UNICHECK

Ім'я користувача:
Моца Андрій Андрійович

ID перевірки:
1007636275

Дата перевірки:
30.04.2021 17:16:49 EEST

Тип перевірки:
Doc vs Internet

Дата звіту:
30.04.2021 17:24:12 EEST

ID користувача:
100006701

Назва документа: **ГАЛ ВІВІЄН ГРЕТА**

Кількість сторінок: **142** Кількість слів: **39940** Кількість символів: **272523** Розмір файлу: **1.63 MB** ID файлу: **1007746222**

4.33%
Схожість

Найбільша схожість: **0.76%** з Інтернет-джерелом (http://mitralanguagetesting.blogspot.com/2010/04/testing-assessing-te

4.33% Джерела з Інтернету	437	Сторінка 144

Пошук збігів з Бібліотекою не проводився

18.3% Цитат

Цитати	256	Сторінка 145
Посилання	1	Сторінка 159

0.02%
Вилучень

Деякі джерела вилучено автоматично (фільтри вилучення: кількість знайдених слів є меншою за **8 слів** та **0%**)

0.02% Вилучення з Інтернету	38	Сторінка 160

Немає вилучених бібліотечних джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	4

ЗАТВЕРДЖЕНО
Вченою радою ЗУІ
Протокол № „3” від „27” квітня 2021 р.
Ф-КДМ-2

Закарпатський угорський інститут ім. Ференца Ракоці ІІ

Кафедра філології

Реєстраційний №_____

Кваліфікаційна робота

ДОСЛІДЖЕННЯ МЕТОДІВ ОЦІНЮВАННЯ НА УРОЦІ АНГЛІЙСЬКОЇ МОВИ У ЗАКАРПАТСЬКИХ УГОРСЬКИХ ШКОЛАХ

ГАЛ ВІВІЄН ГРЕТА
Студент(ка) _2_-го курсу

Освітня програма «Філологія» (мова і література англійська)
Ступінь вищої освіти: магістр

Тема затверджена Вченою радою ЗУІ

Протокол № 7 /27 жовтня 2020 року

Науковий керівник: **Густі Ілона Іштванівна**

д-р філософії, доцент

Завідувач кафедрою: **Берегсасі Аніко Ференцівна**
д-р габілітований, професор

Робота захищена на оцінку _____, «___» _____ 2021_ року

Протокол № _____ / 2021_

ЗАТВЕРДЖЕНО
Вченою радою ЗУІ
Протокол № „3” від „27” квітня 2021 р.
Ф-КДМ-2

Закарпатський угорський інститут ім. Ференца Ракоці ІІ

Кафедра філології

Кваліфікаційна робота

ДОСЛІДЖЕННЯ МЕТОДІВ ОЦІНЮВАННЯ НА УРОЦІ АНГЛІЙСЬКОЇ МОВИ У ЗАКАРПАТСЬКИХ УГОРСЬКИХ ШКОЛАХ

Ступінь вищої освіти: магістр

Виконала: студентка _2_-го курсу

Гал Вівієн Грета
Освітня програма
«Філологія» (мова і література англійська)

Науковий керівник: **Густі Ілона Іштванівна**

д-р філософії, доцент

Рецензент: **Теличко Н.В.**

д-р пед.наук

Берегове
2021

ЗАТВЕРДЖЕНО
Вченою радою ЗУІ
Протокол № „3” від „27” квітня 2021 р.
Ф-КДМ-2

Ferenc Rákóczi II Transcarpathian Hungarian College of Higher Education
Department of Philology

INVESTIGATING ASSESSMENT PRACTICES IN THE ENGLISH CLASSROOM IN TRANSCARPATHIAN HUNGARIAN SCHOOLS

Master's Thesis

Presented by: Vivien Gréta Gál

a 2nd year student

Professional Education program:
Philology (English language and literature)

Thesis supervisor: Ilona Huszti
PhD, Associate Professor

Second reader: Nataliia Telychko
DSc

Beregszász – 2021

ЗМІСТ

CONTENTS

INTRODUCTION

Knowledge of a foreign language is perhaps of unprecedented importance today and this is especially true for English. In addition to language teaching and learning, language assessment plays an increasingly crucial role. This is not surprising, as language has acquired important social, political and economic roles and more and more businesses are making high-stake decisions for employment purposes based on test results.

Assessment is how we identify the needs of our students, keep track of their progress and evaluate our own performance as teachