision" (Hughes 2003: 41).

5.2.3 SCORER RELIABILITY

In the beginning of this chapter, reliability was "defined as consistency of measurement" (Bachman & Palmer 1996: 19), with regard to the scores obtained on different test administrations. The considerations so far have been based on the assumption that scoring itself is always stable and does not vary; neither between different scorers nor between different scoring occasions of the same scorer. However, the assumption of the 'perfect' scoring is, not surprisingly, not realistic. Therefore this section will deal with scorer reliability, a further aspect of the concept of reliability in general.

Scorer reliability comprises two aspects: on the one hand, 'intra-scorer reliability', which refers to the consistency of scores given by one scorer on two occasions

(Hughes 2003: 43), on the other hand, 'inter-scorer reliability', that is, whether different scorers would give the same score on the same test (Hughes 2003: 52). Bachman (1990: 178) used slightly different terms to denote the same concepts, namely 'inter-rater reliability' and 'intra-rater reliability'. These terms are also used by Brown (1996: 206) who defines "intrarater reliability […] [as] an estimate of the consistency of judgements over time".

What can be said for both intra and inter-rater reliability is that as soon as subjective judgement is involved in the scoring of a test, one cannot expect total consistency (Hughes 2003: 43). Purely objective rating can only be assumed when test scoring is not at all subjective and could, for example, also be done by a computer, such as is the case with multiple choice tests. In all other testing situations inconsistencies in scoring are sources of error (Bachman 1990: 178-180). These inconsistencies may either stem from the rating criteria or from the manner of their application. Examples would be the sequence in which tests are scored, and varying focus on different aspects of the rating criteria between scorers. Both Brown (1996) and Bachman (1990) discuss different ways in which inter- and intra rater reliability can be examined. Bachman (1990: 180f.) also provides mathematical formulae to estimate the inter- and intra-scorer reliability. It is important to know that, in analogy to the test reliability coefficient, there is a scorer reliability coefficient, which can be interpreted similarly (Hughes 2003: 43). It can vary from 0 to 1, whereby 1 means the totally objective scoring of, for example, multiple choice tests. "Benchmarks for minimum acceptable inter-rater agreement range from 0.7 to 0.9 on this scale, depending on what is at stake" (McNamara 2000: 58). Subjective test scoring can never be expected to have a coefficient of 1. There are, however, ways with which scorer reliability can be improved so that a coefficient of 0.9 may be obtained (Hughes 2003: 43).

Scorer reliability has also been of importance for the development of the tests presented in this paper. The written compositions were scored independently by two different judges. The scores were then compared to agree on a final score. However, a mathematical estimation of inter-rater reliability was not carried out due to practical reasons. Reliability of scoring affects the overall reliability of a test and will be discussed in more detail in the following chapter and, with regard to the language tests developed for the Austrian "Gymnasium", it will be discussed in section 12.2.1.

5.2.4 HOW TO IMPROVE RELIABILITY

Reliability is a crucial language test quality and is relevant for overall test usefulness: the more reliable the more useful is a particular test. Therefore it is important to know how the reliability value of a language test can be improved. Hughes (2003: 44-50) suggests 15 guidelines with which the reliability of tests and scoring can be increased: the first eight relate to the improvement of consistent performances, the last seven suggestions refer to scoring reliability. In the present section the suggestions will only be outlined theoretically. In section 12.2.1 they will be discussed with regard to the language test developed for the Austrian "Gymnasium".

- Generally speaking it can be said that reliability improves with an increasing number of test items (Hughes 2003: 44). This presupposes that items are independent from others and therefore allow for additional information to be gained from them. Hughes emphasises the relativity of this matter by saying that *enough* samples should be taken: this means neither too many tasks, as this would bore test takers and falsify their performance, nor too few tasks, as this would decrease reliability.

- Items which can be answered by weaker and stronger students or items which cannot be answered by any test takers do not contribute to the reliability of the test and can therefore be excluded (Hughes 2003: 45). These items either have a very low or a very high item facility value and their results are not informative with regard to language ability[7].

- Allowing test takers too much freedom by giving choices and a range of possible answers decreases reliability and should therefore be avoided. It has to be guaranteed, however, that restrictions do not "distort too much the task that we really want to see them perform" (Hughes 2003: 46).

- Tasks are most reliable when there is only one correct answer to them. Therefore items should be worded unambiguously (Hughes 2003: 46). In order to avoid ambiguous items and to avoid acceptable answers which were not anticipated by the item writers all tasks should be pre-tested before the actual administration.

[7] The concept of item facility value will be discussed in chapter 11.

- The importance of providing explicit instructions is closely linked to the preceding suggestion. Both written and spoken instructions must be as clear as possible in order to avoid misinterpretations on the side of the test takers (Hughes 2003: 47).

- Tests should be well laid out and easily legible as test takers should only be tested on their language ability and not on additional tasks which arise through badly typed, small and poorly reproduced texts (Hughes 2003: 47).

- Test takers should get the opportunity to familiarise themselves with testing techniques as unfamiliarity with test formats may result in worse test performance (Hughes 2003: 47).

- Uniformity of conditions, such as timing or acoustic conditions should be aimed at in order to avoid differences in test performances between two administrations (Hughes 2003: 48).

The following pieces of advice relate to scorer reliability.

- Scoring should be as objective as possible (Hughes 2003: 48). Therefore items which allow relatively objective scoring should be used. This does not mean, however, that multiple choice items should be used in all tests. There are also other appropriate test methods such as open-ended questions with a unique correct answer.

- Test takers should be compared to each other as directly as possible. This refers back to the section on freedom in choices. Scoring compositions on one topic will be more reliable than scoring compositions where test takers are allowed to choose from more topics. Therefore the topics should be restricted and guided (Hughes 2003: 49).

- A detailed scoring key should be provided in advance (Hughes 2003: 49). The key should be as detailed as possible including specifications of correct answers and answers which are partially correct. Also the assignment of points should be agreed upon and defined.

- Judges should be trained before the scoring of a test (Hughes 2003: 49). This is of particular importance when scoring is subjective, such as is the case with compositions. Raters should be trained and patterns of scoring should be analysed and compared after each administration.

- Acceptable answers to test items should be defined before the tests are scored (Hughes 2003: 49). In the case of compositions, typical examples representing specific ability levels should be chosen directly after the first test administration. All scorers should then agree upon the scoring before the actual scoring of the test compositions starts. In the case of short answers scorers should discuss any difficulties and unaccepted answers.

- Test takers should be judged anonymously and should not be identified by name (Hughes 2003: 50). Identification of candidates by numbers may reduce expectations of scorers. Also in situations in which candidates are unknown to scorers, inferences from test takers' names may affect the scoring.

- Tests should be scored independently by more than one judge (Hughes 2003: 50). Particularly in cases of subjective scoring Hughes suggests two independent raters to judge the compositions. Ideally, a third person finally compares the scores given to analyse discrepancies.

This summary has presented important aspects which can improve the reliability of a test. In section 12.2.1 it will be analysed which and how many of these guidelines were taken into consideration during the development, the administration and the scoring of the language tests designed for the Austrian "Gymnasium".

5.2.5 RELIABILITY UND VALIDITY

The two preceding sections dealt with validity and reliability, the two main qualities of language tests. After their separate description, their importance and interrelationship will now be summarised before the other four test qualities will be discussed. Reliability and validity are both measurement qualities which are essential for overall test usefulness (Bachman & Palmer 1996: 23). The relationship of reliability and validity can be described as follows: "[t]o be valid a test must provide consistently accurate measurements. It must therefore be reliable" (Hughes 2003: 50). Reliability is therefore a necessary aspect of overall test usefulness (Bachman & Palmer 1996: 23). This relation does, however, not presuppose that it also works vice versa. A reliable test does not automatically have to be valid. On the contrary, a test which is perfectly reliable can be thought of to be not valid as it does not test what it should (Hughes 2003: 50). Hughes (2003: 50) gives an example of this possibility: in order to test writing skills one could ask test takers to write down

translations of 500 words in their own language. This might be a very reliable test, but it might very likely be regarded as not valid. Following these considerations it can be concluded that although reliability is a necessary condition for construct validity it is not sufficient to guarantee test usefulness (Bachman & Palmer 1996: 23). What these considerations try to point out is the interrelated importance of both qualities for test usefulness. When test developers try to design tests which are as reliable as possible, validity should not be regarded as less important (Hughes 2003: 50). If, for example, the restriction of test takers' compositions to a certain word limit decreases validity for the sake of an increasing reliability, it must be considered if validity of the task is still high enough to be able to draw inferences from it. Hughes (2003: 50) summarises the relationship between validity and reliability as follows: "There will always be some tension between reliability and validity. The tester has to balance gains in one against losses in the other". This statement holds also true for the other test qualities. One single test may never perfectly possess all six qualities. Test developers thus always have to find a compromise and decide which qualities seem to be more important than others for a certain purpose.

5.3 AUTHENTICITY

During the last decades there have been numerous attempts to define authenticity with regard to language tests but the definitions have not always used the same terminology, as will be seen in the following brief summary. In 1961, Carroll was one of the first to describe two different approaches to language testing. He distinguished between 'integrative' and 'discrete-point' approaches (Carroll 1961: 37 quoted in Bachman 1990: 300). Although Carroll did not use the term authenticity, the former approach is strongly linked to it. 'Integrative' testing is concerned with the communicative aspect of an utterance rather than with testing certain structures in the language. This idea is still evident in the belief of many language testing researchers who take the view that a thorough measure of a test taker's language ability can only be obtained by creating testing situations which demand similar requirements from the examinees to those demanded in non-test situations (Bachman 1990: 300f.). A similar definition of authenticity is given by Davies et al (1999: 13) where "[a] language test is said to be authentic when it mirrors as exactly as possible the content and skills under test". In this definition it is important to emphasise the restriction 'as exactly as possible' as it is impossible to achieve

complete authenticity in a testing situation which is never a natural occasion. Since test performance is always a specific situation it is crucial to accept that however realistic test tasks seem to be they are never – what McNamara (2000: 9) called – *real*.

Over the last decades the matter of authenticity has been of great importance for test development, which can be seen in the discussion of tests which have been "described variously as 'direct', 'performance', 'functional', 'communicative', and 'authentic'" (Bachman 1990: 301). All these attributes reflect the consideration of the concept of authenticity, which in the last decades has resulted in two main approaches. The first is called the 'real-life' (RL) approach. The second is the 'interactional/ability' (IA) approach (Bachman 1990: 301f.).

In Bachman's view, the RL approach

> considers the extent to which test performance replicates some specified non-test language proficiency […] [and is concerned with] (1) the appearance or perception of the test and how this may affect test performance and test use (so-called 'face-validity'), and (2) the accuracy with which test performance predicts future non-test performance (predictive utility) (Bachman 1990: 301f.).

This RL approach appears to be strongly linked to the concept of validity. The two terms face-validity and predictive utility have been discussed in this paper in the chapter on validity. It is essential for this approach that it does not make a distinction between the language ability which is to be measured and the context of performance (Bachman 1990: 302).

In contrast to this, the IA approach does not look "at non-test language performance *per se* as criterion, […] [but] focuses on what it sees as the *distinguishing characteristic* of communicative language use – the interaction between the language user, the context, and the discourse" (Bachman 1990: 302). Test tasks are therefore designed so that test takers are involved in communicative language abilities. The main difference is that this approach does differentiate between the ability which is intended to be measured, on the one hand, and the performance and the context in which it is observed, on the other hand (Bachman 1990: 303). Its main concern "is with demonstrating the extent to which test performance reflects language abilities, or with construct validity" (Bachman 1990: 303). Despite the differences between the two approaches, in particular with respect to context of testing, Bachman concludes that "[b]oth approaches to authenticity are concerned

with the context of manner in which we elicit a sample of performance – with the characteristics of the testing methods we use" (Bachman 1990: 303). This means that in order to understand and investigate authenticity two test components have to be defined: the non-test communicative language use (or TLU domain) on the one hand, and characteristics of test tasks or test method facets, on the other hand. The TLU domain and test tasks' characteristics and their influence on test performance have been discussed in previous sections of chapter 4.

Bachman and Palmer (1996) do not explicitly distinguish between the two approaches presented above but suggest one definition of authenticity, i.e. "the degree of correspondence of the characteristics of a given language test task to the features of a TLU task" (Bachman & Palmer 1996: 23). The correspondence allows for interpretations of the test scores which can then be generalised beyond test performance to language use in the TLU domains. This generalisability of test scores closely links authenticity to construct validity as the interpretation of scores is a crucial part of construct validation. One further important point why authenticity should be considered in the test development is "its potential effect on test takers' perceptions of the test and, hence, on their performance" (Bachman & Palmer 1996: 24). If students consider a test relevant in terms of content, tasks and correspondence to their TLU domain, they are likely to react positively to it. This idea relates to the concept of face validity which was discussed in section 5.1.3. Although test takers have, of course, no possibility to analyse a language test empirically they can judge it by its content and its appearance. If they consider both content and layout appealing and relevant this may very likely encourage their motivation to take part in it. Summarising the considerations on authenticity, the steps in order to develop authentic tasks are as follows: Firstly, distinguishing features of tasks in the TLU domain have to be identified (Bachman & Palmer 1996: 24). In the second step tasks with these critical features are designed. In chapter 9 on stages of test development this idea will be considered further.

5.4 INTERACTIVENESS

The fourth language test quality is interactiveness. Bachman and Palmer (1996: 25) "define interactiveness as the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task". In other words, interactiveness is the engagement of "language knowledge, metacognitive

strategies, topical knowledge, and affective schemata" (Bachman & Palmer 1996: 25). It is significant, however, that interactiveness does not appear consistently in the literature (neither the *Encyclopedia of Language and Education,* nor the *Dictionary of language testing* define interactiveness; for example McNamara 2000, Brown1996 or Bachman 1990 do not list interactiveness in their indices). However, it seems to be common practice among other linguists not to refer to this concept as a quality on its own, but to consider interactiveness as 'interactional authenticity' in combination with the quality authenticity (Douglas 1997: 116). Bachman and Palmer are among the researchers who explicitly discuss interactiveness. Their description from 1996 will be summarised briefly subsequently. Interactiveness as the test takers' involvement in a test task is not sufficient for tasks to be considered as measures of language ability. Even if test takers were very engaged in mathematical calculation tasks, the performance on these tasks would not be used to draw any inferences on test takers' language ability. This seems to be very obvious, but Bachman and Palmer (1996: 26) use this argument to conclude that interactiveness "provides the vital link with construct validity". Emphasising the relationship between construct validity and interactiveness on the one hand, Bachman and Palmer stress the difference to authenticity on the other hand. While authenticity "pertains to the correspondence between test tasks and TLU tasks, and must consider the characteristics of both kinds of tasks, interactiveness resides in the interaction between the individual […] and the task" (Bachman & Palmer 1996: 26). Despite the minor role interactiveness plays in a majority of the literature, the point that Bachman and Palmer make as they distinguish between authenticity and interactiveness appears to be very important and should therefore be considered in test development. For the development of the language tests for the Austrian "Gymnasium" this distinction made by Bachman and Palmer (1996) was taken into consideration: Test tasks and contents were chosen according to what was assumed to be of interest to the test takers. This was particularly the case with the texts which were used for the writing tests and the topic which was used for the speaking test[8]. The topic of one writing task was, for instance, the introduction of university tuition fees. As most of the test takers were assumed to continue their education at university this topic was considered to be an interesting issue. The topic of the speaking test was the smoking ban in Austria. As this is a current issue in Austria the topic was chosen as it was assumed to evoke interest and

[8] The structure and contents of the language test can be seen in the appendix (cf. section 15.1) and will be described in section 9.5.

commitment on part of the test takers. The choice of these topics was made due to considerations of interactiveness. With these subject matters the distinction between authenticity and interactiveness in the sense of Bachman and Palmer becomes obvious. A discussion of the smoking ban or the introduction of tuition fees in Austria does not necessarily relate to the target language use domain of the test takers, which would represent the language test quality of authenticity, but rather to their personal interests. Therefore the choice of these topics supports the idea of Bachman and Palmer to consider interactiveness as a separate language test quality and not as part of authenticity.

5.5 IMPACT

The next quality of language tests which has to be considered is the effect they may have on educational and social systems and the individuals within these systems (Bachman & Palmer 1996: 29). Wall (1997: 291) defines impact as "the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole". All tests can be said to have effects on people and educational systems as they "are not developed and used in a value-free psychometric test-tube; [but] they are virtually always intended to serve the needs of an educational system or of society at large" (Bachman 1990: 279). This means that as soon as a test is administered and scores are interpreted certain values and goals are implied, which will have consequences (Bachman & Palmer 1996: 30). These consequences of test use can be categorised according to two levels of impact: a micro level, which refers to the effects of test use on individuals and a macro level, which relates to the effects on the educational system or society.

5.5.1 MACRO LEVEL

To begin with the macro level, the influence of tests on society and educational systems has to be considered in the development, the scoring and the drawing of inferences from tests (Bachman & Palmer 1996: 34). These considerations may become particularly important and complex in the field of second or foreign language testing as this most certainly involves more than one culture between which values and goals may vary. Culture specific values are, however, not the only source of variation. Values do also change over time "so that issues such as secrecy and access to information, privacy and confidentiality, which are now

considered by many to be basic rights of test takers, were at one time not even a matter of consideration" (Bachman & Palmer 1996: 34). Impact on the macro level means that language tests may influence educational systems or programs, particularly in the case of high stake tests, on the basis of which important decisions about many test takers are made. High stake tests are tests which have highly influential effects on examinees' future professional or educational careers (Davies et al. 1999: 185). By contrast, one speaks of low stake tests where the test results are not very likely to have substantial effects on the test takers' future lives. In the largest possible sense, tests may also have impact on a society as a whole. This may happen when test results are used to make decisions about "immigration […] [and] certification for professional practice" (Davies et al 1999: 79). Societal impact can also be seen when language tests are used to group pupils into instructional courses (Bachman & Palmer 1996: 34). The use of tests in situations like those may influence not only the people directly involved but also society as a whole. The language test developed for the Austrian "Gymnasium" can be considered a low stake test as the test results will not affect the test takers' future lives substantially. The purpose of the test was not to make decisions about individuals but to enable the people who commissioned the test to base decisions concerning the language tracks of the school on empirical data. Therefore the effects of the language test on the macro level were restricted to the educational system of this particular school. The test results will be used in order to justify maintenance or change of the current language track system of the school[9]. As the test takers were pupils in their final year who already graduated they are not directly affected by possible future changes of the curriculum. Only pupils who will attend this school in the future may be affected by consequences based on the test results.

5.5.2 MICRO LEVEL

In contrast to the macro level, the micro level refers to a test's impact on individuals. This form of impact has often been called washback or backwash[10]. Wall (1997: 291) takes a slightly different view when she states that "'[w]ashback' (also known as 'backwash') is sometimes used as a synonym of 'impact', but it is more frequently used to refer to the effects of tests on teaching and learning". She thus

[9] The language test was repeated in the following year and may have a third administration in the year 2010. Decisions will be based on the results of all three test administrations.
[10] cf. Bachman & Palmer 1996: 30. Please note that the two terms are used interchangeably in the current literature and will also be used synonymously in this paper.

does not restrict washback to effects on individuals but on teaching and learning in general, albeit this definition implies the individuals who are involved in teaching and learning. In the same way McNamara (2000: 73f.) distinguishes between washback and impact. The former affects the instructional setting of learning and teaching, whereas the latter refers to the influence which reaches beyond the scope of the classroom and affect society. From the definitions it can be seen that one aspect in which Wall (1997), McNamara (2000) and also Bachman & Palmer (1996) agree is the strong link between impact and washback. Bachman and Palmer's (1996: 30) argument for this is that washback can be observed in "processes (learning and instruction) […] [which] take place in and are implemented by individuals, as well as educational and societal systems, and they have effects on individuals, educational systems, and society at large". Hughes (2003: 53) defines backwash as "the effect that tests have on learning and teaching" and asserts that backwash can either be positive or negative (Hughes 2003: 1). Negative or harmful backwash occurs when language test items do not reflect the teaching curriculum but are developed on the basis of an outdated approach to language testing. This lack of correspondence between the items and the curriculum results in the need of practising the particular items rather than the skill itself (Davies et al. 1999: 225). One speaks of positive or beneficial backwash, to the contrary, when language testing affects teaching positively. This is, for instance, the case when the implementation of a test which involves measuring test takers' listening skills results in intensive practising of this crucial – but sometimes disregarded – language ability.

As defined above, backwash – either positive or negative – has an influence on individuals. The people who are affected will have an interest in the use of a particular test (Bachman & Palmer 1996: 31). These so-called *stake holders* are all people who are directly or indirectly involved in the test use, such as the test takers, their families and their teachers as well as the test developers and the people who commission the language test (Davies et al. 1999: 184). Among these stake holders some can be said to be more directly affected than others. People who are indirectly affected by tests and the decisions which are made on their basis are, for instance, "test takers' future classmates or co-workers, future employers" (Bachman & Palmer 1996: 31). When we take into account the impact of tests on educational systems and society as a whole as it has been described in the previous section, it may even be argued that all members of a society are indirectly affected by tests. Following the discussion of washback in Bachman and Palmer (1996), only two groups of people who are probably most directly affected by tests will be considered in this

paper: test takers on the one hand, and test users in the form of teachers on the other hand.

Impact on test takers

Bachman and Palmer (1996: 31) state that "[t]est takers can be affected by three aspects of the testing procedure:

1 the experience of taking and, in some cases, of preparing for the test,

2 the feedback they receive about their performance on the test, and

3 the decisions that may be made about them on the basis of their test scores".

The first aspect may influence test takers' characteristics such as topical knowledge, their perception of the TLU domain, their areas of language knowledge and their use of strategies (Bachman & Palmer 1996: 31f.). Preparation for certain tests, particularly high stake tests, may not only take place individually but often also influences teaching by practising test tasks and techniques in school. Taking the test itself may influence test takers' topical knowledge, when, for instance, test tasks offer new information. Test takers' perception of the TLU domain may be particularly influenced when the TLU domain is unfamiliar to the test takers. This is the case when students enrol for programs abroad. Bearing in mind these potential influences test developers have to be critical about the extent to which the test is informative or misleading with regard to test takers' topical knowledge or the TLU domain. Language knowledge may be influenced, either positively, when test takers improve it while taking the test, or negatively, when presented features in the test are misleading as they do not correspond to what test takers learnt before taking the test.

The second aspect, getting feedback, will very likely directly influence test takers. Thus feedback must be presented in a most "relevant, complete, and meaningful" (Bachman & Palmer 1996: 32) way in order to evoke "a positive affective response toward the test on the part of the test takers" (Bachman & Palmer 1996: 32).

The third aspect, the decisions which are made on the basis of the test, may have a range of effects on test takers (Bachman & Palmer 1996: 32). Decisions about passing or failing, employment or non-employment, advancement or non-advancement are most certainly always of great influence on individuals. Therefore it is crucial to consider the fairness of the decisions. Decisions are considered to be

fair when they disregard examinees' group membership. Fairness is also involved in the extent to which test scores are considered to be relevant and appropriate as these scores are then the basis of decision making processes (Bachman & Palmer 1996: 33). Fairness finally also relates to the extent to which test takers are informed about the decisions and the way they are arrived at.

Impact on teachers

Test takers are normally not the only individuals who are influenced by a test. The second main group of people who are affected has earlier in this chapter been referred to as test users. These are individuals who make decisions on the basis of tests. Bachman and Palmer (1996: 33) focus on the background of instructional programs where teachers are the test users that can be said to be most directly involved in the test use. Many teachers probably know the possible influence of tests on their teaching. The term 'teaching to the test' describes what teachers are often confronted with when they have to assess their students with a specific test. Even teachers who want to teach certain other materials or methods may find it hard to avoid practising the methods used in the test. Thus their teaching is influenced by the test students take. Bachman and Palmer (1996: 33) note the important point that one can relate this consideration to the concept of authenticity as described earlier in this paper. If the instructional program is considered the TLU domain and the test tasks do not correspond to the tasks of teaching this may be regarded as a case of low authenticity. What has been said about teaching to the test may also be used for the opposite purpose, namely to achieve positive backwash. When an instructional program and its results are considered dissatisfying by the people who are responsible for it they may want to use a test to positively affect the teaching and thus the instructional program (Bachman & Palmer 1996: 33f.). However, studies (for example by Wall & Alderson 1993) have shown that this assumption cannot be taken for granted.

The two preceding sections were concerned with test impact on test takers and teachers. In the case of the study presented in this paper the backwash effect of the language tests may be considered very small. It can be argued that the test takers were not at all affected by the test; neither by preparing nor by taking it. The reason for this is that they did not know in detail which test formats to expect and could therefore not practise them in advance. With regard to getting feedback as described above it can be said that the test takers did only get feedback when they

asked for it. Examinees who did not want to get any information about their test performance were not confronted with their test results. Besides, the knowledge that the results would not influence their grades or any decisions about their future educational careers may have prevented any negative backwash. The impact on the teachers can be considered to be similar. Although test results may be traced back to their teaching it cannot be said if and in how far their teaching will be influenced by test results.

5.5.3 HOW TO ACHIEVE POSITIVE WASHBACK

Summarising the two previous sections, the impact of tests is or can potentially be highly influential, not only on individuals but also on society and educational systems and therefore have to be considered in the development, scoring and interpretation of test performances. These influences may either be positive or negative but they certainly cannot be avoided. Davies (1990: 1), for instance, claims that "'washback' is so widely prevalent that it makes sense to accept it, to stop regarding is[sic] as negative, and then make it as good as it can be in order to improve its influence to the maximum". There have been several attempts to offer guidelines on how to achieve positive backwash (for example Shohamy 1992, Bailey 1996, Hughes 2003). However, the problem with these guidelines is the fact that there is little evidence of how tests influence teaching and which components lead to positive washback (Wall 1997: 298). Therefore "none of this advice incorporates all we need to know about tests, the uses to which they are put, the importance of the context, the test users, and other factors that are important" (Wall 1997: 298) for positive backwash. Wall concludes that there is a clear need for further research in order to be able to provide guidelines for beneficial washback. Even then, however, there are still practical issues which may restrict the achievement of positive backwash, such as financing specific equipment and qualified assessors. As this is not always possible, beneficial impact is very often limited by considerations of practicality. The subject of practicality as the last of the six qualities of language tests will be discussed in the subsequent section.

5.6 PRACTICALITY

The fact that practicality is discussed last in this chapter on language test qualities does not imply that it is of little importance. Quite to the contrary, practicality is a crucial aspect of language testing as it influences all other test qualities. To a large degree, practicality is the decisive factor as to whether a test is developed and implemented at all (Bachman & Palmer 1996: 35). Components which are included under the term practicality are for instance, "cost of development and maintenance, test length, ease of marking, time required to administer the test […], ease of administration (including availability of suitable interviewers and raters, availability of appropriate room or rooms) and equipment required (computers, language laboratory, etc.)" (Davies et al 1999: 148). There are various different ways of categorising these components which are included under the term practicality. The distinction which seems to be most useful and clear was suggested by Brown (1996) who lists three issues which have to be taken into account when tests are developed.

The aspect of practicality which is probably most important is what Brown (1996: 32) calls the *Cost Issue*. The *Cost Issue* is decisive for the development, design and administration of a test (Bachman & Palmer 1996: 35) and is of particular importance as it affects all other issues of practicality. "Lack of funds can cause the abandonment of otherwise well thought out theoretical and practical positions that teachers have taken" (Brown 1996: 33). In addition to the Cost Issue the second issue of importance is the *Fairness Issue* (Brown 1996: 31). It is concerned with objectivity of assessment, which has already been discussed in the section on reliability. It is interesting that Brown includes fairness within the scope of practicality as most other researchers treat fairness in the context of reliability. The third crucial aspect of the concept of practicality is called *Logistical Issue* and comprises "[e]ase of test construction" (Brown 1999: 33), administration and scoring. Test construction involves the decision of test length and types of test tasks (Brown 1999: 34). Ease of test administration depends on factors such as the time of administration, the number of tests and the kind of equipment and materials which are necessary. The last issue of test scoring refers to how easy and thus cheap scoring is. It is important to note at this point that this issue seems to be in an opposing relationship with the ease of construction (Brown 1999: 35). The easier the construction of a test type, the more difficult is the scoring.

Generally speaking, the reason why practicality is crucial in test development is that, regardless of perfect reliability and validity of a test, it will not be administered if it is not practical (Davies et al. 1999: 148). Therefore practicality is – in addition to reliability and validity – often regarded as the third aspect which is considered in test development (Davies et al. 1999: 148). In the same way, Bachman and Palmer (1996: 36) state "that considerations of practicality are likely to affect our decisions at every stage along the way, and may lead us to reconsider and perhaps revise some of our earlier specifications". Practicality was, of course, also a crucial aspect responsible for many decisions and compromises which were made during the development of the language test described in this paper. Financial considerations as well as issues of personnel and time were to a large degree decisive for decisions on test design and administration. These restrictions to the test development will be described in section 12.1 which discusses the possible effects of practical limitations on the overall usefulness of the test.

The separate discussion of the six qualities of language tests may give the impression that the six qualities function independently from each other. However, this is not the case. The figure in chapter 4 already indicated that the qualities interact, influence and complement each other in terms of the overall test usefulness. With regard to test development it has to be said that "test design involves a sort of principled compromise" (McNamara 2000: 27). "All decisions about test method […] involve a compromise between the desirability of an appearance of authenticity on the one hand and the practicalities imposed by the test situation on the other" (McNamara 2000: 28). This compromise will be considered further in the description of the language test compiled for the Austrian "Gymnasium" in section 12.1.

6 COMMON EUROPEAN FRAMEWORK OF REFERENCE (CEFR)

For tests to be comparable within one country from one school or institution to another or also on an international level between countries there is a need for a framework which allows for comparisons between test results by being based on the same levels of proficiency. One of these frameworks in the European context is the Common European Framework of Reference (CEF/CEFR), which will be described in this chapter. It has to be made clear, however, that a detailed description of the CEFR cannot be given as this would go beyond the scope of this paper. For now it

seems to be useful to give a short introduction of what the CEFR is and then to refer to those aspects of the CEFR which were taken into consideration for the test development for the Austrian “Gymnasium”.

> The Common European Framework provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe. It describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively. […] The Framework also defines levels of proficiency which allow learners’ progress to be measured at each stage of learning and on a life-long basis. (CEFR 2001: 1)

As the name indicates this framework tries to overcome national barriers and therefore “facilitate the mutual recognition of qualifications gained in different learning contexts, and accordingly […] aid European mobility” (CEFR 2001:1). The mutual recognition of qualifications is important for international comparisons and thus promotes “co-operation in the field of modern languages” (CEFR 2001: 1). Another aim of the CEFR is to “assist learners, teachers, course designers, examining bodies and educational administrators to situate and co-ordinate their efforts” (CEFR 2001: 6). The framework can be used variedly in different areas such as planning of language learning programmes or language certification and planning of self-directed learning. For each use the CEFR offers guidelines including, for instance, objectives, content, assessment criteria, selection of materials and assessment criteria, etc. In order for these functions to be fulfilled the CEFR has to be “comprehensive, transparent and coherent” (CEFR 2001: 7). ‘Comprehensive’ means a description of “as full a range of language knowledge, skills and use as possible” (CEFR 2001: 7) for every user to be able to describe their objectives etc. referring to the CEFR. For all skills different stages or levels of proficiency are described in order to be able to calibrate progress. ‘Transparent’ means the clear and explicit formulation of information which should be “readily comprehensible to users” (CEFR 2001: 7). The third characteristic of being ‘coherent’ refers to the description which should be “free from internal contradictions” (CEFR 2001: 7). With regard to educational systems, coherence requires that there is a harmonious relation among their components:

- ✓ the identification of needs;
- ✓ the determination of objectives;
- ✓ the definition of content;

- ✓ the selection or creation of material;
- ✓ the establishment of teaching/learning programmes;
- ✓ the teaching and learning methods employed;
- ✓ evaluation, testing and assessment (CEFR 2001: 7).

In addition to the three characteristics of comprehensiveness, transparency and coherence, the CEFR "should be open and flexible, so that it can be applied, with such adaptations as prove necessary, to particular situations" (CEFR 2001: 7). The CEFR is in itself a very complex framework including descriptions for many purposes, different contents and contexts with regard to various materials and situations. This chapter will only present those aspects which were taken into consideration for the development of the language test for the Austrian "Gymnasium", namely the common reference levels and one table which describes these levels for the skills which were tested.

To begin with the levels of proficiency, there seems to be an agreement on the number and contents of levels which can be applied to a categorisation of language learning (CEFR 2001: 22). There are six levels which provide coverage of language learning proficiency. Different terms are used to refer to these six levels, which will be illustrated in the following figure (cf. Figure 4).

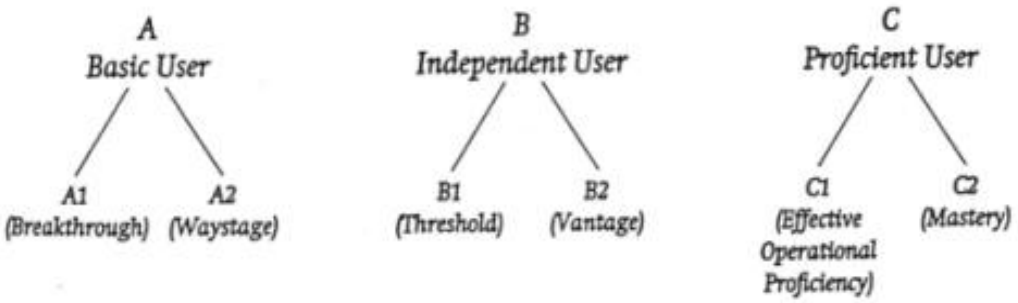

Figure 4: CEFR Levels of Proficiency (CEFR 2001: 23)

The authors are aware of the potential as well as of the restrictions of the descriptions of these levels and state that "[i]t is […] to be expected that these precise formulation of the set of common reference points, the wording of the descriptors, will develop over time as the experience of member states and of institutions with related expertise is incorporated into the description" (CEFR 2001: 23f.). It is also acknowledged that different presentations of the six levels will be

appropriate in different settings. Whilst a simple holistic description may be considered useful to inform non-specialist users and to provide guidelines for teachers and curriculum planners a more complex and detailed description may be requested in other situations (CEFR 2001: 24). The holistic description can be seen in the table below (cf. Table 2).

Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Table 2: CEFR Global Scale (CEFR 2001: 24)

For the purpose of the test development which will be discussed in chapter 9 there was a need for a more detailed description of the common reference levels, going beyond the descriptions in the holistic table above. As the test aimed at the B2 level and intended to test the four skills of listening, reading, writing and speaking, the CEFR was taken into consideration with respect to these four skills at the appropriate levels. The description of the levels and skills which was used for the

test development can be seen in the test specifications in the appendix (cf. section 15.2) and in the table (cf. Table 3) below.

		B2
UNDERSTANDING	Listening	I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programmes. I can understand the majority of films in standard dialect.
	Reading	I can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. I can understand contemporary literary prose.
SPEAKING	Spoken Interaction	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my views.
	Spoken Production	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
WRITING	Writing	I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences.

Table 3: CEFR Level B2 (CEFR 2001: 27)

After this brief description of the framework which was used for the test development for the Austrian "Gymnasium" the subsequent chapter will be concerned with the setting of the school in which the test was carried out and the research assignment.

7 SETTING

As indicated in the introduction the language test presented in this paper was commissioned by the parents' association of an Austrian "Gymnasium". In the beginning of 2008 the parents' association asked the University of Vienna to carry out an objective empirical study in order to compare the two language tracks which are offered at the school. Before a detailed description of the research assignment can be given it is necessary to describe the school and its current situation with regard to the two language tracks.

The school is the oldest "Gymnasium" of Vienna and is situated in the first district. Former graduates include, for instance, Schubert, Schnitzler, Hofmannsthal and Schröding. The school shows pride of its traditional and cultural values and tries to complement them by innovative education. Many pupils' parents have an academic background and decide to send their children to this "Gymnasium" as it has a very good reputation due to various factors. One attractive aspect of this "Gymnasium" is certainly the offer of different language tracks.

The two language tracks offered at the school differ in time and extent of English and French lessons offered. One track starts with English as the first foreign language in the first year. For the sake of practicality this track will be called the English-track in this paper. The pupils in the English-track begin to learn French in the second year where it is only a "Wahlpflichtfach", i.e. two hours a week without exams. From the fourth year onwards French is then taught as a compulsory subject. In addition to these two languages pupils have Latin classes beginning in the third year. As an additional option pupils may choose Greek or French in the fifth year. The second language track works similar to the English-track with the only difference being that pupils start with French as their first foreign language in the first year. This track will be referred to the French-track in this paper. In addition to French classes a variety of extra curricular activities which relate to the French language are offered: a visit to a partner school in France in the first year, 2 weeks in Nantes in the fourth year, an intensive language course for one week in the second or third year and a geography project in which French is the language of instruction. Pupils from the English-track go to Ireland for one week in the seventh year. At the time of test development it was not known if any projects in which English is the language of instruction are offered.

8 STARTING POINT AND RESEARCH ASSIGNMENT

The assignment of the language test in 2008 was not the first attempt to evaluate the two language tracks offered at the school. Some concern for an evaluation of the two language tracks has been existent at the school for some time. Before this research assignment evaluations in the form of questionnaires and interviews had been designed by a group of teachers. However, a full evaluation was never carried out. For this reason the parents' association decided to assign an external team of assessors for an independent evaluation which was to be as objective as possible. The specific interests of the parents' association were the levels of language abilities in both English and French of all pupils of the two language tracks at the school. They were interested to find out if there were any significant differences in the language ability between the two tracks in the final year, i.e. after eight years of instruction.

For this comparison a group of students of the University of Vienna were assigned to develop two language tests at the same level: one test in English and one to test pupils' French language abilities. The two tests were designed to be of the same level of proficiency (B2 according to the CEFR). In order to gain significant information about the language abilities of the pupils a comprehensive test which evaluated all four skills (listening, reading, writing and speaking) was developed. However, it has to be said at this point that the study presented in this paper was not a comprehensive language program evaluation as the language test was the only source of information about the learning outcomes of the pupils of the two language tracks. For the purpose of this paper the main focus will be on a description and analysis of the English test. The French language test will, however, be used to illustrate significant differences in the results. As the test was developed by five project members the pronoun *we* will be used subsequently to refer to the group of test developers. In the next chapter the stages of test development and the test administration will be outlined.

9 TEST DEVELOPMENT AND ADMINISTRATION

The stages of test development have been variously described by numerous scholars in the field of language testing. This chapter will briefly summarise different categorisations of test development and will then refer to the development of the language test for the Austrian "Gymnasium". Although this summary of stages in test development only draws on a limited range of researchers it already shows that there is a variety of suggested stages and activities which do not necessarily completely coincide, depending on the focus and elaborateness of the single researchers. The procedure which seems to be most clear and useful for the development of the test for the Austrian "Gymnasium" is McNamara's (2000).

McNamara (2000: 25ff.) categorises test development into four stages: establishing test content, establishing test method, writing test specifications and finally trialling or trying out. The stages he suggests comprise all important steps in test development and allow for adaptations according to the particular testing situation. For instance, the number of triallings can be varied depending on the test purpose and whether it is a high stake or a low stake test. Other scholars, for instance Anstey (1966), have also proposed four stages of test development which, however, differ from the ones proposed by McNamara as Anstey focuses on trialling. According to Anstey the first stage is the *planning* stage, in which the content, layout, length and scoring are decided on (Davies 1990: 12). This stage thus comprises what McNamara described in stages 1-3. Anstey's stages 2-4 – *prepilot* stage, *pilot* stage and *final validation* stage – are concerned with trialling: first on a "small group of interested people", [...] [then] on a large sample of the same kind of people on whom the test is to be used" (Davies 1990: 12) and finally a trialling of the test's final form. Due to the focus on trialling Anstey's model seems to disregard the stages preceding the trialling and was therefore not particularly useful for the test development for the Austrian "Gymnasium". Bachman and Palmer (1996: 87-91) suggest three main stages of test development each of which involves at least three activities. The three stages are design, operationalization and administration. The first stage involves describing the purpose(s) of the test, defining tasks in the TLU domain, describing test takers' characteristics, defining the construct, developing a plan for evaluating the qualities of usefulness and finally identifying resources. In the second stage test tasks and a blueprint, which can be compared to test specifications, are developed; instructions are written and the scoring method is defined. The final stage comprises procedures for administering tests and collecting

feedback, procedures for analysing test scores as well as archiving. Although, or more to the point, because of the fact that Bachman and Palmer's model is very detailed it appeared to be too complex to be used as a basic framework for the development of the test for the Austrian "Gymnasium". The same is true for Hughes' (2003) model whose steps partly overlap with Bachman and Palmer's stages. Hughes (2003: 58ff.) suggests ten procedures without grouping them into stages, which makes his model less clear and therefore not useful for the study presented in this paper. The ten stages he lists are: stating the problem, writing specifications, writing and moderating items, informal trialling of items on native speakers, trialling on a group of non-native speakers, analysis of results of trialling and making necessary changes, calibration of scales, validation, writing handbooks for test takers, test users and staff and finally training staff.

From the different variants outlined above McNamara's (2000) categorisation into four stages was considered most practicable and clear and was therefore used as the framework for the development of the language test for the Austrian "Gymnasium". Once the test content and method are defined, test specifications are written. On the basis of these test tasks are designed. They are then trialled and changed if necessary. The trialling stage may be repeated in order to achieve satisfying tasks. The following sections will describe the development and administration of the language test presented in this paper with regard to McNamara's model (McNamara 2000).

9.1 ESTABLISHING TEST CONTENT

According to McNamara (2000: 25) establishing test content involves the decisions regarding what is being tested, in other words what the test construct is. This decision is based on "careful sampling from the domain of the test" (McNamara 2000: 25). The domain or TLU domain as it has been called throughout this paper has been discussed in detail in section 4.3. In the case of this language test the domain was defined with regard to language abilities in the areas of speaking, listening, reading and writing. As the teaching in both languages involved practising all four skills to a certain extent it was decided that the test content had to include all four skill areas. Of course, the extent to which the four skills were practised in class was not known to the group of test developers. It had to be assumed that the different teachers who taught the two language tracks favoured certain skills over

others and that their teaching focus varied. Nevertheless, the four skills were assumed to be part of the teaching because they all feature in the curriculum and were therefore relevant areas of language ability to be tested. The detailed description of the tasks for each skill can be seen in the specifications in the appendix (cf. section 15.2). For reasons of practicality the skill of speaking could not be tested on all pupils but only on a selected group of students. This issue will be considered further in the test administration in section 9.5.

9.2 ESTABLISHING TEST METHOD

Once the test content was defined the test method had to be considered. Test method is "the way in which candidates will be required to interact with the test materials, particularly the response format, that is, the way in which the candidate will be required to respond to the materials" (McNamara 2000: 26). Decisions on the test method involve numerous considerations with regard to the six test qualities described in chapter 5. While aiming at highly reliable, valid, authentic, interactive tasks, limits in terms of practicality are imposed on test developers (McNamara 2000: 28). Limitations in terms of practicality were, for instance, the time limit in the case of the speaking test. The time factor was the main reason for selecting a smaller number of test takers, which implied that the results were not as reliable as they had been with a larger test population[11]. For testing the other three skills of listening, reading and writing we decided to use a range of different conventional test formats which are widely used and accepted and which facilitate reliable scoring.

Generally, one can distinguish between various different task types: the three main categories are selected response, constructed response and personal response (Brown & Hudson 2002: 63). Each of these categories includes several item formats which are described in detail, for example, by Brown and Hudson (2002). In the present chapter only those formats which were used in the language test compiled for the Austrian "Gymnasium" will be described. These are all types which belong to the category of selected response items and all items of the constructed response format. Item formats of the category personal response (such as conferences, portfolios or self assessment) will not be considered here. The table below, taken

[11] A discussion of the speaking test can be seen in section 9.5.

from Brown and Hudson (2002: 64), gives an overview of the item formats and the categories of response type to which the item formats belong (cf. Table 4). The formats which were used in the language test presented in this paper are printed in bold.

Response type	Item format
Selected response	**Binary-choice (true/false)**
	Matching
	Multiple-choice
Constructed response	**Fill-in the blank**
	Short answer
	Performance
Personal response	Conferences
	Portfolios
	Self-assessment

Table 4: Overview of Language Test Types (Brown & Hudson 2002: 64)

Selected response

These are items in which students do not construct any language themselves but tick the correct answer from given options (Brown & Hudson 2002: 64). They are also called forced-choice items (Davies et al. 1999: 64). The overall advantages of all the item formats within this response type are the relative ease of objective scoring and the shorter time of administration (Brown & Hudson 2002: 64). The main disadvantage is that these items are rather difficult to create and that students do not actually construct any language. For this reason selected response items seem to be less appropriate for testing writing or speaking, but rather for the receptive skills of reading and listening. As an overall guidelines Brown and Hudson (2002: 65) suggest minimising the chance of guessing by dispersing the answers randomly. As all three item formats of this category have been used for this language test they will be described in some more detail.

Binary choice: Such items offer only two possible answers (yes/true or no/false) from which the test takers may choose (Davies et al. 1999: 215). The disadvantage is fairly obvious: with two possible answers the guessing factor of 50% is very high. It follows that the number of items has to be very large in order to gain reliable

results (Brown & Hudson 2002: 65). Therefore multiple choice questions have come to be favoured over dichotomous questions (Davies et al. 1999: 215). There is, however, a way of improving the true/false items by simply adding a third possible answer, which is “not given”. By adding the third option the chance of guessing is reduced to 33%. This extended form was used in the listening as well as in the reading section of the language test which was developed for this study. The tasks can be seen in the appendix (cf. section 15.1). One example of this item format, taken from the English reading test, can be seen below (cf. Example 1).

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 1: Selected Response Item

It can be argued that the guidelines for binary choice items by Brown and Hudson also apply to the form with three options (yes/no/not given). They suggest that the statements should be rather short and simple and that each question should only test one concept (Brown & Hudson 2002: 67). In addition, expressions which denote *absolutes*, such as *all*, *never*, *always*, etc. should be avoided as they are unlikely to be true and test takers may easily answer the question with their knowledge of the world. For the development of these tasks these suggestions were taken into consideration. For instance, no *absolute* expressions were used and the statements were, to a large extent, rather short.

Matching: This format in which test takers are asked to “match prompts to options” (Brown & Hudson 2002: 67) was used in the reading test. The pupils were asked to match headings to paragraphs within a newspaper article. According to Brown and Hudson (2002: 68) this format has two advantages: it is space-saving and the chance of guessing is rather low (10% for 10 items). However, it is “limited to measuring students’ abilities to associate one set of facts with another” (Brown & Hudson 2002: 67). Brown and Hudson (2002: 65-68) suggest four guidelines for item writers: firstly, it should be a relatively short list in order for the students to be able to keep track of all the items. Secondly, the lists should be homogenous in terms of content and form. Thirdly, all options have to be plausible and fourthly, more options than prompts should be created in order to minimize guessing.

Besides, the examinees should be told whether one option can be used more than once. In the case of our language test the format provided a good opportunity to test pupils' ability to assign headings to the paragraphs of the newspaper article. During the development all four guidelines described above were considered and realised. There were more options than prompts, the options were homogenous, plausible and randomly dispersed (in alphabetical order), and the list was relatively short as there were 14 prompts. The task can be seen in the appendix (cf. section 15.1). One illustrative example of a matching item can be seen below (cf. Example 2).

TASK: 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 be 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 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
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

Example 2: Matching Item

Multiple-choice: This test format was used on more than one occasion in the reading and listening part of the language test complied for the Austrian "Gymnasium". In this item format students are required to choose their answer from a given set of options (Davies et al. 1999: 123f.). The task usually includes a stem, which may be a question, or a statement which is to be completed, and a set of possible answers which includes the key and distractors. One advantage of this format is the relatively low guessing factor or 20%, 25% or 33% for 5, 4 or 3 options respectively. It is particularly useful as it can be used to test a variety of language aspects which, however, implies some disadvantages: multiple-choice formats are often overused or used inappropriately to test, for example, productive skills (Brown & Hudson 2002: 68ff.). Good multiple-choice items are often considered difficult to create. Therefore Brown and Hudson (2002: 65-70) suggest some guidelines to improve items of this format. Firstly, all options have to be grammatically possible answers. They should be of similar length and should not overlap. Additionally, the stem should not contain inadvertent clues by which the number of possible options is reduced. Some phrases, such as "*none of the above, A and B, but not C,* or *all of the above*" (Brown & Hudson 2002: 70) should be avoided as they require multiple cognitive processes which may confound test results. Finally, wordiness in the options should be avoided by reducing them to their distinctive parts only. The multiple-choice items of the English test were developed bearing in mind the suggested guidelines. Grammaticality was assured for all distractors. The length of the answers was relatively consistent and they were randomly dispersed. The multiple-choice items can be seen in the appendix (cf. section 15.1). One example taken from the English reading test can be seen below (cf. Example 3).

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.

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.

Example 3: Multiple Choice Item

Constructed Response

In contrast to the selected response format where test takers only have to choose an answer, this type of item requires examinees to create language by formulating their own answers (Davies et al. 1999: 32). Therefore this format seems most appropriate for testing the productive skills of writing and speaking (Brown & Hudson 2002: 71f.). One advantage of this format is that there is virtually no chance of guessing. This advantage, however, is gained at the cost of disadvantages during the scoring procedure. In contrast to selected response items the constructed response format is not only time-consuming but also difficult to score as scoring is partly subjective. Therefore it is very important for all judges to agree on a range of possible correct answers before the scoring is undertaken. By defining the correct answer in advance an attempt is made to render scoring as fair as possible (Brown & Hudson 2002: 72). The three item types which belong to the category of constructed response items are fill-in the blank, short answers and performance. All of them were used in the English language test (reading, listening and writing) and will be described subsequently.

Fill-in the blank: In this format test takers are required to fill in words or phrases in a given language context (Brown & Hudson 2002: 73). Fill-in tasks may vary with regard to the number of words missing and the structure of the language context. The advantages are the relatively easy development of such tasks, the short time of administration and their flexible use. There are, however, also disadvantages. One is that this format is limited to testing only single words or short phrases (Brown & Hudson 2002: 72). Additionally, in some situations there may be a large number of possible answers, which makes scoring difficult. In all cases the given language context should be sufficient and the blanks should be of same length and not too close together. This format was only used in the listening part of the language test

where the test takers had to listen for details and fill in the exact word(s). By restricting the answer to only one correct possibility scoring was facilitated and made objective. The tasks can be seen in the appendix (cf. section 15.1). One illustrative example can be seen below (cf. Example 4).

TASK 2: Insert the exact words you heard in the video. The number of words you need to fill in is shown by the number in brackets.

1. Unfortunately, Dawson's songs weren't nominated, as they weren't _____ (1).

Example 4: Fill-in the Blank Item

Short answer: This item format requires phrases and sentences to be produced by the test takers in response to statements or questions (Brown & Hudson 2002: 74). The advantages and disadvantages are very similar to the fill-in format. The variety of possible answers is even greater with this item type and theoretically, each test taker could produce an answer different from all the others. In order to avoid a wide range of answers the questions have to be worded very specifically and directly. To give the examinees an idea of the form a possible answer may take can be helpful. This item format was used in the reading part and the listening part of the language test and can also be seen in the appendix (cf. section 15.1). The following example taken from the English reading test will illustrate this item format (cf. Example 5).

TASK 5: Short Answer Questions

Answer the following questions 1-10 in very short phrases or with keywords. Indicate your answers below the questions.

1. Does Roger live in the same city in which he used to live during his childhood?

……..

Which lines tell you the answer?

……

Example 5: Short Answer Item

The problem of various different answers to one question was encountered frequently during the scoring process. Therefore it was of great importance for our project group to agree on a set of possible answers among the judges in order to be as fair as possible. This range of correct answers was defined before the scoring process and was then slightly changed during the process as unexpected but correct answers were encountered.

Performance: This item format requires test takers to produce a longer stretch of language, either spoken or written, or in a combination of more skills (e.g. reading and writing, listening and speaking) (Brown & Hudson 2002: 74). Thus this format seems most appropriate to test mainly the productive skills. One advantage is that the tasks simulate authentic language use, which may provide beneficial backwash (Brown & Hudson 2002: 72). Disadvantages arise with regard to practicality: performance tasks are difficult to create, take relatively long to administer and they may cause logistical problems. One main issue of this item format is subjectivity in scoring which affects reliability. Therefore such language performances have to be scored by several judges, which may result in higher costs. Performance tasks may take various different forms, such as "essays, problem solving assignments, role-playing, group tests, and communicative tasks" (Brown & Hudson 2002: 74). For the present language test this format was used for testing the productive skills of writing and speaking. The writing tasks took the forms of an argumentative essay and a personal letter. The speaking task was defined as a group discussion with role cards which the test takers were allowed to use. The exact instructions can be seen in the appendix (cf. section 15.1).

This section has described some very common test formats and their use in the language test developed in the context of the present study. In the construction of the test the guidelines described were considered and could be realised in most cases. The choice of format in the different test parts was largely based on considerations of usefulness: for the productive skills of writing and speaking none of the selected response formats were used as both skills are more reliably tested with constructed response items. For testing the receptive skills of reading and listening, however, all types of selected response formats were used. By including all of them the project team attempted to create a balanced variety of test formats in order to account for different preferences among the test takers. We tried to avoid misleading test results which would arise from a predominance of a particular test format. The same is true for the constructed response, which was used for testing receptive and productive language skills. Once the test method was agreed upon the next step in test development was the writing of the test specifications.

9.3 TEST SPECIFICATIONS

Test specifications may take different forms and may go through various stages of development (e.g. Lynch & Davidson 1994; Davidson & Lynch 2001; Brown & Hudson 2002). For the language test discussed here the specifications were created according to the guidelines of Brown and Hudson (2002: 87ff.) and therefore contain the following elements. Adapting from Popham (1981) Brown and Hudson (2002: 88ff.) suggest a general descriptor, which describes the overall purpose and aim of the test, specific test descriptors for each area or skill tested and finally, item specifications which are illustrated and defined by a general description, a sample item, prompt attributes, response attributes and, if necessary, specification supplements. The components which make up test specifications will be briefly discussed subsequently. It has to be noted however, that in the test specifications of the present study the general descriptor and the specific test descriptors were combined. Therefore the contents of the general descriptor can be seen in the specific test descriptors of each skill. (Full specifications can be seen in the appendix (cf. section 15.2)).

General descriptor: The general descriptor gives information on the objectives of the test and the test format (Brown & Hudson 2002: 87f.). It may also define what test takers who pass the test are able to do in each category of the test. As indicated above the general descriptor of the present study was described separately for each skill combining the general and the specific test specifications. Therefore an example of the general descriptor will be given after the description of the specific test descriptors.

Specific test descriptors: They should accompany the general descriptor and define which areas will be tested (Brown & Hudson 2002: 88). Within these areas components of the test may be listed and the level of proficiency should be defined there. The specific test descriptors for each skill tested at the present study can be seen in the appendix (cf. section 15.2). The following example combines the general descriptor and the specific test descriptors of the English reading test (cf. Table 5).

GENERAL TEST DESCRIPTOR	
subtitle	✓ Test of English Reading Comprehension Competence
description	✓ The test is designed to evaluate mastery of reading English at the matura level (final exam level), i.e. in the 12th school year. ✓ The test, a paper-and-pencil test, lasts 50 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 concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. [They] can understand contemporary literary prose." (Common European Framework of Reference for Languages: Learning, teaching, assessment, 2001: 27)[12] ✓ The test covers the English language skill area of reading and in particular reading comprehension competence.
structure	✓ The test consists of two parts, each of them dealing with a different text type: a newspaper article from a quality newspaper and a literary text. ✓ Examinees have to apply certain reading strategies in order to cope with the texts and the responses (e.g. skimming, scanning). The examinee demonstrates mastery of academic reading abilities, such as understanding the core context of the different text types, being able to detect detailed information as well as to gain a broad overview. The global and detailed understanding will be tested in different test formats (e.g. selected response, personal response, short answers).

Table 5: Test Descriptor of the English Reading Test

Item specifications: Item specifications consist of five components. Full specifications can be found in the appendix (cf. section 15.2). Some illustrative examples will be given here.

a) General description: this is a short and simple description "of the behavior being examined and how that behavior will be assessed" (Brown & Hudson 2002: 90). Thus it is analogous to the general test descriptor except that it only focuses on single test items and not on the test as a whole (cf. Example 6).

Reading Test Specifications – Newspaper Article	
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: o skimming for main idea

[12] Council for Cultural Co-operation, Education Committee, Modern Languages Division. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

	○ scanning for specific information ✓ extensive reading comprehension: ○ summary of a single text ○ answering questions according to the text

Example 6: General Description

b) Sample item: This example item should give test writers an idea of what test items should look like and provide the context for understanding further details in the item specifications (Brown & Hudson 2002: 94). Furthermore, Brown and Hudson suggest that sample items should accompany any instructions and directions given to test takers (cf. Example 7).

SAMPLE TEST ITEMS
a) Multiple Choice Exercise: ACCORDING TO THE TEXT, THE GOAL OF THE NEW READING TEST IS *a) to facilitate the entry to secondary education for underprivileged children* *b) to guarantee that children succeed on of the most basic stages in their cognitive development, namely the ability to read.* *c) to increase literacy among pupils from deprived backgrounds.* *d) to prevent children from future learning difficulties due to reading deficits.*
b) Matching Exercise: ***A*** *A Conservative government would particularly support underprivileged pupils.* ***B*** *A new system in learning to read will be implemented.* ***C*** *Abolishing the key stage one test will challenge pupils to focus on practising reading.* ***D*** *Abolishing the key stage one test will put the focus back on teaching reading.* ***E*** *…….*

Example 7: Sample Item

c) Prompt attributes: They define in detail what the prompts should contain (Brown & Hudson 2002: 94). These guidelines attempt to narrow down the variability of items which test writers produce in order to get congruent items by different item writers. Statements in the prompt

attributes section may refer to the length of the prompt, the genre, the linguistic level and the structure of the prompt (cf. Example 8).

PROMPT ATTRIBUTES / TEXT SPECIFICATIONS	
source	✓ the text is chosen from a quality newspaper from an English speaking country; broadsheets like for example: The Observer, The Times, The New York Times, etc.
topic	✓ the text will, in general, be unfamiliar to the test takers but the topic might be more or less familiar due to the test takers own interest in news ✓ the text shall not require any special or former knowledge from the test taker concerning vocabulary or the topic itself *(Do not choose a text, for example about the development of a political matter, which would require the information of prior headlines or articles of the media business.)* ✓ the text shall be controversial so that test takers can produce an opinion-based piece of writing afterwards
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
format	✓ include paragraph numbers (1, 2, 3, etc) on the left hand side of the text

Example 8: Prompt Attributes

d) Response attributes: In this section of the specifications the responses of the test takers are defined; in the case of selected response items the options from which test takers choose are described in detail (Brown & Hudson 2002: 94) (cf. Example 9). In the case of constructed response items the section provides guidelines for a response which is considered correct. It is important in this section to define exactly what is evaluated and how it is scored (cf. Example 10).

RESPONSE ATTRIBUTES / TEST ITEM SPECIFICATIONS
The student will respond to the newspaper article in 3 different ways: a) a multiple choice exercise b) a dichotomous exercise c) and a matching exercise
a) Multiple Choice Exercise: Considering the length and complexity of the text there will be about 6-8 multiple choice

exercises. Students are required to answer them with regard to the text. For each multiple choice item there are four possible answers, among them is at least one correct answer. The distractors are constructed in a way so that all options are grammatically possible. If more than one option can be correct, this will be stated explicitly in the rubric of the task. All given answers are ordered alphabetically and labelled with the letters a – d. The questions itself will be given number 1, 2, 3, etc.
b) Dichotomous Exercise: A statement with information on the text is given. The students have to decide if the statement is true, false or not given. Considering the length of the text there will be about 8-10 dichotomous items. The statements are numbered and appear in the order of the text.
c) Matching Exercise: A one-sentence summary of each paragraph is provided. Additionally there are four wrong summaries which serve as distractors. The distractors are constructed in a grammatically correct way and are only different from the correct version in a minor but decisive word or idea. The summarising sentences are ordered alphabetically according to the initial letter of the first sentence and then labelled with letters (A, B, C, etc). Students have to match one summary to each paragraph.

Example 9: Response Attributes / Test Item Specifications

SCORING
a) Multiple Choice Exercise: Test taker gets **half a point** for every correctly ticked or not ticked answer. *→ MAXIMUM OF 2 POINTS/QUESTION*
b) Dichotomous Exercise: Test taker gets **two points** for every correct answer.
c) Matching Exercise: Test taker gets **two points** for each correctly matched statement.

Example 10: Scoring

e) Specification supplement: This is a section in which important additional information on the content of items may be stated (Brown & Hudson 2002: 95). This is particularly the case when item writers need to be aware of certain limits for the content which items may contain; for instance, when gerund is tested, a list of verbs that can be used in the items can be provided. A specification supplement was not necessary for the study presented in this paper.

Once the test and item specifications have been written Brown and Hudson (2002: 95) suggest an evaluation of the specifications by experts in the field of language

teaching methodology. They should then evaluate the specifications with regard to item quality and content, which can be done with a form, provided by Brown and Hudson (2002). The next step involves revising and improving the specifications on the basis of the experts' comments. The final version of the specifications should then be given to the item writers who develop items which are based on the specifications. In the case of the language test presented here these recommendations could not be realised due to reasons of practicality. As the tests were to be developed in a relatively short time there was little time for revisions and evaluations of the test specifications. Furthermore, the specifications were not given to independent item writers as in this project the writers of the specifications were at the same time the item writers. This may have been the reason why there were no significant misunderstandings to be expected during the realisation of the test specifications. Although test specifications are relatively time consuming to produce they are important as they "provide a clear enough description so that any trained item writer using them will be able to generate items very similar to those written by any item writer" (Brown & Hudson 2002: 96). Once items have been written on the basis of the test specifications the test should be piloted before the administration.

9.4 TRIALLING

The fourth stage of test development involves trialling the test. As described earlier in the present chapter, piloting may be done differently and ideally more than once. Anstey (1966), for example, suggests piloting to be done three times: first on a "small group of interested people, [...] [then] on a large sample of the same kind of people on whom the test is to be used" (Davies 1990: 12) and finally a trialling of the test's final form. In the case of the language test presented here the trialling was carried out only once and followed McNamara's (2000) guidelines for trialling which will be outlined subsequently.

McNamara (2000: 32) suggests an appropriate trial population which resembles the actual test takers in important aspects, such as age, proficiency, background, etc. A trial population of 100 test takers is appropriate for discrete point test items but often more people will be required. The number of test items in the trialling should also be greater than needed in the final test (McNamara 2000: 60). In the case of the language test for the "Gymnasium" the trial population was much smaller than the number suggested by McNamara. The French test was trialled in a "Gymnasium" in

Lower Austria with 21 pupils taking part in the listening test and 20 pupils participating in the reading test. The English test was piloted in a "Gymnasium" in Upper Austria with 13 pupils. The examinees of the triallings had a relatively similar learning background, had the same age and were in the same form as the actual test population. The level of proficiency was expected to be about the same as the level of the target test population although there was a tendency to assume that – if discrepancies in the performances were to be observed – it would rather be the actual test population who would perform slightly better than the trialling population. This means that apart from the number of students, which was for reasons of practicality rather small, the conditions suggested by McNamara were fulfilled.

The results of the trialling should then be analysed statistically with regard to item facility, item discrimination, reliability coefficient and rater reliability, whereby the last concerns subjective judgements of speaking and writing (McNamara 30, 55ff.). For the scoring of these two skills it is important to train judges during this stage of test development and to evaluate the rating scales which are used. The piloting for the language test was only carried out for the receptive skills of listening and reading. This means that suggestions concerning testing for productive skills were not taken into account during the trialling stage and the rater reliability can therefore not be discussed with regard to the trial test[13]. Statistical analysis was carried out with regard to item facility only. Item discrimination, which "allows us to see if individual items are providing information on candidates' abilities consistent with that provided by the other items on the test" (McNamara 2000: 60) was not carried out in the trialling stage of our test development. Therefore item discrimination will not be included in the discussion.

Item facility, or item difficulty, gives information about the level of difficulty of a certain item by calculating the percentage of test takers who answered the item correctly (McNamara 2000:60f.). This calculation is done by dividing the number of test takers who answered the item correctly by the number of all test takers. The outcome can be expressed in numbers or percentages (1.0 = 100%). There is a certain range of percentages within which the item is considered acceptable. The ideal item facility would be of 50% (= .5); normally, however, the acceptable item facility ranges from .33 to .67. The insights from the analysis of the item facility were used for the revision of the test for the "Gymnasium". This means that items whose

[13] Inter-rater checks were carried out during the scoring of the actual test administration. The written compositions were scored by two different judges who compared their scores and agreed on a final score. The scoring procedure will be described in section 9.6.

item facility was either too low (i.e. less than 33%) or too high (i.e. more than 67%) were mostly deleted from the test. Some items with rather low item facility were kept in the test as they may "ease candidates into the test and to allow them a chance to get over their nerves" (McNamara 2000: 61). In the same manner, some rather difficult items (with a relatively low item facility) were kept in the test in order to be able to distinguish the very able test takers. A statistical analysis of the reliability of the test was not carried out in the trialling stage. It was, however, conducted for the final test. A discussion of the reliability of the final test can be seen in chapter 12.

In addition to statistical analyses, McNamara (2000: 32) suggests a feedback from the trial test population, which can be gathered in the form of a questionnaire. This feedback can be of great relevance as test takers may find it very easy to spot problematic items and unclear instructions which test developers may find difficult to see. This feedback was not gathered in the trialling of the language test for the "Gymnasium", but problematic and unclear instructions could be seen in the sheer variety of answers. For instance, some multiple choice items were answered unexpectedly and not consistently by the trial test population, which was ascribed to unclear instructions. These items were therefore deleted. In the case of one short answer question of the English reading test the trialling showed that the wording of the question was not explicit enough, which was improved after the trialling.

After the trialling stage the test needs to be revised, changed and improved on in the light of the trials. This, however, means that the newly revised test needs to be trialled again (McNamara 2000: 32). For reasons of practicality this could not be done in the case of the language test presented in this paper. If and in how far this may have influenced the reliability and facility value of the final test version will be discussed in chapter 11. The following section will describe the actual test administration and the test population of the English and the French language tests.

9.5 DESCRIPTION OF TEST ADMINISTRATION AND TEST TAKERS

The structure of the test administration was identical for both language tests with regard to test length, timing and order of test parts. Only the contents and the topics of the single test parts were different[14]. The English test was carried out on a

[14] The English test is attached in the appendix (cf. section 15.1). The French language test and a detailed discussion of it can be seen in *Sprachen testen - Hören und Lesen: Prüfen und Testen nach*

Tuesday morning from 9 am till 12 am. The order of the test parts was as follows: Firstly, the listening test was carried out. The second test part was the reading test and finally, the writing test was carried out. The speaking test was conducted with selected pupils and was carried out after the writing test.

To begin with the listening test, it consisted of three listening texts, two of which were audio files, one text was a video. All texts were played twice to the test takers. The instructions for the test tasks were written on the test sheets. The same text had been spoken and recorded from an English native speaker and was played to the test takers in addition to the written instructions. The listening test took 40 minutes which included pauses for revision. Item formats to be answered included multiple-choice items, short answer questions, true/false/not given items and fill-in items. Examinees were allowed to use dictionaries. After the listening test the examinees were asked to fill in a feedback sheet about the length, level of difficulty, interest and layout of the test. This feedback was used to gather information about possible improvements of the tests. It can be seen in the appendix (cf. section 15.3).

After a short break the reading test followed as the second part. It consisted of two texts: one newspaper article and one literary text. The test takers were allowed to use their dictionaries in order to answer questions in the form of multiple-choice items, short answer questions and true/false/not given items. This part of the test took 50 minutes. The feedback sheet which test takers filled in after the listening part was attached to the reading part as well.

The third part was the writing test. The test takers had to create two written compositions on the basis of the texts used in the reading part. One text was an argumentative essay on the topic of university tuition fees. The second text was a personal letter from the point of view of the literary protagonist to a former class-mate. The composition of both tasks had to be completed within 60 minutes. The use of dictionaries was allowed.

The fourth test part was the speaking test which was carried out in the form of a group discussion on the topic of the introduction of a smoking ban in Austria. This topic was discussed by four pupils each: two of them were supposed to be against the smoking ban and two of them should approve the introduction of a smoking ban.

---

europäischem Standard im Rahmen eines Sprachstandsvergleichs eines Wiener Gymnasiums - objektive, reliable, valide und authentische Leistungsmessung? (Diploma thesis by Silvia Schweinberger).

The test takers were given role cards with arguments to support their positions. However, the discussion was not a role-play as the examinees did neither have to stick to their position, nor to their arguments. These role cards were merely intended to support the test takers in their argumentation. Due to practical reasons this part could only be carried out with a smaller number of test takers. They were chosen from the two language tracks with regard to their language abilities estimated by their teachers. The intention of this selection was to get a broad and reliable picture of the ranges of language ability present in both language tracks. In the selection process it was attempted to create a balance with regard to the sex of the test takers. Ideally, two female and two male examinees should have comprised one group. However, during the test administration this intention could not be realised. Several test takers refused to take part in the speaking test. Therefore alternative examinees had to be found. Due to this the balance between female and male test takers, and more importantly, the balance between test takers from the two language tracks could not be perfectly sustained.

The structure and administration of the French test was nearly identical. Therefore only the differences will be briefly outlined. Generally speaking, there were fewer examinees taking part in the French test[15]. It was carried out one day after the English test from 10 am till 1 pm. In contrast to the English listening test the French test consisted of two videos and one audio file. The reading and the writing parts matched exactly. The speaking test was also carried out with a selected group of pupils. However, only few test takers agreed to take part in it. It was significant that pupils from the English-track, in particular, refused to participate in the group discussion. Consequently, alternative examinees had to be found, which resulted in a lack of test takers from the English-track[16].

As indicated in chapter 8, the test takers were pupils of an Austrian “Gymnasium” in their last year. The three classes of this year were divided into two language tracks. The English-track comprised two classes and consisted of 39 pupils. The French-track was comprised by one class with 26 pupils. As the two languages were tested on different days, the number of examinees varied between the two languages. The English test was taken by 34 pupils from the English-track and 16 from the French-track. In the French test 17 pupils from the English-track and 20 pupils from the

[15] The numbers of test takers of the English and the French tests can be seen in the table below (cf. Table 6)

[16] This imbalance will be considered in section 10.4 when the results of the speaking test are described.

French-track took part. In the following table the total number of pupils from the English- and the French-track is shown next to the number of pupils who took part in the tests. For the analysis only results from examinees who took all parts of the tests (i.e. listening, reading, writing) were used. The skill of speaking was tested in a smaller group of selected pupils and is not included in the following table (cf. Table 6).

	Total number of pupils	Test takers in the English test	Test takers in the French test
English-track	39	34	7
French-track	26	16	19
Total	65	50	26

Table 6: Number of Pupils and Test Takers in the English and French Tests (cf. Binder et al. 2008: 9)

9.6 SCORING

The scoring procedure is a crucial component of language testing and needs to be considered carefully in order to be able to make reliable and valid interpretations on the basis of test scores. The present section will describe the scoring methods applied in the test presented in this paper. The scoring of each item type can be seen in the specifications in the appendix (cf. section 15.2). With several item formats, such as short-answer questions, fill-in questions or true/false/not given items the scoring was relatively easy and unproblematic. The discussion here will focus on problems encountered during the scoring of the multiple-choice items and will then describe the scoring methods of the speaking and the writing tests in detail.

The multiple-choice items of the present language test were scored in a rather unconventional manner. It seems to be common practice in the field of language testing to award one point for each correctly answered multiple-choice item. This method was, however, not applied in the present test. As with several items there was more than one correct answer to be ticked, which was considered to be rather difficult for the test takers. In cases of, for instance, two correct answers (out of four possible answers) it was 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 possible answers: incorrect – partly correct – fully correct. This scoring method awarded half a point for each correctly ticked or not-ticked answer. This means that a fully correct answer was awarded with 2 points. In the course of the planning and developing stages of the language test this method was assumed to be a satisfying measure of fully and partially correct answers. The problems with this scoring method were only encountered during the actual scoring procedure. It turned out to be the case that most test takers got at least 1 or 1.5 of 2 possible points for each multiple-choice item as the correct decision not to tick an answer was awarded with half a point as well. This means that even test takers who did not tick any answer got at least 0.5 or 1 point. The result was that 0 points were hardly ever assigned, which distorted the overall results with regard to this item format. It also distorted the statistical analysis of the item facility which will be described in section 11 as for the statistical calculation only fully correct answers were considered. Partially correct answers were not included in the analysis, which reduced the facility value of several items. In a further analysis the facility value could be recalculated taking into account also the partially correct answers. For the study presented in this paper a recalculation was, however, not undertaken. As the scoring method had been defined in the test specifications it was maintained during the scoring despite the unsatisfactory results. The only change that was made during the scoring procedure concerned test takers who did not tick any answer. These items were then marked with a NA (no answer) and were awarded with 0 points. For future tests this scoring method will probably not be used again. A satisfactory alternative would be the common practice of assigning 0 points for incorrect or partly correct answers and 1 point for fully correct answers. 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.

With regard to the scoring of the writing and the speaking test, Brown and Hudson (2002: 72) give some advice on how to facilitate scoring, and thus reduce subjectivity in the case of performance tasks (cf. section 9.2). Firstly, the task should be clearly instructed and defined (Brown & Hudson 2002: 77). The assessors should know in advance how to score the test takers' performances and design a scoring scale. This will largely depend on the decision whether a holistic or analytic scoring system is used. Holistic scoring is a marking procedure "whereby raters judge a stretch of discourse (spoken or written) impressionistically according to its

overall properties rather than providing separate scores for particular features" (Davies et al. 1999: 75). Analytic scoring to the contrary awards scores for particular features instead of assigning one overall score (Davies et al. 1999: 7). For the scoring of the present language test these first two recommendations were taken into consideration. The instructions for the speaking and the writing test were worded very carefully and specifically and the scoring system was agreed on in advance. In the cases of both the speaking and the writing test analytic rating scales with four or more categories were used. Both scales can be seen in the appendix (cf. section 15.2).

The next three guidelines suggested by Brown and Hudson (2002: 77) are concerned with inter- and intra-scorer reliability, which was discussed in section 5.2.3. In order to achieve reliable results, raters should score all answers of a test part in one session and, if possible, anonymously. Language performances should always be scored by at least two different judges and finally, raters should be made aware of the possibility of *bluffing*. This means that test takers may try to either write around the topic or write so much that raters are likely to award higher scores. Both strategies may distort the accuracy of the results. These guidelines were considered in the test development for the present study. For the writing task, two different raters judged each performance independently. *Bluffing*, as described above, was tried to be counteracted by defining a word limit for an acceptable length of the essays. It is obvious that subjectivity can never be totally excluded from scoring compositions. In the study it might have been an advantage that the pupils were unfamiliar to the judges, which means that the raters did not have any preconceived expectations as to their performance. By considering the guidelines above we attempted to reduce subjectivity as much as possible in order to allow for fair scoring. After this summary of the procedures used in the scoring of the language test for the Austrian "Gymnasium" the next chapter will present the results from the actual implementation of the language tests.

10 RESULTS

This chapter will present the results from the language test in both English and French. The results for each skill and each language will be shown separately for the test takers from the English-track and the French-track. The analysis will include a description of the *mean* and, in the case of listening and reading, the statistics of

the *range* of scores and the *standard deviation*. In the final section the results will be summarised. Although the French language test has not been included in the discussion of the paper, the present chapter will include the French test results as the comparison between the two languages was the main purpose of the test. The results will then be analysed with regard to practicality and reliability in the next chapter. This analysis will show in how far the test results presented in this chapter can be said to be reliable.

10.1 LISTENING

As indicated in section 9.5 the listening test was the first part to be taken. The examinees had to answer questions on three different listening comprehension texts. The results of the three texts were merged and calculated for the two language tracks. For this calculation the results of all examinees who took part in the listening test were used. A total number of 50 examinees took part in the English listening test. Of these 50 test takers 34 were from the English-track and 16 were from the French-track. The French listening test was taken by 36 examinees. 20 of them were from the French-track and 16 were from the English-track. The following two figures present the results of the two language tracks on the English listening test and the French listening test in percentages (cf. Figure 5 and Figure 6). The percentages express the average score, i.e. the *mean* of the two language tracks with regard to the maximum of points.

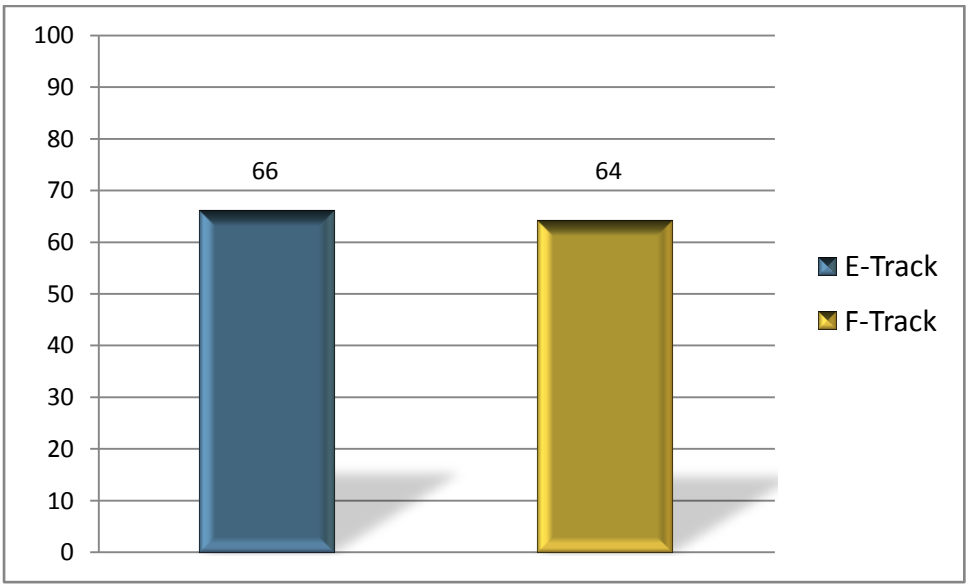

Figure 5: Results of the English Listening Test from the English-Track and the French-Track in %

The figure above (cf. Figure 5) shows that the English-track got an average of 66% of the maximum score in the English listening test. The French-track was slightly weaker in the English listening test with a percentage of 64%. The following figure (cf. Figure 6) shows the results of the French listening test. Subsequently, both results will be complemented by an analysis of the range of scores and the standard deviation.

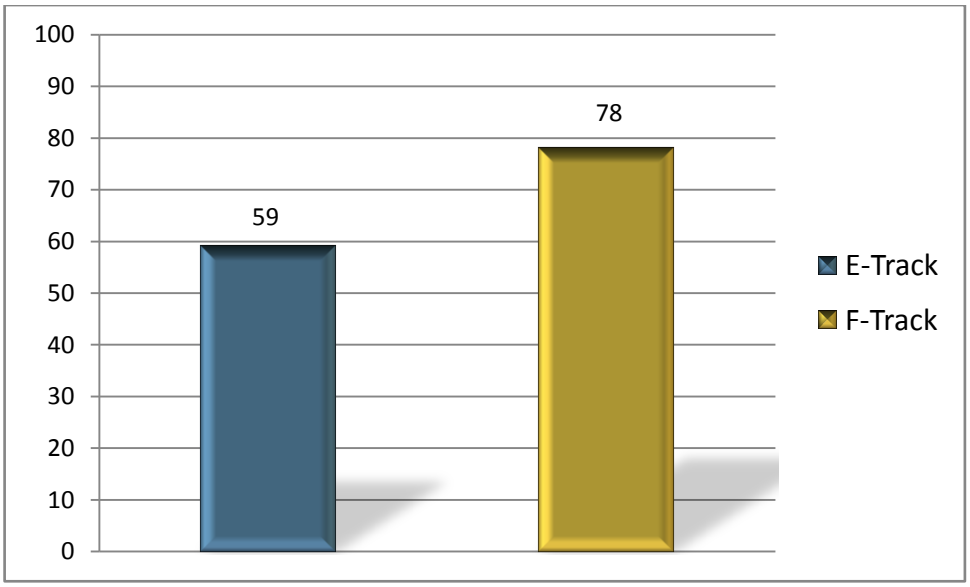

Figure 6: Results of the French Listening Test from the English-Track and the French-Track in %

The second figure (cf. Figure 6) shows the results from the French listening test. In contrast to the results on the English listening test there are significant differences. The English-track scored an average of 59%; the French-track exceeded them by 19% with an average score of 78%.

The figures show that the French-track got almost identical results in the English listening test and exceeded the English-track clearly in the French listening test. In addition to the results of the mean scores presented above the following statistical analysis will illustrate the range of scores and the standard deviation for the three classes. Class A is the French-track and classes B and C comprise the English-track. Both statistics give information about the spread of scores, i.e. how widely the scores of all test takers are spread out. The *range* shows the difference between the lowest and highest scores from all test takers (Alderson et al. 1995: 95). A small range means that the test takers' performances are rather homogeneous with a very narrow spread. A wide spread implies the opposite. The weakness of the spread of scores is that it only considers the top and the bottom scores but not the scores of

all test takers. Therefore the *standard deviation* should be calculated as an additional measure of dispersion. The standard deviation indicates the average amount of difference between the test takers' scores and the mean (Alderson et al. 1995: 95). Both statistics are used to give a more precise picture of test takers' performances (cf. Table 7).

Listening test			
		English (maximum of 99 points)	**French** (maximum of 101 points)
A		16 test takers	20 test takers
	Range	45.5 – 76.5 = 31	61 – 96.5 = 35.5
	Standard deviation	10.55	10.31
B		14 test takers	10 test takers
	Range	38.5 – 82.5 = 44	41 – 89.5 = 48.50
	Standard deviation	13.17	14.59
C		20 test takers	6 test takers
	Range	54 – 85 = 31	47 – 63 = 16
	Standard deviation	9.50	7.22

Table 7: Range and Standard Deviation of the Listening Test in English and French (cf. Binder et al. 2008: 11)

It is interesting to see that the French-track showed relatively homogeneous results. The spread of scores of the French listening test ranges from 61 (i.e. slightly above 60% of the maximum points of 101) to 96.5, i.e. 97.5% of the maximum points. The spread of scores of the English listening test is similar. The lowest score is 45.5 points, i.e. 45.0% of the maximum points. The top score on the English listening test was 76.5, i.e. 75.7% of the maximum points of 99. Although the highest score on the English listening test is not as high as the top score on the French listening test the French-track showed balanced performances with a relatively small standard deviation of 10.55 for the English test and 10.31 for the French test. The figures of the range and the standard deviation imply that the performance of the French track was relatively homogeneous.

The statistical analysis of the results from the English-track, comprised by classes B and C, is rather problematic. Due to the small number of participants, particularly of class C, the statistics cannot be said to be perfectly reliable. Nevertheless, a certain tendency is to be observed: the range of scores and the standard deviation of class B are relatively high, which implies a wide range of different levels of performances within this class. Although the top scores on the English listening test from classes B and C (i.e. 82.5 and 85) are higher than the top score from the French-track (i.e. 76.5) the mean of both language tracks is almost the same (66% for the English-track and 64% for the French-track). The reason for this is the wide range and the high standard deviation of the English-track and the relatively homogeneous performance of the French-track. With regard to the French listening test it can be said that the top scores of both classes from the English-track (i.e. 89.5 for class B and 63 for class C) were lower than the highest score of the French-track (i.e. 96.5) and the English-track also showed a lower mean (cf. Figure 6). These statistics support the impression from the mean in the figures above: the French-track generally performed more homogeneously than the English-track and got higher scores in the French-listening test.

10.2 READING

The reading test was the second part of the language test and included two texts of different genres. The overall results were calculated for the two language tracks. For the calculation only the results of examinees that completed both reading texts were used. A total number of 51 examinees took part in the English reading test. Of these 51 test takers 35 were from the English-track and 16 were from the French-track. The French reading test was taken by 34 examinees. 20 of them were from the French-track and 14 were from the English-track. The following two figures present the results of the two language tracks on the English reading test and the French reading test in percentages (cf. Figure 7 and Figure 8). The percentages express the average score, i.e. the *mean* of the two language tracks with regard to the maximum of points.

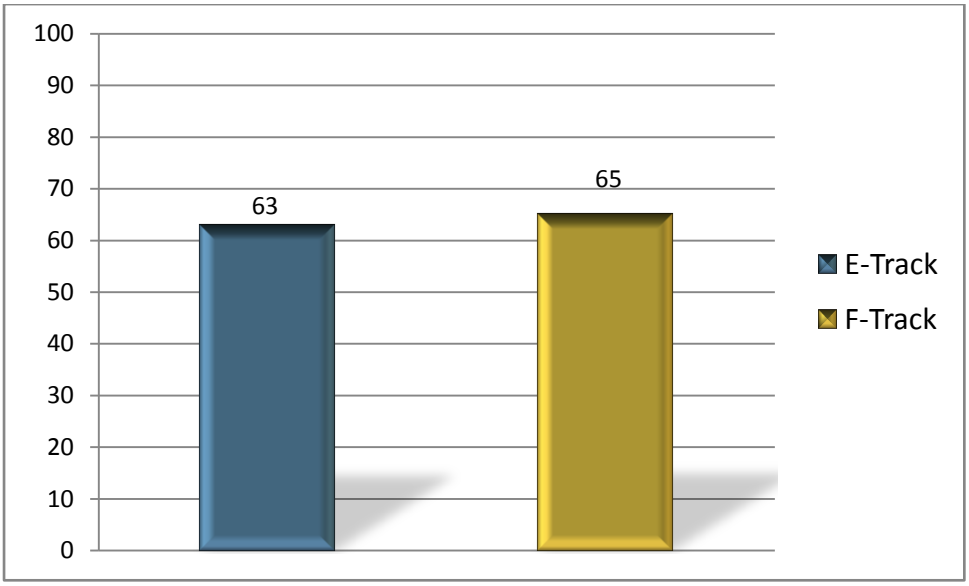

Figure 7: Results of the English Reading Test from the English-Track and the French-Track in %

The figure above (cf. Figure 7) shows that the results of the English reading test were almost identical for the English-track and the French-track. The former got an average of 63%, the latter scored 65% on average.

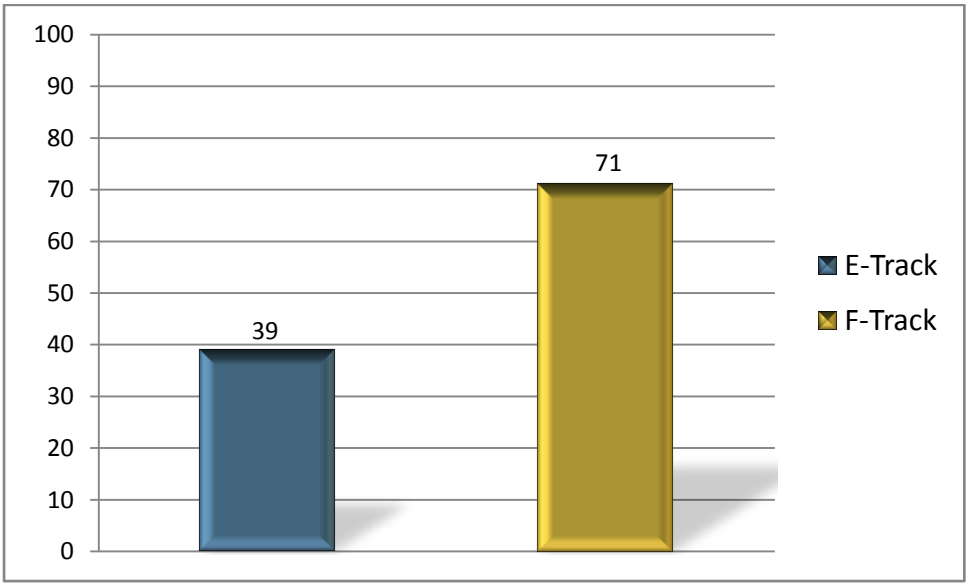

Figure 8: Results of the French Reading Test from the English-Track and the French-Track in %

In a similar manner to the results of the French listening test, the results of the French reading part are very different to the results of the English reading part (cf. Figure 8). The English-track scored an average of 39% whereas the French-track achieved nearly double this value with 71%.

The results from this part of the test show the same tendency as described for the listening part. The English reading test was accomplished almost equally well by the English-track and the French-track. To the contrary, the French reading test was mastered significantly better by the French-track than by the English-track. The calculation of the mean will again be complemented by an analysis of the spread of scores and of the standard deviation. The following table (cf. Table 8) shows the range of scores and the standard deviation with regard to the reading tests in English and French per class.

Reading test			
		English (maximum of 83 points)	**French** (maximum of 70 points)
A		16 test takers	20 test takers
	Range	41.5 – 62 = 20.5	36.5 – 67.5 = 31
	Standard deviation	6.23	9.28
B		14 test takers	8 test takers
	Range	32.5 – 71 = 38.50	9 – 40.5 = 31.5
	Standard deviation	11.6	11.5
C		21 test takers	6 test takers
	Range	26.5 – 67 = 40.5	10 – 60 = 50
	Standard deviation	9.26	17.18

Table 8: Range and Standard Deviation of the Reading Test in English and French (cf. Binder et al. 2008: 15)

Similar to the listening test the French-track showed relatively homogeneous results in the reading test. The scores of the French reading test are spread from 36.5 (i.e. slightly above 50% of the maximum points of 70) to 67.5, i.e. 96% of the maximum points. The spread of scores of the English reading test is similar. 41.5 points are 50% of the maximum points. The highest number of scores on the English reading test was 62 which is a percentage of 75% of the maximum points of 83. The rather low standard deviation is a further supportive indication of the homogeneous performances of the French-track.

The results for classes B and C are rather different. The performances among the test takers of the English-track are not as homogeneous as the performances of the

French-track. This can be seen in the spread of scores and the standard deviation which is in all parts of the reading test higher than the deviation of the French-track. Class B's scores on the English reading test range from 32.5 to 71 of a maximum of 83 points. This means that the best score is higher than the best score of the French-track. The lowest score, however, is a percentage of only 39.2%. The same tendency can be seen in the French reading test. Class B got scores from 9 to 40.5, i.e. from 12.9% to 57.9% with a standard deviation of 11.5. Similar results were scored by class C. The scores on the English reading test range from 26.5 to 67, i.e. from 31.9% to 80.7%. The standard deviation is relatively low. In contrast to the English reading test the scores on the French reading test show a rather high standard deviation and a range of scores from 10 to 60, i.e. from 14.3% to 85.7%. This spread is the highest of all parts of the reading test and implies rather diverse performances of the test takers from class C. Compared to the French-track it can be said that both classes of the English-track show less homogeneity in their performances.

10.3 WRITING

The writing test was the third part of the test. It required the test takers to create two different compositions on the basis of the texts used in the reading test. The compositions were judged by two independent assessors according to a scoring grid which can be seen in the appendix (cf. section 15.2). The number of test takers in the English writing test was the same as for the English reading test: a total number of 51 test takers, 36 from the English-track and 15 from the French-track. The number of test takers in the French writing test was much smaller: of 27 examinees only 8 test takers were from the English-track and 19 were from the French-track. The following figures show the results of the two language tracks on the English writing test and the French writing test in percentages (cf. Figure 9 and Figure 10). The percentages express the average score, i.e. the *mean* of the two language tracks with regard to the maximum of points.

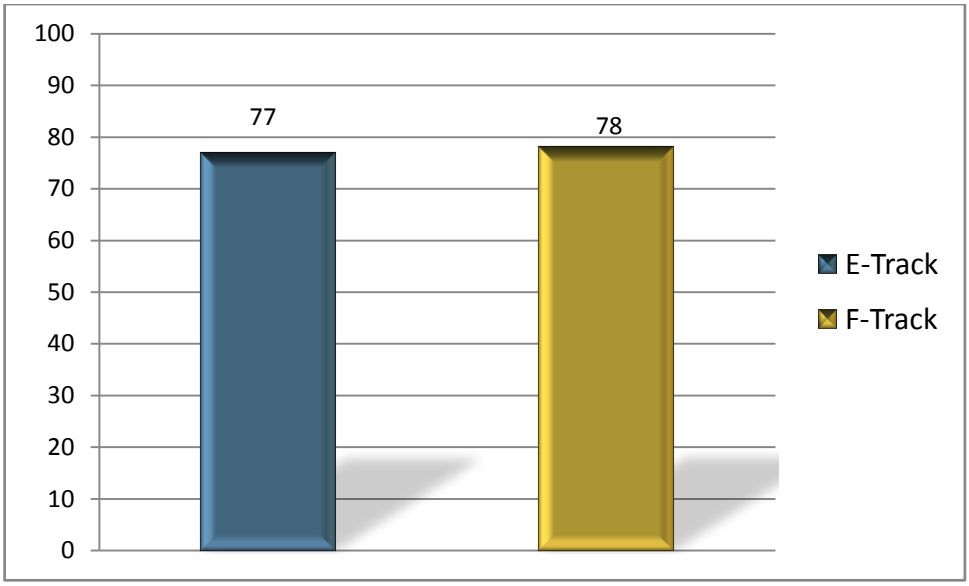

Figure 9: Results of the English Writing Test from the English-Track and the French-Track in %

The figure above (cf. Figure 9) shows that the results of the English writing test are almost identical. The French-track exceeded the English-track by only 1%. The English-track scored an average of 77%; the French-track an average of 78%. The following figure (cf. Figure 10) will present the results of the French writing test.

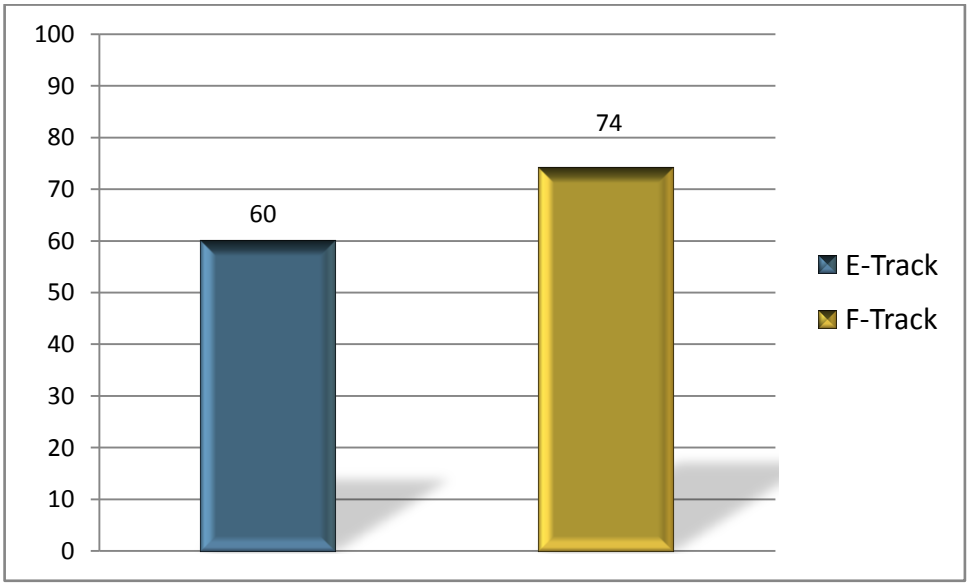

Figure 10: Results of the French Writing Test from the English-Track and the French-Track in %

In contrast to the writing results in English, there is a greater difference between the two tracks in the French writing test. The English-track scored 60% while the French-track scored an average of 74%. The results of this test part are similar to all results of the French language test above although the difference between the two language tracks is smaller than in the other skills. The French-track clearly

exceeded the English-track in the French writing task and scored almost identical results in the English writing test.

10.4 SPEAKING

For the last part of the test selected test takers were examined with regard to their speaking skills. For the English speaking test 11 test takers were examined. Six of them were from the English-track, five from the French-track. For the French speaking test 10 pupils were examined. Only one of them came from the English-track, nine were from the French-track. This unbalanced number was not the intention of the test developers. Half of the test takers were supposed to be from the English-track but most of them did not participate. Therefore alternate test takers from the French-track were examined in order to compensate the drop out rate. As indicated in section 9.5 not all examinees could be tested in this last part due to time limits. Therefore several pupils from both language tracks were selected to take part in the speaking test. The selection of test takers was intended to provide a balanced picture of the speaking abilities of the two language tracks. Due to the small number of test takers a presentation of the results in percentages may distort the results. Therefore the results of the speaking test will be presented in absolute numbers.

According to the CEFR the level of speaking ability of pupils in their last year should be B2. Therefore B2 was defined as the target level of this speaking test. With regard to the English speaking test most of the examinees reached the target level B2; five even reached the level C1. Four of the five test takers who achieved a language level of C1 came from the English-track. Thus it can be said that they showed a better level of language competence in the skill of speaking (cf. Table 9). With regard to the French speaking test the results are as follows: three examinees from the French-track reached the language level B2, three even reached level C1. The test taker from the English-track showed a speaking ability of B1, which is below the target level of B2. Due to the small number of test takers it is clear that these results cannot be said to be reliable and cannot be used as a generalisation on the English-track. The high drop out rate from pupils of the English-track does, however, show their reluctance to take part in the French speaking test. The following table (cf. Table 9) will present the number of test takers on the speaking test and the results with regard to the CEFR.

	Total number of test takers	Test takers from English-track	English-track at or above B2	Test takers from French-track	French-track at or above B2
English speaking test	11	6	6	5	4
French speaking test	10	1	0	9	6

Table 9: Results of the Speaking Tests from the English-Track and the French-Track

10.5 SUMMARY

The results presented above show that the English-track and the French-track achieved very similar scores in the English reading, listening and writing tests. By contrast, the results on the French test parts were significantly different. The greatest difference in scores was to be observed in the reading test in French; but also the results in the French listening and writing tests were rather dissimilar. The French-track scored better in all French test parts and additionally, exceeded the English-track minimally in the English writing test. Only the English speaking test was accomplished better by the English-track than by the French-track. These results imply that the French-track of this year showed a better overall language performance than the English-track. Reasons for this tendency may be seen in the extent of extracurricular activities in the French-track (cf. chapter 7). Moreover, the results above support the idea of learning French as the first foreign language. This appears to be an advantage as, currently, the English language seems to be present in every-day conversations, music and films and may therefore be strongly influential on individuals even outside any language program. French, on the other hand, does not seem to be particularly present in the Austrian society outside the classroom. Therefore it may be of great importance to start teaching and learning French explicitly at a very early stage. Of course, the results of this language test will be considered by the people who commissioned the test and it will be their concern to decide on any changes to be made with regard to the two language tracks of the “Gymnasium”. After this summarising presentation of the test results the next chapters will be concerned with an analysis of the test itself with regard to classical item analysis and Rasch analysis.

11 CLASSICAL ITEM ANALYSIS: FACILITY VALUE

Traditionally, classical item analysis, which comprises facility value, discrimination index and an analysis of the distractors of multiple-choice items, is carried out in order to gather information about the quality of test tasks (Hughes 2003: 225). By means of the facility value the level of difficulty of each item can be calculated. Item discrimination "allows us to see if individual items are providing information on candidates' abilities consistent with that provided by the other items on the test" (McNamara 2000: 60) and finally, the analysis of the distractors provides information about the manner and quality of multiple-choice items. Due to the limited scope of this paper the analysis of this chapter will only be undertaken for one of the three components, namely the facility value. It will be discussed with regard to a comparison between the pilot test and the final test version which was administered. Problematic items, i.e. items with a facility value above 67% or below 33%, from the final test version will be compared to the facility value of the trialling. By doing this the usefulness of the trialling will be examined. It has to be noted, however, that the trialling population was rather small as only 13 pupils took part in it. For a more reliable picture of the trialling a larger test population would have been desirable.

This analysis can merely present a limited selection of specific items as a detailed examination of all items would go beyond the scope of this paper. For the following discussion only items of the English listening test and the English reading test with a significantly high or low item facility value will be analysed. The language test and the relevant items can be seen in the appendix (cf. section 15.1). As indicated in section 9.6 partial scoring was applied with multiple-choice items. For the calculation of the facility value, however, only fully correct answers were used, which implies that the results of the multiple choice items may sometimes be rather low although some test takers nearly got the full range of points.

11.1 THE ENGLISH LISTENING TEST

The overall number of items from the English listening test is 48, spread over 10 tasks. Tasks 1 to 3 accompany the first listening comprehension text, tasks 4 to 7 accompany the second listening comprehension text and tasks 8 to 10 accompany the third text. Of these 48 items 19 were within the acceptable range of .33 to .67. 12 had a higher facility value and 17 were too easy. The facility value for all items was calculated for all test takers disregarding the language track.

It is significant that half of the items of the first listening comprehension text (tasks 1-3) were very difficult and their facility values range from .04 to .26. The four items with the lowest facility value will be compared to the pilot test subsequently. The following table will show the items which will be discussed with their facility values of the trialling and the actual test administration (cf. Table 10). The test can be seen in the appendix[17] (cf. section 15.1).

Number of item	Facility value of test	Facility value of trialling
Item1/2	.06	.00
Item2/6	.18	Not in the trial test
Item2/7	.04	.23 !
Item3/1	.22	.00

Table 10: Facility Values of Selected Items of the English Listening Test (Text 1)

From the table above it can be seen that a comparison between the trialling and the actual test administration is difficult in some cases. There is one item (Item 2/6) with a very low facility value in the actual test administration which does not appear in the trialling. The reason for this is that this item was used in a different form (i.e. with different answers) in the trialling and resulted in a very low facility value. Therefore the item was changed. Unfortunately, the revised item was not piloted again[18]. This might have shown that it was still too difficult. In this case a second piloting might have been very useful as an indication to the facility values to be expected. Item 2/7 is accompanied by an exclamation mark. This exclamation mark signifies that the item was slightly changed after the trialling. For item 2/7 this means that in the trial test version the question was asked in a true/false/not given format. After the trialling the test format was changed to a fill-in format in order to avoid too many different item formats. The result was a decreasing facility value of only .04. It would have been interesting to see how high the value would have been if the item had remained the same. As was generally the case, the revised item was not piloted before the actual test administration.

Item 1/2 and item 3/1 were used in the trialling and showed a facility value of .00, which means that no one answered these items correctly. They were, however, kept in the test and the facility values of both items increased. Nevertheless, they were

[17] In order to find the items which will be discussed they will be numbered the way they were in the test (e.g. task 3, item 4 = item3/4).
[18] This explanation also applies to two similar items (item 9/3 and item 9/8) which will be presented in subsequent tables.

still too difficult and could have been deleted from the test. In both cases the trialling gave a very precise picture of the facility value to be expected.

The items of the second listening comprehension text (tasks 4-7) were rather easier (cf. Table 11). Six of 15 items had a facility value of .68 and above. From these items the three items with the highest facility value will be considered further in the subsequent analysis.

Number of item	Facility value of test	Facility value of trialling
Item5/5	.9	.76
Item5/6	.8	.30
Item6/2	.84	.76 !

Table 11: Facility Values of Selected Items of the English Listening Test (Text 2)

Item 5/5 and item 5/6 are very interesting. The items were used in the same format in both test administrations and both resulted in a significant increase of the facility values. Item 5/5 rose from .76 to .9, which means that the value of the trialling was already above the acceptable limit of .67 and rose even more. To the contrary, item 5/6 was even below the acceptable benchmark of .33, rose to .8 and resulted in a facility value which was above the limit. This item, in particular, shows that test trials do not necessarily predict the facility value of the actual test administration satisfactorily. Item 6/2 was very interesting to analyse. The exclamation mark indicates that the item format used in the trial test was different to the format of the actual test administration. In the trialling the question was designed as a multiple choice item with only two given answers. The trialling resulted in a relatively high facility value of .76. As a result the item was changed into a fill-in format without any given answers. Interestingly, the facility value even increased after the change which was intended to make it more difficult.

The third listening comprehension text (tasks 8-10) was the longest. Of 20 items 7 were within the acceptable range, 10 were too difficult and three had a facility value above .74 and more. From these items the three items with the lowest facility value will be analysed. The most difficult item had a facility value of .02, which means that only one test taker answered this item correctly.

Number of item	Facility value of test	Facility value of trialling
Item9/3	.06	Not in the trial test
Item9/5	.02	.00 !
Item9/8	.12	Not in the trial test

Table 12: Facility values of selected items of the English listening test (Text 3)

From the table above it can be seen that there are again two items which were not used in the trial test (cf. Table 12). These items are item 9/3 and item 9/8. The reason why they were not used in the trial test has been stated above. The results of the actual test administration showed that both items were still too difficult and should have been trialled after their revision. Item 9/5 is one of the few examples of a rather precise prediction of the facility value in the actual test administration. The item was a multiple choice item with four given answers. In the trialling none of the test takers answered the item correctly. For this reason the item was revised by changing one of the distractors. Nevertheless, the results of the actual test administration were similarly low.

All items discussed show that a second trialling should in all cases be carried out when items are revised in order to gather facility values which are likely to predict the values to be expected in the actual test administration. However, some items which were not changed but used identically in both test administrations did not show similar facility values in the trialling and the test administration either. With regard to our testing experience it seems to be the case that items with extremely low facility values do not show significantly higher results with a different test population. However, it has to be emphasised once more that the trial population was very small and therefore perhaps not reliable enough.

11.2 THE ENGLISH READING TEST

The English reading test consisted of 46 items spread over five tasks. Of these 46 items 17 items had a facility value within the acceptable range of .33 and .67. Nine were too difficult and 20 items had a facility value of more than .67. This shows that the reading test was in general too easy for the test takers. The following analysis will consider items with the highest and the lowest facility values.

Of the seven items of the first task three were acceptable. The other four were slightly too difficult. Their facility values range from .2 to .31. Interesting for the

analysis of this task are two items with a facility value of .31. This value is very close to the acceptable benchmark of .33, which means that the items are only slightly too difficult. As they are so close to the benchmark the comparison of the facility value of the trialling is very interesting. It shows if the test takers from the trialling achieved similar results or if they accomplished the items better which would result in a higher facility value which would then be within the acceptable range.

Number of item	Facility value of test	Facility value of trialling
Item1/2	.31	.23
Item1/4	.31	**.38**

Table 13: Facility Values of Selected Items of the English Reading Test (Task 1)[19]

The first two items are very interesting (cf. Table 13). Their facility values in the actual test administration were close to the benchmark of .33. In the trialling the numbers are relatively different. The first item is too difficult with a value of .23 whereas the second item was within the acceptable range in the trialling. The first item was kept in the test because of the tentative assumption that the actual test population would perform slightly better than the trialling population (cf. section 9.4). The actual test showed that this assumption was incorrect as the result of item 1/4 was worse for the test takers of the actual test than for the trialling population.

Task 2 was very balanced with three acceptable items, three too easy and three that were too difficult. Of interest are the two items on either extreme, i.e. the one with the lowest and the one with the highest facility value. They will be analysed subsequently.

Number of item	Facility value of test	Facility value of trialling
Item2/5	.09	.15
Item2/6	.86	.076

Table 14: Facility Values of Selected Items of the English Reading Test (Task 2)

In the trialling of item 2/5 a facility value of .15 was achieved. Although this number is, of course, very low it is still higher than the value which was achieved in the actual test. The greatest difference between the two facility values of an item can be

[19] Facility values which are within the acceptable range of .33 to .67 are in bold print in the tables.

seen with item 2/6. Whereas the trialling population achieved a value of .076, which means that only one test taker answered this item correctly, the examinees of the actual test administration achieved a facility value of .86. This significant difference of .784 is impossible to explain and calls into doubt the trialling procedure. The wording of the question was the same for both test administrations and also the text to which the question refers was identical.

The third task consisted of 13 items, seven of which were acceptable. Two had a facility value below .33 and four were too easy. The items from this task are not considerably significant and will not be analysed in detail.

The items of task 4 are much more significant. All of the nine items were too easy with facility values of .82 and above. All items with a facility value of .9 and above will be compared to the facility value of the trialling in order to find out if the trialling population achieved similarly good results and if the items should have been removed perhaps (cf. Table 15).

Number of item	Facility value of test	Facility value of trialling
Item4/3	.92	.69
Item4/4	.92	**.61**
Item4/5	.94	.76
Item4/7	.9	.92
Item4/8	.9	.84

Table 15: Facility Values of Selected Items of the English Reading Test (Task 4)

The first three items of task 4 show the same tendency. The facility values of the trialling test were much lower than the values of the actual test administration. Item 4/3 was only slightly above the limit of .67 and item 4/4 was even within the acceptable range. Item 4/5 was above the limit but still significantly lower than the facility value of the actual test administration. For these three items the assumption that the actual test population would perform better than the trialling population seems to have been borne out by the test results. The last two items are not significantly different with regard to the two test administrations. The numbers show that both test populations performed almost identically and that both test items are too easy. They should have been deleted from the test as their high facility values show that they do not distinguish well between stronger and weaker test takers.

Task 5 was relatively balanced again with four items within the acceptable range and four items above and will not be discussed in detail.

To sum up, it can be said that the facility values of the trialling were only in some occasions good indicators of the performance which was to be expected from the actual test population. For five of nine items the facility values differ significantly. In the case of item 1/4 and item 4/4 the facility values of the trialling was within the acceptable range while the values of the actual test administration were too low for the former and too high for the latter item.

The comparison between the facility values of the English listening and the English reading tests shows that the results of the trialling should – if at all – only be taken as rough indicators for the performance to be expected. The facility values of the trialling can, of course, not be considered perfect indicators. This means that items with facility values close to the benchmarks of the acceptable limit should be kept in the test as an exact correspondence between the trialling and the test population is not very likely. For a useful and reliable prediction the trialling population would have to be significantly larger than it was in the case of this trialling. The number of the examinees taking part in the trialling was certainly too small in order to provide reliable results. Furthermore, the trialling should ideally be carried out in a school with language tracks similar to those of the school in which the test is to be administered. One further recommendation concerns the number of trial tests. In the case of the study presented in this paper the trying out was only conducted once. After significant changes of the test items and test length the final test version should have been trialled a second time in order to gather reliable results. Summarising, the results of the trialling of the present project cannot be said to be satisfactory indicators of the performance to be expected during the actual test administrations. The main reason for the limited usefulness is the fact that due to practical considerations the trialling could only be carried out once with too small a trialling population. For further testing projects the trialling needs to be considered more carefully.

12 ANALYSIS WITH REGARD TO PRACTICALITY AND RELIABILITY

While the last chapter was concerned with the classical item analysis, this chapter will analyse the language test developed for the Austrian “Gymnasium” with regard

to two of the six test qualities, which have been discussed extensively in chapter 5: practicality and reliability. The reason for this selection is the focus on an analysis of the reliability of the test, which will be discussed qualitatively and statistically. The fact that practicality often affects and restricts language test development and, in particular, reliability has been referred to above and will be considered further in the following section.

12.1 PRACTICALITY

Throughout the test development phase decisions concerning test format and scoring had to be made on the basis of practicality. In section 5.6 general considerations regarding practicality have been discussed. In this chapter the limitations and restrictions which had to be faced in the course of test development will be summarised. It can be said that most of them had an influence on reliability, which is why they are briefly presented here. The restrictions will be categorised into three distinct groups: before the test administration – test administration – scoring.

Generally speaking, many decisions have to be made before the actual test administration. Hughes (2003: 47), for example, suggests making test takers familiar with the test format in advance. In the case of our language test this was unfortunately not possible. The project started only a few months before the actual test administration, which made it difficult – if not impossible – to spend much time on preparing the test takers for the test. The test takers came from three different classes with different time tables and teachers. This means that some pupils may have been familiar with the test format in advance as their teachers may have introduced them to it in the course of teaching. Moreover, teachers might not have been happy to dedicate more time to practicing test formats. Within any school system projects such as our language test are not easy to put into practice, which is, in particular, due to time constraints.

A second decision before the test administration concerned rater training. Hughes (2003: 49) suggests training judges before actually scoring test performances. Again, time was the main problem with this issue. The judges were all – except for one native speaker – students of English. Due to the fact that the assignment was at short notice there was not enough time to train the judges. This had to be done during the scoring itself. Grading was then compared between different raters and

thus tried to be made as reliable as possible. Of course, training in advance would have been a desirable aspect.

During the test administration the most obvious restriction was the testing of the speaking skills. The decision to carry out the speaking test with selected pupils was made for reasons of practicality. While all other test parts could be taken at the same time by all examinees the speaking part could only be conducted successively. In each session four pupils were tested and scored by the same judges. Limitations in terms of practicality were that testing all pupils would have been very time-consuming. Time limits were set by the school authorities on whom collaboration depended. Moreover, there were difficulties to have a room at our disposal in which to test speaking. If there had been an additional room there would still have been a problem with the judges. We wanted the same judges to score all test takers in order to increase inter-rater reliability.

During the scoring process there was one obvious restriction. The written compositions were only scored by two different raters. Hughes (2003: 50) suggests at least two independent scorers. Ideally, a third person finally compares the scores given to analyse possible discrepancies. Such a third person was not involved in the scoring process of this language test.

This section has presented the main restrictions which became operative for reasons of practicality. Decisions which had to be made on the basis of practicality may very likely have affected the overall reliability of the language test. Therefore it will be the subject of the next section to attempt an analysis of the reliability of the test.

12.2 RELIABILITY

As discussed at length in section 5.2, reliability is a crucial language test quality which is decisive for overall test usefulness. For the purpose of an analysis of the overall usefulness of the test developed for the present project the following examination of the reliability will be done on two levels. Firstly, the language test will be analysed according to Hughes' advice on how to improve reliability (cf. section 5.2.4). Secondly, the reliability coefficient will be analysed statistically. For this analysis two different calculations which were done applying the Rasch model will be presented. The first calculation will show the reliability of the test in the form in

which it was administered. The second calculation will show how reliability may be improved by changing or deleting weak items.

12.2.1 ANALYSIS ACCORDING TO HUGHES

As described in section 5.2.4 there are several guidelines which may help to improve the overall reliability of a language test. Hughes (2003: 44ff.) gives 15 pieces of advice, eight of which refer to the overall reliability, seven concern the scoring reliability. As they have already been described they will only be listed again and then discussed separately with regard to the language test developed in the present project.

1. Take enough samples of behaviour
2. Exclude items which do not discriminate well between weaker and stronger students
3. Do not allow candidates too much freedom
4. Write unambiguous items
5. Provide clear and explicit instructions
6. Ensure that tests are well laid out and perfectly legible
7. Make candidates familiar with format and testing techniques
8. Provide uniform and non-distracting conditions of administration
9. Use items that permit scoring which is as objective as possible
10. Make comparisons between candidates as direct as possible
11. Provide a detailed scoring key
12. Train scorers
13. Agree acceptable responses and appropriate scores at outset of scoring
14. Identify candidates by number, not name
15. Employ multiple, independent scoring

ad 1) Take enough samples of behaviour

The more items a test has, the more reliable the results will be (Hughes 2003: 44). This does not, however, mean that there is no upper limit to the number of items. There is a fine line between having enough items and having too many. In the language test for the Austrian “Gymnasium” the number of items was reduced after the piloting. The trialling version included, for example, five listening comprehension texts, which was certainly too much as the test takers’ concentration decreased after three texts. Therefore the number of listening tasks was reduced to three texts with a total of 48 items. The two shorter texts were accompanied by 13 and by 15 tasks. The longer text was accompanied by 20 items. The number of tasks seemed to be appropriate and the length of 40 minutes for all three listening tasks (including time for preparation and revision) seemed to be manageable to the test takers. The number of items in the reading part was also reduced after the trialling. Unclear items and items with too low or too high item facility were deleted. The final number of 46 items for both reading texts was relatively easy to complete within 50 minutes by most test takers. The feedback from the test takers supports this assumption as a great majority considered the test length to be ‘okay’ (cf. Question 4 in the appendix; section 15.3).

ad 2) Exclude items which do not discriminate well between weaker and stronger students

Items which do not discriminate well either have a rather high facility value, i.e. they are answered correctly by most test takers or they have a very low facility value, i.e. they are answered correctly by very few test takers. Item facility has already been discussed in detail in section 9.4 and in chapter 11. Most of the items which did not discriminate well were deleted after the trialling, some of them, however, were kept in the test since they may help test takers overcome their nervousness (McNamara 2000: 61).

ad 3) Do not allow candidates too much freedom

Tasks which allow for very individual answers by the test takers are more difficult to score and therefore less reliable. In the language test presented in this paper selected response items were used predominantly for testing the receptive skills of listening and reading. These types of items limit the possibility of answers and are therefore more reliable. For testing the productive skills, however, constructed

response items were used. They are less reliable but more direct. This means that the risk of less reliable results was taken for the benefit of direct testing.

ad 4) Write unambiguous items

Ambiguous items were detected during the trialling stage and were then deleted. For instance, one multiple-choice item was intended to have only one correct answer. However, the scoring of the trialling showed that the formulation of this item was not explicit and clear enough, which allowed for two correct answers. Consequently, this item was deleted.

ad 5) Provide clear and explicit instructions

Instructions were attempted to be made very explicit and, in the case of the listening part, repeated several times. The instructions for the listening part were given in written and oral mode. Instructions for the writing and speaking tasks were presented in written mode and were formulated in a very specific and clear way in order to avoid ambiguity. Questions during the actual test administration are arguably not to be ascribed to unclear instructions but rather on the fact that some test takers did not read the instructions carefully enough.

ad 6) Ensure that tests are well laid out and perfectly legible

The test tasks were written in a clearly legible type font, point 12 or 11. The instructions were separated visually by putting them into frames. Layout was also a question on the feedback sheet which was given to the test takers (cf. appendix; section 15.3). It turned out that about 50% of every class considered the layout to be good. The reason for some test takers to consider the layout worse is very likely to be due to the fact that the sheets of the reading test were stapled together. Some test takers found it impractical and would have preferred paper clips which were used to hold together the sheets of the listening part.

ad 7) Make candidates familiar with format and testing techniques

Unfortunately, this was not possible in the course of the language test developed for the Austrian “Gymnasium”. The test takers came from somewhat different learning backgrounds as they were taught by different teachers with varying focus. Therefore it was not known to the project group whether or not test formats similar to the ones

used in the language test were familiar to the test takers. Familiarisation of the test takers with the test format in advance was not possible for reasons of practicality. However, it can be assumed that the format of performance was familiar to the test takers as this is a traditional format to test the writing skill.

ad 8) Provide uniform and non-distracting conditions of administration

In the language test the focus was on a comparison between students' abilities in English and French. The two tests were administered on two days but with the same conditions on both days. The test was taken in the same room and the test parts appeared in the same order. It may be argued that the test takers had an advantage on the second day as they were already familiar with the testing procedure. This, however, is then true for all candidates, which does not distort results between the two language tracks. It may lead to better results in the test on the second day. The effect of familiarisation after encountering a particular test format once is, however, to be questioned.

ad 9) Use items that permit scoring which is as objective as possible

Hughes' advice to use items which can be scored objectively does not mean that multiple choice items are the only format allowed. There are also other types of selected response items and constructed response items which permit relatively objective scoring. In the case of the language test for the Austrian "Gymnasium" the majority of items were selected response items. Short answers and fill-in the blank formats were used as well, the range of possible answers was, however, very limited, which increased objectivity.

ad 10) Make comparisons between candidates as direct as possible

This means that, in particular with performance tasks, the topics should be limited and clearly defined in order to facilitate comparing test takers' results (Hughes 2003: 49). This issue is only interesting with regard to the writing tasks in the language test. The tasks were designed so that there was little choice for the test takers. One task had to be fulfilled in the same manner by all of them. For the second writing task the examinees could choose between two headings. The format and structure, as well as the requirements of content were the same. Thus the comparison can be said to be very direct.

ad 11) Provide a detailed scoring key

A scoring key is of importance for scoring constructed response items where more than one answer may be correct. In particular with performance items a detailed key has to be used in order to minimize subjectivity of scoring. For all types of constructed response items of the test a scoring key was developed in advance. For the productive skills of writing and speaking analytic scoring keys were used. They can be seen in the appendix (cf. section 15.2). A list of acceptable answers to short answer items were agreed upon in advance, then revised and extended during the process of scoring.

ad 12) Train scorers

For scoring in which subjectivity is an issue raters should be trained how to score and how to use the scoring keys. In an ideal case patterns of scoring should be compared and analysed after each administration (Hughes 2003: 49). For the purpose of the language test developed for the Austrian “Gymnasium” rater training in advance was, unfortunately, not possible. Comparing and analysing scoring between different judges was, however, done in the case of written compositions and in the case of the speaking test.

ad 13) Agree acceptable responses and appropriate scores at outset of scoring

This has already been partly discussed in point 11. Hughes’ (2003: 49) advice to select archetypical examples of certain levels of proficiency was not realised. However, assigning appropriate scores was already done before the administration of the test and was included in the test specifications. Although in some cases the scoring turned out to be rather dissatisfactory it was retained as outlined in the specifications.

ad 14) Identify candidates by number, not name

Hughes (2003: 50) claims that inferences from names may affect assessors, even when test takers are unfamiliar to the judges. In the case of the language test in the Austrian “Gymnasium” the examinees were unfamiliar to the judges. Unfortunately, it was not always possible to score tests without knowing the names of the test takers. For the scoring of the listening and reading parts this was not of primary concern as most items were selected response items and therefore highly objective

to score. For the scoring of written compositions it may be argued that it is to a large extent the handwriting rather than the names which may affect judges' scoring.

ad 15) Employ multiple, independent scoring

Multiple, independent scoring increases reliability and was used in the case of written and spoken productions. All written compositions were scored independently by two different judges. The scores were then compared and discussed in order to come to a conclusion. The speaking test which was held in the form of a group discussion was scored by three non-native speakers and one native-speaker. The final score was again decided in the course of discussions among the judges.

The first part of this chapter was concerned with ways with which reliability can be improved. From the discussion above it can be seen that most suggestions were considered and put into practice in both the test development and scoring. The statistical analysis of the next section will show the overall reliability coefficient of the test.

12.2.2 RASCH ANALYSIS

The calculation was done for the listening and the reading part of the English test by using the Rasch model, which was described in section 5.2.1[20]. For the purpose of this chapter only selected results from the analysis will be presented as a thorough discussion of all aspects would go beyond the scope of this paper. The relevant results can be seen in the appendix (cf. section 15.4).

The following analysis of the language test will firstly present the overall reliability coefficient of the English reading and the English listening test. The reliability coefficient expresses "the extent to which the persons are reliably different in ability from each other [...] [which] is the Rasch equivalent of the usual (classical test theory)" (McNamara, lecture notes, 2008). Then the items which were detected to be the most problematic due to their *fit* value will be presented. Finally, the second statistical analysis which was done without the designated problematical items will show if and in how far the reliability coefficient improved compared to the first analysis.

[20] The actual calculations were carried out at the University of Melbourne. I would like to thank Professor Tim McNamara for making resources available to carry out this task.

As described in section 5.2.1 the reliability coefficient for listening tests should ideally be in a range of .80 to .89 (Lado 1961 quoted in Hughes 2003: 39). For reading tests the reliability coefficient should be above .9. The analysis of this language test showed that the reliability coefficient of the listening test was .72 and the coefficient of the reading test was .73. The first value is relatively satisfying, but could still be improved. On the other hand, the reliability coefficient of the reading test is rather low with regard to the range given above. Consequently, problematic items with a high *fit* value were identified. In the case of the English listening test these were four items (23, 28, 37, 46). Their *fit* values were 1.20, 1.86, 1.22 and 1.28. The items with the highest *fit* value of the English reading test were items 9, 10, 16, 25 and 39. Their *fit* values were 1.22, 1.18, 1.17, 1.26 and 1.23. The items can be seen in the tests in the appendix (cf. section 15.1). For the reason of their high *fit* values these items were considered negative for the overall reliability estimate. Once the weak items were identified they were deleted and the statistical analysis was carried out a second time.

The result was an increase in the reliability coefficient. For the English listening test the reliability coefficient rose by .07 from .72 to .79. A similar increase was to be observed for the English reading test. The coefficient rose from .73 to .79 which means an improvement of .06. Although the reliability coefficients of both test parts can still be improved further the deletion of the most problematic items showed an increase in the reliability coefficient. According to the acceptable range of the coefficient for reading tests, i.e. above .90 (Lado 1961 quoted in Hughes 2003: 39), the English reading test would still need further revision. Lienert & Raatz (1998: 14) however, accept a lower reliability value for the purpose of comparisons. According to them a coefficient between .5 and .7 is high enough in order to make reliable comparisons. As the purpose of these tests was a comparison between the two language tracks and not an exact measure of individual test takers' abilities the reliability coefficient of both parts of the test are high enough. This means that for the language test in the Austrian "Gymnasium" the results from this analysis were sufficient to justify the use of the language tests for the purpose of a comparison between the two language tracks with regard to English and French. However, further analyses with regard to test takers' inconsistencies and further problems of test items could be made in order to increase the reliability coefficient even more.

13 CONCLUSION

This thesis has presented the development of two language tests in English and French with the purpose of language program evaluation in an Austrian "Gymnasium". The research assignment was commissioned by the parents' association of the school who asked for an empirical and independent comparison of the pupils' language abilities of the two language tracks of the school. The focus of this paper was, however, not on the test results but on the description of test development and on a critical analysis of the reliability and overall usefulness of the tests. Only if the tests themselves show a high level of reliability, reliable interpretations on the basis of test scores can be made.

The test results were presented and discussed with regard to their significant differences. Although the results indicate a certain trend of language abilities of the two language tracks it will be the concern of the parents' association and the school authorities to draw conclusions from the results and to change the program accordingly. In order to base crucial decisions on a reliable source of information the test was continued in the following year and may have a final administration in the third year in a row. By doing so a broader and more reliable picture of the two language tracks can be achieved by counterbalancing class specific differences.

With regard to the CEFR it can be questioned in how far the items of the test correspond to the level B2. The CEFR was used as the basis for the choice of texts and item formats, the source of texts and the scoring of the speaking test. Although literature on the levels of certain test items is being developed, this field has not been fully investigated yet and will have to be the subject of further studies.

The main concern of this thesis was whether the reliability of the tests was high enough in order to be able to draw reliable conclusions from the test results. For this reason, the English test was analysed qualitatively and statistically. For the qualitative analysis the stages of test development and considerations of the six language test qualities were outlined with regard to the specific language test. From the discussion it can be seen that most suggestions which are standard in the field of language testing were put into practice. Of particular importance for the test development were the suggestions from Hughes (2003) who is concerned with language test development in school settings. Therefore his advice on how to improve reliability was discussed thoroughly. Nevertheless, or more to the point, because of the fact that the test was conducted in a school setting certain limitations

to the test development, administration and scoring had to be faced. These limitations of the study which can be said to have affected the reliability of the test were presented in the chapter on practicality as they were due to practical issues such as time constraints and financial considerations. Among them were, for instance, a second revision of the test and particular test items, a second trialling, an analysis of item discrimination and a Rasch analysis before the actual test administration. Taking into account all these limitations in a revision of the test and a second administration this would very likely improve the reliability value of the test further. Nevertheless, the purpose of the test was justified by its reliability value which was calculated statistically.

For the statistical analysis the classical item analysis and the Rasch method were used. The former showed the extent of correspondence of the facility value between the trialling and the actual test administration. The latter presented the reliability value of the test and showed the potential of improving this value. The results show that the reliability values of the English test could still be improved. However, according to Lienert and Raatz (1998: 14), who consider a reliability coefficient from .5 to .7 high enough for the purpose of comparisons, it can be concluded that the reliability value of the test achieved a satisfying level. Therefore the test can be said to be reliable enough in order to be used as evidence for the comparison of the pupils' language abilities of the two language tracks in the Austrian "Gymnasium".

14 BIBLIOGRAPHY

ALDERSON, J. C. 2005. *Assessing reading*. Cambridge: Cambridge Univ. Press.

ALDERSON, J. C. 2006. *Diagnosing foreign language proficiency: the interface between learning and assessment*. London: Continuum.

ALDERSON, J. C.; BERETTA, A. (eds.). 1992. *Evaluating second language education*. Cambridge: Cambridge Univ. Press.

ALDERSON, J. C.; CLAPHAM, C.; WALL, D. 1995. *Language Test Construction and Evaluation*. Cambridge: University Press.

ALDERSON, J. C.; URQUHART, A. H. 1985. "The effect of students' academic discipline on their performance on ESP reading tests". *Language Testing* 2, 2: 192-204.

ALTE 1998. *Multilingual Glossary of Language Testing Terms*. Cambridge: Cambridge Univ. Press.

ANASTASI, A. 1988. *Psychological Testing.* (6th edition). New York: Macmillan.

ANASTASI, A.; URBANA, S. 1997. *Psychological Testing.* (7th edition). New Jersey: Prentice Hall.

BACHMAN, L. F. 1986. "The test of English as a foreign language as a measure of communicative competence". In STANSFIELD, C. W (ed.). 1986. *Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference*. Princeton, NJ: Educational Testing Service, 69-88.

BACHMAN, L. F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford Univ. Press.

BACHMAN, L. F. (ed.). 1998. *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge Univ. Press.

BACHMAN, L. F. 2004. *Statistical analyses for language assessment*. Cambridge: Cambridge Univ. Press.

BACHMAN, L. F.; PALMER, A. S. 1982. „The construct validation of some components of communicative proficiency". *TESOL Quarterly* 16, 4: 449-65.

BACHMAN, L. F.; PALMER, A. S. 1996. *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford Univ. Press.

BAILEY, K. M. 1996. "Working for washback: A review of the washback concept in language testing". *Language Testing* 13(3), 257-279.

BINDER, M.; DERNTL, V.; HEIDER, M.; LOCH, H.; SCHWEINBERGER, S.; Sweeney-Novak, S.; Dalton-Puffer, C. 2008. *Vergleichende Sprachstandserhebung der 8. Klassen 2007/08* Project Report: MS University of Vienna (FDZE).

BM:UKK. 2008. "Development of education in Austria". http://www.bmukk.gv.at/medienpool/17147/develop_edu_04_07.pdf (17 April, 2009).

BROWN, A. 2005. *Interviewer variability in oral proficiency interviews*. Frankfurt am Main/Wien: Lang.

BROWN, J. D. 2004. *Language assessment: principles and classroom practices*. White Plains, N.Y.: Longman.

BROWN, J. D. 1989. "Language program evaluation: A synthesis of existing possibilities". In JOHNSONS, K. (ed.). *Program design and evaluation in language teaching*. London: Cambridge University, 222-241.

BROWN, J. D.; HUDSON, T. 2002. *Criterion-referenced language testing.* Cambridge: Cambridge Univ. Press.

BUCK, G. 2002. *Assessing listening*. Cambridge: Cambridge Univ. Press.

BYGATE, M. (ed.). 2001. *Researching pedagogic tasks: second language learning, teaching and testing*. Harlow: Longman.

CARROLL, J. B. 1961. "Fundamental considerations in testing for English language proficiency of foreign students". In Conference on Testing the English

Proficiency of Foreign Students. *Testing the English Proficiency of Foreign Students*. Washington, DC: Center for Applied Linguistics, 30-40.
CATTELL, R. B. 1964. "Validity and reliability: a proposed more basic set of concepts". *Journal of Educational Psychology* 55, 1-22.
COHEN, A. D. 1994. *Assessing Language Ability in the Classroom*. (2nd edition). New York: Heinle and Heinle.
COOK, T.D; REICHARDT, C. S. (eds*.).* 1979*. Qualitative and quantitative methods in evaluation research.* Beverly Hill, California: Sage.
CORSON, D. (gen. ed.). 1997. *Encyclopedia of language and education.* (7) Clapham, C. (ed.). 1997. *Language testing and assessment.* Dordrecht: Kluwer.
COUNCIL FOR CULTURAL CO-OPERATION/MODERN LANGUAGES DIVISION. 2001. *Common European framework of reference for languages: learning, teaching, assessment*. Council for Cultural Co-operation, Education Committee, Modern Languages Division. Cambridge: Cambridge Univ. Press.
CRONBACH, L. J. 1990*. Essentials of Psychological Testing*. (4th edition). New York: Harper and Row.
CUMMING, A.; BERWICK, R. (eds.). 1996. *Validation in language testing*. Clevedon: Multilingual Matters.
DAVIDSON, F.; LYNCH, B. K. 2001. *A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University.
DAVIES, A. 1990. *Principles of Language Testing*. Oxford: Basil Blackwell.
DAVIES, A. et al. 2000. *Dictionary of language testing*. Cambridge: Cambridge Univ. Press.
DOUGLAS, D. 1997. "Language for Specific Purposes Testing" In CORSON, D. (gen. ed.). *Encyclopedia of language and education* (7). Dordrecht: Kluwer: 112-120.
DOUGLAS, D. 2000. *Assessing languages for specific purposes*. Cambridge: Cambridge Univ. Press.
FULCHER, G. 2003. *Testing second language speaking*. London: Pearson.
FULCHER, G.; DAVIDSON, F. 2007*. Language testing and assessment: an advanced resource book*. London: Routledge.
GAGE, N. L. 1989. "The paradigm wars and their aftermath". *Educational Researcher* 18(7), 4-10.
GUBA, E.G. (ed.). 1990a. *The paradigm dialog*. Newbury Park, California: Sage.
GUBA, E.G. 1990b. "The alternative paradigm dialog". In GUBA, E.G. (ed.). *The paradigm dialog*. Newbury Park, California: Sage: 17-27.
GUMPERZ, J. J.; HYMES, D. H. (eds.). 1972. *Directions in Sociolinguistics: The Ethnography of Communication*. New York: Holt, Rinehart and Winston.
HOWE, K. R. 1988. "Against the quantitative-qualitative incompatibility thesis: or dogmas die hard". *Educational Researcher* 17(1), 10-16.
http://www.bmukk.gv.at/medienpool/17147/develop_edu_04_07.pdf
HUGHES, A. 2003. *Testing for language teachers*. (2nd edition). Cambridge: Cambridge Univ. Press.
HYMES, D. H. 1972. "Models of interaction of language and social life" In GUMPERZ, J. J.; HYMES, D. H. (eds.). *Directions in Sociolinguistics: The Ethnography of Communication*. New York: Holt, Rinehart and Winston: 35-71.
JOHNSONS, K. (ed.). *Program design and evaluation in language teaching*. London: Cambridge University.
JONG DE, J. H. A. L.; STEVENSON, D. K. (eds.). 1990. *Individualizing the assessment of language abilities*. Clevedon: Multilingual Matters.
LADO, R. 1961. *Language Testing*. London: Longman.
LIENERT, G.; RAATZ, U. 1998. *Testaufbau und Testanalyse*. (6. Auflage). Weinheim: Psychologie Verlags Union.

LINN, R. L. 1988. *Educational Measurement*. (3rd edition). New York: American Council on Education/Macmillan.
LUMLEY, T. 2005. *Assessing second language writing: the rater's perspective*. Frankfurt am Main/Wien: Lang.
LUTJEHARMS, M; CULHANE, T. (eds.). *Practice and Problems in Language Testing 3*. Brussels: Free University of Brussels.
LYNCH, B. K. 1990a. "A Context-Adaptive Model of Program Evaluation". *TESOL Quarterly* 24(1): 23-42.
LYNCH, B. K. 1990b. "The Author responds". TESOL *Quarterly* 24(4): 764-767.
LYNCH, B. K. 1996. *Language program evaluation: theory and practice.* Cambridge: Cambridge Univ. Press.
LYNCH, B. K. 1997. *Language program evaluation: theory and practice*. Cambridge: Cambridge Univ. Press.
LYNCH, B. K. 2003. *Language assessment and programme evaluation*. Edinburgh: Edinburgh Univ. Press.
LYNCH, B. K.; Davidson, F. 1994. "Criterion-referenced language test development: Linking curricula, teachers and tests". *TESOL Quarterly* 28, 727-743.
MCNAMARA, T. F. 1996. *Measuring second language performance*. London: Longman.
MCNAMARA, T. F. 1997. "Performance Testing*"*. In CORSON, D. (gen. ed.). *Encyclopedia of language and education* (7). Dordrecht: Kluwer: 131-140.
MCNAMARA, T. F. 2000. *Language Testing*. Oxford: Oxford Univ. Press.
MCNAMARA, T. F.; ROEVER, C. 2006. *Language testing: the social dimension*. Malden, Mass: Blackwell.
MESSICK, S. A. 1988. "Validity". In LINN, R. L. *Educational Measurement*. (3rd edition). New York: American Council on Education/Macmillan: 13-103.
PALMER, A. S.; GROOT, P. J. M. et.al. (eds.). 1981. *The Construct Validation of Tests of Communicative Competence*. Washington, DC: TESOL.
POPHAM, W. J. 1981. *Modern educational measurement*. Englewood Cliffs, NY: Prentice-Hall.
REICHARDT, C. S; COOK, T.D. 1979. "Beyond qualitative versus quantitative methods" In COOK, T.D.; REICHARDT, C.S. (eds*.). Qualitative and quantitative methods in evaluation research.* Beverly Hill, California: Sage: 7-30.
SHOHAMY, E. 1984. "Does the testing method make a difference?" *Language Testing* 1, 2, 147-70.
SHOHAMY, E. 1992. "Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning". *The Modern Language Journal* 76(4), 513-521.
SMITH, J. K. 1988. "The evaluator/researcher as person vs. the person as evaluator/researcher"*. Educational Researcher* 17(2), 18-23.
SMITH, J. K.; HESHUSIUS, L. 1986. "Closing down the conversation: the end of the quantitative-qualitative debate among educational inquirers"*. Educational Researcher* 15(1), 4-12.
STANSFIELD, C. W (ed.). 1986. *Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference*. Princeton, NJ: Educational Testing Service.
STEVENSON, D. K. 1981. "Beyond faith and face validity: the multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests". In PALMER, A. S.; GROOT, P. J. M. et.al. (eds.). 1981. *The Construct Validation of Tests of Communicative Competence*. Washington, DC: TESOL: 37-61.
STEVENSON, D. K. 1982. "The sociolinguistics of language testing: a tester's perspective". In LUTJEHARMS, M; CULHANE, T. (eds.). *Practice and Problems in Language Testing 3*. Brussels: Free University of Brussels: 4-22.

STEVENSON, D. K. 1985. “Authenticity, validity and a tea party”. *Language Testing* 2, 1, 41-7.
WALBERG, H. J. (ed.). 1992. *The international encyclopedia of educational evaluation*. Oxford: Pergamon Pr.
WALL, D. 1997. “Impact and Washback in Language Testing”. In CORSON, D. (gen. ed.). *Encyclopedia of Language and Education* (7). Dordrecht: Kluwer: 291-302.
WALL, D.; ALDERSON, J. C. 1993a. “Examining Washback: The Sri Lankan Impact Study”. *Language Testing* 10/1: 41-69.
WALL, D.; ALDERSON, J. C. 1993b. “Does Washback Exist?”. *Applied Linguistics* 14/2:115-129.
WEIGLE, S. C. 2002. *Assessing writing*. Cambridge: Cambridge Univ. Press.
WEIR, C. J. 1993. *Understanding and Developing Language Tests*. New York et al.: Prentice Hall.
WEIR, C. J. 2005. *Language testing and validation: an evidence-based approach.* Basingstoke: Palgrave Macmillan.
WOODS, A.; FLETCHER, P.; HUGHES, A. 1986. *Statistics in language studies*. Cambridge: Cambridge University Press.

15 APPENDIX

15.1 ENGLISH LANGUAGE TEST

Name:_____ Class:_____

Listening Comprehension

Listening comprehension: short video

You will see the following video (about 2 minutes long) twice.
- **You will have 2 minutes to read through the questions beforehand.**
- **Now you will watch the video for the first time.**
- **Then you will have 1 minute to answer the questions.**
- **Afterwards, you will watch the video for the second time.**
- **Then you will again have 1 minute to answer the questions.**

© MTV News: Daily Headlines
Kimya Dawson's Life Is Peachy Post-'Juno'

TASK 1: Tick the correct option(s) in the following Multiple Choice Tasks. Bear in mind that more than one statement can be correct.

1. The presenter thinks that
a. Ellen Page is the best new coming actress.
b. Ellen Page will win an Oscar.
c. "Juno" is one of the most special movies.
d. "Juno" is definitely the best movie of the year.

2. According to the presenter, the script of "Juno" can be described as
a. irresistible.
b. clever.
c. somewhat polarising.
d. very exciting.

3. According to the presenter, the music of Kimya Dawson
a. describes the thoughts and feelings of Juno.
b. is not very intellectual.
c. is deceptively childlike.
d. provides the right musical voice for Juno.

4. Kimya Dawson
a. enjoyed working on "Juno".
b. dislikes the director.
c. is a big fan of Ellen Page.
d. is not very happy to be on the interview.

5. She also adds that the hype about "Juno" is
a. great.
b. horrible.
c. okay.
d. surreal.

TASK 2: Insert the exact words you heard in the video. The number of words you need to fill in is shown by the number in brackets.

1. Unfortunately, Dawson's songs weren't nominated, as they weren't

_____ (1).

2. The reason for this is that the director only chose her _____

_____ (2).

3. Surprisingly, the album was _____ _____ (2) on the

Billboard Album Charts.

4. Some of the music was written as long as _____ _____

_____ (3).

5. The presenter thinks the way Dawson's songs _____ (1) the

character of Juno is almost eerie.

6. Dawson says that the reason for this is that her music has _____

_____ (2) to those of the film.

7. Dawson adds that she does not want to _____ _____ (2) with a major

record label.

TASK 3: Tick the correct option(s) in the following Multiple Choice Task. Bear in mind that more than one statement may be correct.

The video addresses:

a. Dawson's own experiences as a teenager
b. the parallels between Dawson's songs and the character of Juno
c. the reasons why "Juno" is such a special movie
d. the songs Juno plays in the movie

Listening comprehension: short video

You will see the following video (about 2 minutes long) twice.
- **You will have 2 minutes to read through the questions beforehand.**
- **Now you will watch the video for the first time.**
- **Then you will have 1 minute to answer the questions.**
- **Afterwards, you will watch the video for the second time.**
- **Then you will again have 1 minute to answer the questions.**

<u>TASK 4</u>: Tick ☑ the correct option(s) in the following Multiple Choice Tasks. Bear in mind that more than one answer may be possible!

1. The Maya developed a great society in Central and South America more than
a. 20 000 years ago.
b. 10 000 years ago.
c. 5 000 years ago.
d. 2 000 years ago.
e. 1 000 years ago.

2. Their society completely collapsed somewhere around
a. 900 BC.
b. 800 BC.
c. 100 AD.
d. 800 AD.
e. 900 AD.

3. The Maya
a. built towering pyramids.
b. built towering temples.
c. crafted artwork of ceramic.
d. crafted artwork of marble.
e. crafted artwork of stone.

4. They made advances in
a. arithmetic
b. astrology
c. astronomy
d. maths
e. physics

<u>TASK 5</u>: True or false? Decide whether the following statements are true or not, according to what you hear. If there is not enough information to answer, choose "not given".

1	The Maya mapped planetary movements.	□ True □ False □ Not given

2	In spite of their advances in various scientific fields, they could not create accurate calendars.	□ True □ False □ Not given
3	Experts once thought the Maya were a peace-loving, gentle civilisation.	□ True □ False □ Not given
4	The Maya were ruled by tribal chieftains.	□ True □ False □ Not given
5	The Mayan society performed rituals, violence and bloodletting.	□ True □ False □ Not given
6	Some of the violent customs of the Maya included human sacrifices in order to be able to predict the future.	□ True □ False □ Not given

TASK 6: Insert the exact words you heard in the video. The number of words you need to fill in is shown by the number in brackets.

1. In one ball game they would _____ _____ (2) your head.

2. At some point in history the Mayans began to abandon their

_____ (1) cities.

3. Today, roughly 6 million Mayan _____ (1) live in countries

like Mexico, Guatemala and Honduras.

4. Some Mayan descendents still speak Mayan languages and continue to carry

forward _____ (1) parts of this amazing culture.

TASK 7: What is the newscaster talking about? Tick the sentence that best summarizes the content of the news item!

He talks about:

a. the fact that an epidemic wiped out the Maya by millions, forcing the survivors to flee their great cities.
b. the fact that their own cruelty put an end to the Mayan civilisation.
c. the fact that the Maya's overuse of local water and land clearing strategies made it impossible to feed the population.
d. the fact that there are only various hypotheses on why this well-developed people essentially disappeared.
e. the fact that the Maya left their sophisticated cities because of foreign invasion and devastating wars.
f. the fact that mass revolts resulted in peasants overthrowing their rulers and escaping to the countryside.

Listening comprehension: long audio

You will hear the following audio file (about 10 minutes long) twice.
- **You will have 3 minutes to read through the questions beforehand.**
- **Now you will hear the recording for the first time.**
- **Then you will have 1 minute to answer the questions.**
- **Afterwards, you will hear the recording for the second time.**
- **Then you will again have 1 minute to answer the questions.**

© BBC Radio 4: Best of Today
Binge Drinking (Guest: Kevin Barron, Chair of Health Select Committee)

<u>TASK 8</u>: Read through the statements. Are they "true" or "false"? If there is not enough information to answer, choose "not given".

1	According to the police officer, it is necessary to enforce the law more strictly.	□ True □ False □ Not given
2	He also wonders why alcohol is sold cheaply in areas which are known to have many problems.	□ True □ False □ Not given
3	The police officer says that it needs to be researched why alcohol is advertised in certain ways to young people.	□ True □ False □ Not given
4	The drinks manufacturers should pay more taxes.	□ True □ False □ Not given
5	According to the spokesman of the drinks company, there is evidence that high prices will decrease the consumption.	□ True □ False □ Not given
6	He also adds that alcohol sales to other countries will be reduced if taxes are increased.	□ True □ False □ Not given
7	The spokesman of the drinks company says that higher taxes will be harmful for their business.	□ True □ False □ Not given
8	The prime minister refused to talk about the alcohol issue on the radio show.	□ True □ False □ Not given
9	According to the economics editor, binge drinkers will certainly buy less alcohol after the prices have been put up.	□ True □ False □ Not given
10	There are too many people stealing alcohol.	□ True □ False □ Not given

Roger Dean Kiser: **The Bully**

I walked into the Huddle House restaurant in Brunswick, Georgia and sat down at the counter. I picked up a menu and began to look at the various items trying to decide if I wanted to order breakfast or just go ahead and eat lunch.

A

I looked up and turned to the side to see a rather nice looking woman standing before me.

B

"Is your name Roger by any chance?" she asked me.

"Yes." I responded, looking rather confused as I had never seen the woman before.

"My name is Barbara and my husband is Tony," she said, pointing to a distant table near the door leading into the bathrooms.

C

"I'm sorry. I'm, ah. I'm ah, confused. I don't think that I know you guys. But my name is Roger. Roger Kiser," I told her.

D

"Tony Claxton. Tony from Landon High School in Jacksonville, Florida?" she asked me.

"I'm really sorry. The name doesn't ring a bell." I said.

E

I finally decided to order breakfast and a cup of decaffeinated coffee. I sat there continually racking my brain trying to remember who this Tony guy was.

"I must know him," I though to myself. "He recognizes me for some reason." I picked up my coffee and took a sip. All of a sudden it came to me like a flash of lighting.

"Tony. TONY, THE BULL." I mumbled, as I swung myself around on my stool and faced in his direction.

F

"The bully of my seventh grade geography class," I thought.

How many times that guy had made fun of my big ears in front of the girls in my class! How many times this son-of-a-gun had laughed at me because I had no parents and had to live in an orphanage! How many times this big bully slammed me up against the lockers in the hallway just to make himself look like a big man to all the other students!

He raised his hand and waved at me. I smiled, returned the wave and turned back around and began to eat my breakfast.

"Jesus. He's so thin now. Not the big guy that I remember from back in 1957," I thought to myself.

G

Tony had accidentally hit several plates knocking them off the table as he was trying to get into his wheelchair which had been parked in the bathroom hallway while they were eating. The waitress ran over and started picking up the broken dishes and I listened as Tony and his wife tried to apologize.

As Tony rolled by me, being pushed by his wife, I looked up and I smiled.

H

"Roger" he said, as he nodded his head forward.

I

[...]

"You remember. Don't you?" he said, looking directly into my eyes.

"I remember, Tony," I said.

"I guess you're thinking 'What goes around comes around'," he said, softly.

"I would never think like that, Tony," I said, with a stern look on my face.

9. Mr. Barron adds that, from his point of view,

a. alcohol is advertised to young people just as cigarettes used to be advertised.
b. parents should control their children´s drinking.
c. teachers need better preparation for this problem.
d. the government needs to look at this issue in depth soon.

<u>TASK 10</u>: Tick the correct option(s) in the following Multiple Choice Task. Bear in mind that more than one statement can be correct.

These points are addressed in the discussion:

a. The opinion of the drinks companies
b. The opinion of the ministry
c. The opinion of the police
d. What changes nightclubs could make

Reading Test

Tuition fees gain allure in cash-hit European campuses

Luke Harding in Berlin
Monday October 13, 2003
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. [...]

Name: ..... ***Class:*** .....

E-mail address : .....

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.

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
   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 eduacation 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
   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
   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 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
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

Roger Dean Kiser: **The Bully**

I walked into the Huddle House restaurant in Brunswick, Georgia and sat down at the counter. I picked up a menu and began to look at the various items trying to decide if I wanted to order breakfast or just go ahead and eat lunch.

A

I looked up and turned to the side to see a rather nice looking woman standing before me.

B

"Is your name Roger by any chance?" she asked me.

"Yes." I responded, looking rather confused as I had never seen the woman before.

"My name is Barbara and my husband is Tony," she said, pointing to a distant table near the door leading into the bathrooms.

C

"I'm sorry. I'm, ah. I'm ah, confused. I don't think that I know you guys. But my name is Roger. Roger Kiser," I told her.

D

"Tony Claxton. Tony from Landon High School in Jacksonville, Florida?" she asked me.

"I'm really sorry. The name doesn't ring a bell." I said.

E

I finally decided to order breakfast and a cup of decaffeinated coffee. I sat there continually racking my brain trying to remember who this Tony guy was.

"I must know him," I though to myself. "He recognizes me for some reason." I picked up my coffee and took a sip. All of a sudden it came to me like a flash of lighting.

"Tony. TONY, THE BULL." I mumbled, as I swung myself around on my stool and faced in his direction.

F

"The bully of my seventh grade geography class," I thought.

How many times that guy had made fun of my big ears in front of the girls in my class! How many times this son-of-a-gun had laughed at me because I had no parents and had to live in an orphanage! How many times this big bully slammed me up against the lockers in the hallway just to make himself look like a big man to all the other students!

He raised his hand and waved at me. I smiled, returned the wave and turned back around and began to eat my breakfast.

"Jesus. He's so thin now. Not the big guy that I remember from back in 1957," I thought to myself.

G

Tony had accidentally hit several plates knocking them off the table as he was trying to get into his wheelchair which had been parked in the bathroom hallway while they were eating. The waitress ran over and started picking up the broken dishes and I listened as Tony and his wife tried to apologize.

As Tony rolled by me, being pushed by his wife, I looked up and I smiled.

H

"Roger" he said, as he nodded his head forward.

I

[...]

"You remember. Don't you?" he said, looking directly into my eyes.

"I remember, Tony," I said.

"I guess you're thinking 'What goes around comes around'," he said, softly.

"I would never think like that, Tony," I said, with a stern look on my face.

TASK 4: Matching

You must choose which of the paragraphs 1-5 fit into the gaps *A* - *I* in the literary text. There is one paragraph for each gap. There are more gaps than paragraphs. Indicate your answers in the table below.

1) All of a sudden I heard the sound of dishes breaking so I spun around to see what had happened.

2) "Excuse me," said someone, as they touched me on the shoulder.

3) I looked in the direction that she was pointing but I did not recognize the man who was sitting alone at the table.

4) She turned and walked back to her table and sat down.

5) "Tony" I responded, as I nodded my head, in return.

A	
B	
C	
D	
E	
F	
G	
H	
I	

TASK 5: Short Answer Questions

Answer the following questions 1-10 in very short phrases or with keywords. Indicate your answers below the questions.

1. Does Roger live in the same city in which he used to live during his childhood?

.....

Which lines tell you the answer?

.....

2. How does Tony move in his wheelchair inside the restaurant?

3. Is Barbara sure that the man she asks is Roger? ..... Which words tell it? .....
4. What does the saying "What goes around comes around" in line 49 mean?
5. What do you learn about the living conditions in Roger's childhood?
6. What were the reasons for Tony to bully Roger? a) b) c) d)
7. Which line tells that Roger has forgiven Tony?
8. Which meal does Roger have in the restaurant?

Writing Test

TASK 1: Personal Response

Tuition fees are now a reality in Austria. Choose one of the following statements and write an argumentative essay on the topic. Use the chosen statement as the heading for your essay. Refer to the article above by supporting your main points with stated examples, opinions and given information. Your essay should not exceed a length of 250-300 words. Make sure you structure your text with paragraphs, introduction and conclusion.

A) Bad – Worse – Tuition fees

B) Tuition fees are a good thing: you get something for your money!

TASK 2: Personal Response

Imagine you are Roger Kiser. Write a letter to a former class-mate in which you tell him/her about the meeting with Tony, the Bull. Include some memories of your time at school with Tony and describe Tony's situation as it is now. Include information from the text! Make sure you follow the style of a personal letter. It should not exceed a length of 150-200 words. Use a separate sheet of paper.

Speaking Task

Smoking Policy in Austria

You can use any information you might have about this topic. Keep in mind:
- **You might stick to your role if you want, but you don't necessarily have to.**
- **Discussion time is between 10 and 15 minutes.**
- **You have to let the other test takers, including members of your group, take turns and hear them out.**

Role card Team 1

Find arguments to support your point of view described below:

- Smoking can have severe negative effects on your health
- Even second – hand smoking can cause cancer or other diseases
- Restaurant and bar staff are exposed to the harmful influence of smoking for several hours a day against their will
- Austria is lagging behind in terms of the current smoking policy in the European Union

Speaking Task

Smoking Policy in Austria

You can use any information you might have about this topic. Keep in mind:
- **You might stick to your role if you want, but you don't necessarily have to.**
- **Discussion time is between 10 and 15 minutes.**
- **You have to let the other test takers, including members of your group, take turns and hear them out.**

Role card Team 2

Find arguments to support your point of view described below:

- Smoking, especially in cafés, is part of Viennese lifestyle
- Austria is too cold a country to ask people to smoke outdoors
- Building smoking rooms means additional high costs
- Heavy losses are to be expected for bars and restaurants
- Small bars and cafés may not even be able to offer their guests smoking areas, and therefore might lose many customers

15.2 LANGUAGE TEST SPECIFICATIONS

Listening Test Specifications– General Specifications

GENERAL TEST DESCRIPTOR	
subtitle	✓ Test of English Listening Competence
description	✓ The test is designed to evaluate mastery of English listening comprehension at the matura level (final exam level), i.e. in the 12th school year. ✓ The test, a paper-and-pencil test, lasts 50 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 understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. [They] can understand most TV news and current affairs programmes. [They] can understand the majority of films in standard dialect." (Common European Framework of

	Reference for Languages: Learning, teaching, assessment, 2001: 27)[21] ✓ The test covers the English language skill area of listening comprehension.
structure	✓ The test consists of three parts, each of them dealing with a different text type: 1 (2 min) short video on popular culture 1 (2 min) short video on scientific topic 1 (10 min) long audio on contemporary topic ✓ The examinee demonstrates mastery of understanding authentic standard speech by answering Multiple Choice Questions, Short Answer Questions, True/False/Not Given Questions and Global Understanding Questions.

PROMPT ATTRIBUTES / TEXT SPECIFICATIONS	
source	✓ the videos / audio texts which are presented are taken from different broadcasting stations of English speaking countries. All texts are spoken in a standard variety. MTV News Science Channel BBC Radio today
topic	✓ the text will, in general, be unfamiliar to the test takers but the topic might be more or less familiar due to the test takers own interest in news ✓ the text shall not require any special or former knowledge from the test taker concerning vocabulary or the topic itself *(Do not choose a text, for example about the development of a political matter, which would require the information of prior headlines or articles of the media business.)*
length	✓ short text: 2 min ✓ long text: 10 min ✓ ideally ~180 words/minute ✓ short text: 13-15 items ✓ long text: 20 items
format	✓ all instructions are on the task sheet and read out ✓ questions are in chronological order

[21] Council for Cultural Co-operation, Education Committee, Modern Languages Division. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

	✓ only one type (MC, short answer,…) per task ✓ MC Questions in alphabetical order ✓ MC Questions: students get information on whether only one or more answers can be correct in the instructions ✓ Short Answer Questions: students know how many words to fill in ✓ Global Understanding Questions at the end of each listening
sequence	✓ Preparation time: short texts: 2 min long text: 3 min ✓ Listening for the first time ✓ 1 minute to answer questions ✓ Listening for the second time ✓ 1 minute to answer questions

RESPONSE ATTRIBUTES/ TEST ITEM SPECIFICATIONS

a) Multiple Choice: tick right answer(s)
b) Short Answer Questions: fill in right word(s)
c) True/False/Not Given: tick right answer
d) Global Understanding Question: tick right answer

SCORING

- Multiple Choice: 0,5 points/right choice
- Short Answer Questions: 2 points/correct item
- True/False/Not Given: 2 points/correct item
- Global Understanding Questions: 2 points/right choice

SAMPLE TEST ITEMS

a) Multiple Choice Questions:

According to the presenter, the music of Kimya Dawson

a. describes the thoughts and feelings of Juno.
b. is not very intellectual.
c. is deceptively childlike.
d. provides the right musical voice for Juno.

b) Short Answer Questions:

The presenter thinks the way Dawson's songs complement (1) the

character of Juno is almost eerie.

c) True/False/Not Given

The police officer says that it needs to be researched why alcohol is advertised in certain ways to young people	□ True □ False □ Not given

d) Global Understanding Question

These points are addressed in the discussion:

a. The opinion of the drinks companies
b. The opinion of the ministry
c. The opinion of the police

What changes nightclubs could make

Reading Test Specifications– General Specifications

GENERAL TEST DESCRIPTOR	
subtitle	✓ Test of English Reading Comprehension Competence
description	✓ The test is designed to evaluate mastery of reading English at the matura level (final exam level), i.e. in the 12th school year. ✓ The test, a paper-and-pencil test, lasts 50 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 concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. [They] can understand contemporary literary prose." (Common European Framework of Reference for Languages: Learning, teaching, assessment, 2001: 27)[22] ✓ The test covers the English language skill area of reading and in particular reading comprehension competence.
structure	✓ The test consists of two parts, each of them dealing with a different text type: a newspaper article from a quality newspaper and a literary text. ✓ Examinees have to apply certain reading strategies in order to cope with the texts and the responses (e.g. skimming, scanning). The examinee demonstrates mastery of academic reading abilities, such as understanding the core context of the different text types, being able to detect detailed information as well as to gain a broad overview. The global and detailed understanding will be tested in different test formats (e.g. selected response, personal response, short answers).

Reading Test Specifications – Newspaper Article

PROMPT ATTRIBUTES / TEXT SPECIFICATIONS	
source	✓ the text is chosen from a quality newspaper from an English speaking country; broadsheets like for example: The Observer, The Times, The New York Times, etc.
topic	✓ the text will, in general, be unfamiliar to the test takers but the topic might be more or less familiar due to the test takers own interest in news ✓ the text shall not require any special or former knowledge from the

[22] Council for Cultural Co-operation, Education Committee, Modern Languages Division. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

	test taker concerning vocabulary or the topic itself *(Do not choose a text, for example about the development of a political matter, which would require the information of prior headlines or articles of the media business.)* ✓ the text shall be controversial so that test takers can produce an opinion-based piece of writing afterwards
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
format	✓ include paragraph numbers (1, 2, 3, etc) on the left hand side of the text

RESPONSE ATTRIBUTES / TEST ITEM SPECIFICATIONS

The student will respond to the newspaper article in 3 different ways:

d) a multiple choice exercise

e) a dichotomous exercise

f) and a matching exercise

a) Multiple Choice Exercise:

Considering the length and complexity of the text there will be about 6-8 multiple choice exercises. Students are required to answer them with regard to the text. For each multiple choice item there are four possible answers, among them is at least one correct answer. The distractors are constructed in a way so that all options are grammatically possible. If more than one option can be correct, this will be stated explicitly in the rubric of the task. All given answers are ordered alphabetically and labelled with the letters a – d. The questions itself will be given number 1, 2, 3, etc.

b) Dichotomous Exercise:

A statement with information on the text is given. The students have to decide if the statement is true, false or not given. Considering the length of the text there will be about 8-10 dichotomous items. The statements are numbered and appear in the order of the text.

c) Matching Exercise:

A one-sentence summary of each paragraph is provided. Additionally there are four wrong summaries which serve as distractors. The distractors are constructed in a grammatically correct way and are only different from the correct version in a minor but decisive word or idea. The summarising sentences are ordered alphabetically according to the initial letter of the first sentence and then labelled with letters (A, B, C, etc). Students have to match one summary to each paragraph.

SAMPLE TEST ITEMS

a) Multiple Choice Exercise:

ACCORDING TO THE TEXT, THE GOAL OF THE NEW READING TEST IS

a) *to facilitate the entry to secondary education for underprivileged children*
b) *to guarantee that children succeed on of the most basic stages in their cognitive development, namely the ability to read.*
c) *to increase literacy among pupils from deprived backgrounds.*
d) *to prevent children from future learning difficulties due to reading deficits.*

b) Matching Exercise:

A *A Conservative government would particularly support underprivileged pupils.*

B *A new system in learning to read will be implemented.*

C *Abolishing the key stage one test will challenge pupils to focus on practising reading.*

D *Abolishing the key stage one test will put the focus back on teaching reading.*

E *…….*

SCORING

a) Multiple Choice Exercise:

Test taker gets **half a point** for every correctly ticked or not ticked answer.

→ *MAXIMUM OF 2 POINTS/QUESTION*

b) Dichotomous Exercise:

Test taker gets **two points** for every correct answer.

c) Matching Exercise:

Test taker gets **two points** for each correctly matched statement.

Reading Test Specifications– Literary Text

GENERAL DESCRIPTION

item description	Skill area description: READING Text type: LITERARY TEXT 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: o skimming for main idea o scanning for specific information ✓ extensive reading comprehension: o summary of a single text o answering questions according to the text

PROMPT ATTRIBUTES / TEXT SPECIFICATIONS

source	✓ the text is taken from a literary work or short story collection. ✓ the test takers are not familiar with the text they are confronted with.
topic	✓ the topic chosen does not require any specific cultural background knowledge and is neutral in order not to create an emotional or personal involvement or negative reaction. *For instance, religious topics or topics related to ethnicity are avoided.*
length	✓ 400 – 600 words ✓ The different passages will consist of unsimplified text at an advanced level of English. The texts will also include unnecessary information that the student will not need to answer the follow-up tasks.
format	✓ the text is presented in an easily readable font (i.e. at least size 12 in Arial or Times New Roman) ✓ Every five lines the number will be shown next to the line on the left side.

RESPONSE ATTRIBUTES / TEST ITEM SPECIFICATIONS

The test for the literary text has to be answered in the following ways:

a) matching exercises
b) short answer questions

a) Matching exercise:

The student will be asked to match sentences, which are taken from the text, to their original place in the text. There are more possible places indicated than sentences. The sentences are given numbers, the places where too fill them in within the text are labelled with letters.
b) Short answers exercise: The student will be asked to respond to the set of questions in the form of short answers. The short answers should be done in the form of phrases or keywords only. No full sentences are required here.
SAMPLE TEST ITEMS
A) SHORT ANSWER QUESTIONS *Who is meant by the "group" (line 69)?*

Writing Test Specifications– General Specifications

GENERAL TEST DESCRIPTOR	
subtitle	✓ Test of English Writing Competence
description	✓ The test is designed to evaluate mastery of writing English at the matura level (final exam level), i.e. in the 12th school year. ✓ The test, a paper-and-pencil test, lasts 60 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 write clear, detailed text on a wide range of subjects related to [their] interests. [They] can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. [They] can write letters highlighting the personal significance of events and experiences." (Common European Framework of Reference for Languages: Learning, teaching, assessment, 2001: 27)[23] ✓ The test covers the English language skill area of writing.
structure	✓ The test consists of two parts, each of them dealing with a different text type:

[23] Council for Cultural Co-operation, Education Committee, Modern Languages Division. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

	an argumentative essay and a letter. ✓ The examinee demonstrates mastery of academic writing skills, such as summarising unsimplified texts or creating expository compositions on a topic presented in a reading passage.

Writing Test Specifications – Argumentative Essay

PROMPT ATTRIBUTES / TEXT SPECIFICATIONS	
source	✓ the text which is the input for the writing task is the same as in the reading test: it is chosen from a quality newspaper from an English speaking country; broadsheets like for example: The Observer, The Times, The New York Times, etc. ✓ Two headings are stated in the task description. From these two the students have to choose and respond to only one. ✓ The length of the essay should not exceed a length of 250-300 words.
topic	✓ the text will, in general, be unfamiliar to the test takers but the topic might be more or less familiar due to the test takers own interest in news ✓ the text shall not require any special or former knowledge from the test taker concerning vocabulary or the topic itself *(Do not choose a text, for example about the development of a political matter, which would require the information of prior headlines or articles of the media business.)* ✓ the text shall be controversial so that test takers can produce an opinion-based piece of writing afterwards
length	✓ 800 – 1000 words ✓ the text is presented in an unsimplified version
format	✓ include paragraph numbers (1, 2, 3, etc) on the left hand side of the text

RESPONSE ATTRIBUTES/ TEST ITEM SPECIFICATIONS
The student will respond to the newspaper article with an argumentative essay. One of the two stated headings is the heading for their essay. The essay will include arguments and examples from the text to support their own arguments. The response is correct if a) the essay corresponds to the chosen heading b) arguments and information from the text are included c) the register is appropriate for an essay

d) the length of the essay agrees with the given limit

e) if the formal structure (paragraphs, introduction and conclusion) is realised

SCORING

The essay will be scored according to a scoring grid. Aspects of language that are scored are to be seen in the attached grid.

Students can get a total of 24 points.

Students can get 6 points for each category as can be seen in the following scoring grid will.

Furthermore, **one point** will be **deducted** in case of **disregarding given text length** (for part 2: ± 10% of words).

SAMPLE TEST ITEMS

Argumentative Essay:

Choose one of the following statements and write a critical letter to the editor on the topic of the newspaper article. Use the chosen statement as the heading for your letter. Refer to the article above by supporting your main points with stated examples, opinions and given information. Your essay should not exceed a length of 300-350 words.

a) Reading tests at 6 will ruin our children's self-confidence!

b) The earlier you read the better you develop

Writing Test Specifications – Letter

PROMPT ATTRIBUTE / TEXT SPECIFICATION	
source	✓ the text which is the input for the writing task is the same as in the reading test[24]: it is taken from a literary work or short story collection.
format	✓ the instruction for the writing task is attached to a separate sheet of paper. The writing task is to be written on this sheet. The length of the task is to be stated clearly. The length of the letter should not exceed a length of 150-200 words.

RESPONSE ATTRIBUTES/ TEST ITEM SPECIFICATIONS

The student will respond to the literary work with a personal letter. The letter will include information from the text.

The response is correct if

a) the letter includes information from the text

b) the register is appropriate for a personal letter

c) the length of the letter agrees with the given limit

d) if the formal structure is realised

SCORING

The letter will be scored according to a scoring grid. Aspects of language that are scored are to be seen in the attached grid.

Students can get a total of 24 points.

Furthermore, **one point** will be **deducted** in case of **disregarding given text length** by more than 50 words.

SAMPLE TEST ITEMS

Personal letter:

Imagine you are Roger Kiser. Write a letter to a former class-mate in which you tell him/her about the meeting with Tony, the Bull. Include some memories of your time at school with Tony and describe Tony's situation as it is now. Include information from the text! Make sure you follow the style of a personal letter. It should not exceed a length of 150-200 words. Use a separate sheet of paper.

[24] Please see specifications for the reading test

Beurteilung der Textproduktion in der Oberstufe

	AUSDRUCK Reichhaltigkeit und Anwendungsbereitschaft richtige Wahl und Verwendung Rechtschreibung; Verständlichkeit		**GRAMMATIK** richtige Anwendung; Fehler Verständlichkeit Vielfalt an Strukturen, Risikobereitschaft		**ERFÜLLUNG DER AUFGABENSTELLUNG*** INHALT :Sachwissen und Relevanz Textformat, Länge, Sprachregister		**AUFBAU** Absätze, Lesbarkeit; Satzlogik Satzverbindungen *editing*, Interpunktion
6	• Sehr großer Wortschatz; • sehr große Anwendungsbereitschaft; • Richtige und idiomatische Form, Auswahl und Anwendung des Vokabulars; • kaum Rechtschreibfehler; • Bedeutung der Wortwahl immer klar.	**6**	• Sicherer Gebrauch der passenden Strukturen und Formen, sehr guter Satzbau; • kaum Fehler bei Verbformen, Gebrauch der Zeiten, Plural, Wortstellung, Präpositionen, etc.; • Bedeutung des Satzes immer klar; • sehr große Vielfalt an Strukturen; • sehr große Anwendungsbereitschaft.	**6/ 3**	• Die Aufgabe wurde umfassend und eigenständig gelöst; • sehr großer Ideenreichtum und Anzahl an schlüssigen Argumenten; • sehr gute Sachkenntnis; • Text bezieht sich auf Themenstellung; • immer richtiges Format, Länge und Register.	**6/ 3**	• Sehr guter Aufbau; • sinnvolle Absätze, eindeutig lesbar; • sehr guter und variantenreicher Gebrauch von Satzverbindungen; • keine Flüchtigkeitsfehler; • richtige Interpunktion,
5	• Großer Wortschatz; • große Anwendungsbereitschaft; • meistens richtige und idiomatische Form, Auswahl und Anwendung des Vokabulars; • wenige Rechtschreibfehler; • Bedeutung der Wortwahl meistens klar.	**5**	• meistens richtiger Gebrauch der passenden Strukturen und Formen, guter Satzbau; • wenige Fehler bei Verbformen, etc.; • Bedeutung des Satzes meistens klar; • große Vielfalt an Strukturen; • große Anwendungsbereitschaft.	**5/ 2,5**	• Die Aufgabe wurde nahezu eigenständig gelöst; • großer Ideenreichtum und Anzahl an schlüssigen Argumenten; • gute Sachkenntnis; • Text bezieht sich fast immer auf Themenstellung; • richtiges Format, Länge und Register.	**5/ 2,5**	• Guter Aufbau; • meist sinnvolle Absätze, eindeutig lesbar; • guter Gebrauch von Satzverbindungen; • kaum Flüchtigkeitsfehler; • meistens richtige Interpunktion.
4	• Angemessener Wortschatz; • angemessene Anwendungsbereitschaft • einige Fehler in der Idiomatik, Auswahl und Anwendung des Vokabulars; • einige Rechtschreibfehler; • Bedeutung der Wortwahl nicht immer klar.	**4**	• einige Fehler im Gebrauch der Strukturen und Formen, angemessener Satzbau; • einige Fehler bei Verbformen, etc.; • Bedeutung des Satzes nicht immer klar; • angemessene Vielfalt an Strukturen; • angemessene Anwendungsbereitschaft.	**4/ 2**	• Die Aufgabe wurde angemessen gelöst; • angemessener Ideenreichtum und angemessene Anzahl an schlüssigen Argumenten; • angemessene Sachkenntnis; • Text bezieht sich meistens auf Themenstellung; • Format, Länge und Register meistens richtig.	**4/ 2**	• angemessener Aufbau; nicht immer gut strukturiert • Absätze manchmal vergessen oder nicht sinnvoll, großteils lesbar; • angemessener Gebrauch von Satzverbindungen; • einige Flüchtigkeitsfehler; • wenige Fehler in der Interpunktion.
3	• Ausreichender Wortschatz; • gerade noch ausreichende Anwendungsbereitschaft • beträchtliche Anzahl an Fehlern in der Idiomatik, Auswahl und	**3**	• beträchtliche Anzahl an Fehlern im Gebrauch der Strukturen und Formen, nicht immer angemessener Satzbau; • beträchtliche Anzahl an Fehlern bei	**3/ 1,5**	• Die Aufgabe wurde gerade noch ausreichend gelöst; • ausreichender Ideenreichtum und ausreichende Anzahl an schlüssigen Argumenten;	**3/ 1,5**	• Ausreichend strukturiert; • Absätze häufig vergessen oder nicht sinnvoll, stellenweise unlesbar; • ausreichender und meistens

	Anwendung des Vokabulars; • beträchtliche Anzahl an Rechtschreibfehlern; • Bedeutung der Wortwahl manchmal unklar.		Verbformen, etc.; • Bedeutung des Satzes manchmal unklar; • mäßige Vielfalt an Strukturen; • mäßige Anwendungsbereitschaft.		• ausreichende Sachkenntnis; • Text bezieht sich nicht immer auf Themenstellung; • Format, Länge und Register meistens richtig.		richtiger Gebrauch von Satzverbindungen; • beträchtliche Anzahl an Flüchtigkeitsfehlern und • Fehlern in der Interpunktion.
2	• beschränkter Wortschatz; • beschränkte Anwendungsbereitschaft • häufige Fehler in der Idiomatik, Auswahl und Anwendung des Vokabulars; • häufige Rechtschreibfehler; • Bedeutung der Wortwahl oft unklar.	**2**	• häufige Fehler im Gebrauch der Strukturen und Formen, einfacher Satzbau; • häufige Fehler bei Verbformen, etc.; • Bedeutung des Satzes manchmal unklar; • geringe Vielfalt an Strukturen; • geringe Anwendungsbereitschaft.	**2/ 1**	• Die Aufgabe wurde nicht angemessen gelöst; • geringer Ideenreichtum und geringe Anzahl an schlüssigen Argumenten; • wenig Sachkenntnis; • Text bezieht sich häufig nicht auf Themenstellung; • Format, Länge und Register manchmal unzureichend.	**2/ 1**	• Nicht angemessener Aufbau; • Absätze falsch gesetzt, häufig unlesbar; • wenige und teilweise falsche Satzverbindungen; • viele Flüchtigkeitsfehler; • viele Fehler in der Interpunktion.
1	• sehr beschränkter Wortschatz; • viele Fehler in der Idiomatik, Auswahl und Anwendung des Vokabulars; • viele Rechtschreibfehler; • Bedeutung der Wortwahl meistens unklar; • Übersetzung aus dem Deutschen.	**1**	• viele Fehler im Gebrauch der Strukturen, Formen, und Satzbau; • viele Fehler bei Verbformen, etc.; • Bedeutung meistens unklar, Übersetzung aus dem Deutschen; • keine Vielfalt an Strukturen; • keine Anwendungsbereitschaft.	**1/ 0,5**	• Die Aufgabe wurde nicht eigenständig gelöst; • Mangel an Ideen , irrelevante Argumentation; • keine Sachkenntnis; • Text bezieht sich selten auf Themenstellung; • Format, Länge und Register oft unzureichend.	**1/ 0,5**	• Unlogisch, unzusammenhängend im Aufbau; • keine Strukturierung durch Absätze, unlesbar; • falsch verwendete Satzverbindungen: • sehr viele Flüchtigkeitsfehler: • sehr viele Fehler in der Interpunktion.

* Veto Kategorie: Erhält der Kandidat hier 0 Punkte, ist der ganze Text mit 0 Punkten zu bewerten ("Themenverfehlung)

Anmerkungen zur Bewertungsskala:

a) Die Zahlen **1 – 6 sind Punkte** und entsprechen nicht der umgekehrten Notenskala. Der 3-Punkte-Level entspricht einer „genügenden´´ Leistung in jedem Teilbereich. Wird ein Teilbereich unterschritten, muss in einer anderen Kategorie kompensiert werden.

b) **Veto Kategorie:** Eine Themenverfehlung oder eindeutige Missachtung der Textsorte bewirkt null Punkte für den gesamten Text.

c) Die **Gewichtung** der Kategorien *Erfüllung der Aufgabenstellung* und *Aufbau* ist flexibel:
Erreichbar sind 6 Punkte bzw. 3 Punkte – je nach Stellenwert dieser Kategorien in der jeweiligen Aufgabenstellung.

Speaking Test Specifications– General Specifications

GENERAL TEST DESCRIPTOR	
subtitle	✓ Test of English Speaking Competence
description	✓ The test is designed to evaluate mastery of spoken English at the Matura level (final exam level), i.e. in the 12th school year. ✓ The test is 10 minutes long and takes place as a free group discussion. ✓ The level of the test corresponds to the B2 level of proficiency of the Common European Framework of Reference. For further specifications, see the scoring grid below. ✓ The test covers the English language skill area of speaking.
structure	✓ The examinees take part in open group discussions, two teams of two people discuss with each other. ✓ The examinees demonstrate their ability to produce speech and take part in a conversation.

Speaking Test Specifications – Group discussion

PROMPT ATTRIBUTE / TEXT SPECIFICATION	
topic	✓ The topic will, in general, be familiar to the test takers. ✓ The topic shall not require any special or former knowledge from the test taker concerning vocabulary or the topic itself. *(Do not choose a text, for example about the development of a political matter, which would require the information of prior headlines or articles of the media business.)*
format	✓ The discussion group consists of four examinees that are split into two teams. These two teams receive "role cards" and should consequently argue for or against a topic. In the course of the discussion, the examinees may come to an agreement with the opposing team. ✓ All instructions are on the task sheet and are read out. ✓ One of the examiners starts the discussion of by asking the examinees to introduce their names and their point of view of the topic according to the role card.
length	✓ 10 minutes preparation time ✓ 10-15 minutes discussion time
sequence	✓ Introduction of task ✓ Preparation time in teams: 10 minutes ✓ Start of discussion ✓ Discussion time: 10-15 minutes

RESPONSE ATTRIBUTES/ TEST ITEM SPECIFICATIONS
The examinees should speak with the other members of their discussion group and show the ability to communicate with them in English at an adequate level. The response is correct if the examinee

a) speaks in a coherent and cohesive manner with only a little jumpiness.
b) has a good lexical range.
c) maintains a relatively high degree of grammatical accuracy.
d) can adjust to normal changes of direction and style in conversation.
e) can produce stretches of language with a fairly even tempo.

SCORING

The discussion will be scored according to a scoring grid (attached below). There are two non-native examiners and a native examiner who listen to the discussion and take notes on the scoring sheet during the discussion. Afterwards, the scores of each examinee are discussed between the examiners in order to come to a common final score.

SAMPLE TEST ITEMS

Role card: Smoking Policy in Austria

You can use any information you might have about this topic. Keep in mind:

- Preparation time is 10 minutes – discuss your arguments with your partner.
- You may stick to your role if you want to, but you don't necessarily have to.
- Discussion time is between 10 and 15 minutes.
- You have to let the other test takers, including members of your group, take turns and hear them out.

Find arguments to support your point of view described below:

- Smoking can have severe negative effects on your health
- Even second – hand smoking can cause cancer or other diseases
- Restaurant and bar staff are exposed to the harmful influence of smoking for several hours a day against their will
- Austria is lagging behind in terms of the current smoking policy in the European Union
- …

Criteria list for assessing interactive speaking

	2 points	1 point	0 points
Coherence and Cohesion	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some ‚jumpiness' in a long contribution.	Can connect phrases in a simple way. Can link a series of shorter elements into a connected, linear sequence of points. Can produce simple connected speech.
Score			

Lexical range	Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic exprecions and colloquialisms.	Has a good range of vocabulary for matters connected to most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can sometimes still cause hesitation and circumlocution.	Can understand the main points of clear standard input on familiar matters regularly encountered. Sufficient vocabulary to express him/herself with some hesitation and circumlocution.
Score			

Grammatical accuracy	Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot.	Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.
Score			

Discourse competence	Can adjust what he/she says and the means of expressing it to the situation and the recipient and adopt a level of formality appropriate to the circumstances. Readily available range of interaction strategies.	Can adjust to the changes of direction, style and emphasis normally found in conversation. Can vary formulation of what he/she wants to say. Can intervene appropriately in discussions and initiate and end discourse.	Can initiate, maintain and close simple face-to-face conversation. Can repeat back part of what has been said to confirm mutual understanding. Can enter unprepared into conversation on topics that are familiar.
Score			

Spoken fluency	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for expressions, there are few noticeably long pauses. Can interact with a degree of fluency and spontaneity that makes regular interaction quite possible.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production. Can briefly give reasons and explanations.
Score			

Overall Score: ☐ **according CEFR-level:** ☐ **key:**

10 – 8 points => C1
7 – 4 points => B2
3 – 0 points => B1

15.3 FEEDBACK SHEET

Feedback zum Test

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

1. Wie gut findest du diese Art von Test?

sehr gut	gut	weniger gut	schlecht

2. Wie interessant waren die Texte bzw. Themen?

sehr interessant	eher interessant	eher uninteressant	uninteressant

3. Wie schwer war der Test?

sehr schwer	schwer	leicht	sehr leicht

4. Wie findest du die Länge des Tests?

zu lang	lang	okay	kurz	sehr kurz

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

sehr gut	gut	weniger gut	schlecht

Weitere Kommentare:

15.4 RASCH ANALYSIS

English listening test: first calculation

QUEST: The Interactive Test Analysis System
Current System Settings
all on all (N = 50 L = 48 Probability Level=0.50)

Data File	=	ListeningEnglish.txt
Data Format	=	items 1-48
Log file	=	LOG not on
Page Width	=	120
Page Length	=	65
Screen Width	=	78
Screen Length	=	45
Probability level	=	0.50

Maximum number of cases set at 100000
VALID DATA CODES 0 1 2 3 4 5 6
GROUPS
1 all (50 cases): All cases
SCALES
1 all (48 items): All items

DELETED AND ANCHORED CASES:
No case deletes or anchors

DELETED AND ANCHORED ITEMS:
No item deletes or anchors
RECODES

QUEST: The Interactive Test Analysis System
Case Estimates
all on all (N = 50 L = 48 Probability Level=0.50)

Summary of case Estimates

Mean	0.49
SD	0.35
SD (adjusted)	0.30
Reliability of estimate	0.72

Fit Statistics

Infit Mean Square		Outfit Mean Square	
Mean	1.06	Mean	1.02
SD	0.26	SD	0.31

Infit t		Outfit t	
Mean	0.25	Mean	0.13
SD	1.04	SD	0.76

0 cases with zero scores
0 cases with perfect scores

Item 23: item 23
Infit MNSQ = 1.20
Disc = -.12

Categories	0 [0]	2 [2]	missing
Count	8	42	0
Percent (%)	16.0	84.0	
Pt-Biserial	0.12	0.12	
Mean Ability	0.61	0.47	NA
StDev Ability	0.42	0.33	NA
Step Labels		1	2
Thresholds		0.41	0.41
Error		0.20	0.20

Item 28: item 28
Infit MNSQ = 1.86
Disc = -.03

Categories	1 [0]	2 [1]	3 [2]	4 [3]	5 [4]	6 [5]	missing
Count	9	8	1	4	7	21	0
Percent (%)	18.0	16.0	2.0	8.0	14.0	42.0	
Pt-Biserial	0.14	0.03	0.08	0.26	0.25	0.25	
Mean Ability	0.62	0.47	0.34	0.16	0.28	0.58	NA
StDev Ability	0.27	0.23	0.00	0.31	0.51	0.28	NA
Step Labels		1	2	3	4	5	6
Thresholds			0.13	0.26	0.31	0.40	0.61
Error			0.45	0.42	0.42	0.42	0.41

Item 37: item 37
Infit MNSQ = 1.22
Disc = 0.04

Categories	0 [0]	2 [2]	missing
Count	30	19	1
Percent (%)	61.2	38.8	
Pt-Biserial	-0.02	0.02	
Mean Ability	0.49	0.49	0.59
StDev Ability	0.40	0.25	0.00
Step Labels		1	2
Thresholds		0.73	0.73
Error		0.15	0.15

Item 46: item 46
Infit MNSQ = 1.28
Disc = -.06

Categories	0 [0]	1 [1]	2 [2]	3 [3]	4 [4]	missing
Count	2	10	19	10	6	3
Percent (%)	4.3	21.3	40.4	21.3	12.8	
Pt-Biserial	-0.01	0.18	0.12	0.20	0.21	
Mean Ability	0.41	0.68	0.43	0.34	0.69	0.40
StDev Ability	0.23	0.30	0.34	0.31	0.37	0.11
Step Labels		1	2	3	4	
Thresholds		-1.53	-0.19	0.90	1.67	
Error		0.78	0.55	0.51	0.56	

English listening test: second calculation

QUEST: The Interactive Test Analysis System
Current System Settings
all on all (N = 50 L = 48 Probability Level=0.50)

Data File	=	ListeningEnglish.txt
Data Format	=	items 1-48
Log file	=	LOG not on
Page Width	=	120
Page Length	=	65
Screen Width	=	78
Screen Length	=	45
Probability level	=	0.50

Maximum number of cases set at 100000
VALID DATA CODES 0 1 2 3 4 5 6
GROUPS
1 all (50 cases): All cases
SCALES
1 all (48 items): All items

DELETED AND ANCHORED CASES:
No case deletes or anchors

DELETED AND ANCHORED ITEMS:
No item deletes or anchors
RECODES

QUEST: The Interactive Test Analysis System
Case Estimates
all on all (N = 50 L = 48 Probability Level=0.50)

Summary of case Estimates

Mean	0.53
SD	0.47
SD (adjusted)	0.42
Reliability of estimate	0.79

Fit Statistics

Infit Mean Square		Outfit Mean Square	
Mean	0.99	Mean	1.02
SD	0.21	SD	0.40

Infit t		Outfit t	
Mean	-0.03	Mean	0.16
SD	0.89	SD	0.77

0 cases with zero scores
0 cases with perfect scores

English reading test: first calculation

QUEST: The Interactive Test Analysis System
Current System Settings
all on all (N = 51 L = 46 Probability Level=0.50)

Data File	=	ReadingEnglish.txt
Data Format	=	items 1-46
Log file	=	LOG not on
Page Width	=	120
Page Length	=	65
Screen Width	=	78
Screen Length	=	45
Probability level	=	0.50

Maximum number of cases set at 100000
VALID DATA CODES 0 1 2 3 4 5 6
GROUPS
1 all (51 cases): All cases
SCALES
1 all (46 items): All items

DELETED AND ANCHORED CASES:
No case deletes or anchors

DELETED AND ANCHORED ITEMS:
No item deletes or anchors
RECODES

QUEST: The Interactive Test Analysis System
Case Estimates
all on all (N = 51 L = 46 Probability Level=0.50)

Summary of case Estimates

Mean	0.64
SD	0.37
SD (adjusted)	0.31
Reliability of estimate	0.73

Fit Statistics

Infit Mean Square		Outfit Mean Square	
Mean	1.01	Mean	0.99
SD	0.28	SD	0.41

Infit t		Outfit t	
Mean	0.02	Mean	-0.06
SD	1.28	SD	1.10

0 cases with zero scores
0 cases with perfect scores

Item 9: item 9
Infit MNSQ = 1.22
Disc = 0.04

Categories	0 [0]	1 [1]	2 [2]	missing
Count	32	3	16	0
Percent (%)	62.7	5.9	31.4	
Pt-Biserial	-0.04	0.03	0.02	
Mean Ability	0.64	0.66	0.65	NA
StDev Ability	0.42	0.24	0.23	NA
Step Labels		1	2	
Thresholds		0.94	1.09	
Error		0.47	0.47	

Item 10: item 10
Infit MNSQ = 1.18
Disc = -.04

Categories	0 [0]	1 [1]	2 [2]	missing
Count	37	1	13	0
Percent (%)	72.5	2.0	25.5	
Pt-Biserial	0.06	-0.12	-0.03	
Mean Ability	0.65	0.33	0.63	NA
StDev Ability	0.34	0.00	0.44	NA
Step Labels		1	2	
Thresholds		1.19	1.24	
Error		0.50	0.50	

Item 16: item 16
Infit MNSQ = 1.17
Disc = 0.02

Categories	0 [0]	1 [1]	2 [2]	missing
Count	37	1	13	0
Percent (%)	72.5	2.0	25.5	
Pt-Biserial	-0.02	0.11	-0.01	
Mean Ability	0.63	0.90	0.65	NA
StDev Ability	0.37	0.00	0.35	NA
Step Labels		1	2	
Thresholds		1.19	1.24	
Error		0.50	0.50	

Item 25: item 25
Infit MNSQ = 1.26
Disc = -.06

Categories	0 [0]	1 [1]	2 [2]	missing
Count	27	2	22	0
Percent (%)	52.9	3.9	43.1	
Pt-Biserial	0.08	-0.06	-0.05	
Mean Ability	0.67	0.53	0.62	NA
StDev Ability	0.35	0.37	0.38	NA
Step Labels		1	2	
Thresholds		0.70	0.80	
Error		0.45	0.45	

Item 39: item 39

Infit MNSQ = 1.23
Disc = 0.20

Categories	0 [0]	1 [1]	2 [2]	3 [3]	4 [4]	missing
Count	6	1	4	13	27	0
Percent (%)	11.8	2.0	7.8	25.5	52.9	
Pt-Biserial	-0.21	-0.17	0.03	0.13	0.05	
Mean Ability	0.44	0.23	0.69	0.72	0.65	NA
StDev Ability	0.48	0.00	0.41	0.26	0.35	NA
Step Labels		1	2	3	4	
Thresholds		-0.14	-0.07	0.09	0.55	
Error		0.54	0.54	0.53	0.47	

English reading test: second calculation

QUEST: The Interactive Test Analysis System
Current System Settings
all on all (N = 51 L = 46 Probability Level=0.50)

Data File	=	ReadingEnglish.txt
Data Format	=	items 1-46
Log file	=	LOG not on
Page Width	=	120
Page Length	=	65
Screen Width	=	78
Screen Length	=	45
Probability level	=	0.50

Maximum number of cases set at 100000
VALID DATA CODES 0 1 2 3 4 5 6
GROUPS
1 all (51 cases): All cases
SCALES
1 all (46 items): All items

DELETED AND ANCHORED CASES:
No case deletes or anchors

DELETED AND ANCHORED ITEMS:
No item deletes or anchors
RECODES

QUEST: The Interactive Test Analysis System
Case Estimates
all on all (N = 51 L = 46 Probability Level=0.50)

Summary of case Estimates

Mean	0.77
SD	0.47
SD (adjusted)	0.42
Reliability of estimate	0.79

Fit Statistics

Infit Mean Square		Outfit Mean Square	
Mean	1.01	Mean	0.98
SD	0.31	SD	0.45

Infit t		Outfit t	
Mean	0.02	Mean	-0.04
SD	1.27	SD	1.07

0 cases with zero scores
0 cases with perfect scores

Derntl Victoria

geboren am	20.08.1985 in Linz
Adresse:	Lehenbrunn 19, 4320 Perg
E-Mail:	v_derntl@gmx.at
Familienstand:	ledig
Staatsangehörigkeit:	österreichisch
Eltern:	Heinz und Elfriede Derntl
Geschwister:	Christian und Stephanie Derntl

AUSBILDUNG

Universität Wien: Lehramtsstudium Englisch und Deutsch

September 2003 bis Juli 2009

Erasmus Auslandsaufenthalt in Manchester (UK)

September 2006 – März 2007 (Manchester Metropolitan University)

BORG Perg

1999 – 2003: Matura mit Auszeichnung

Hauptschule 1 Perg

1995 – 1999

Volksschule Pergkirchen

1991 – 1995

ERFAHRUNG

Schulen in Wien und Oberösterreich

Unterrichtserfahrung in den Fächern Deutsch und Englisch

‚LernQuadrat'

Beschäftigung als Nachhilfelehrerin für Deutsch und Englisch

‚Schülerhilfe'

Beschäftigung als Nachhilfelehrerin für Deutsch und Englisch

Institut für Suchtprävention

Ausbildung zur Peer-Trainerin

Tätigkeit in Fahrschulgruppen im Rahmen der "Peer Drive Clean"-Initiative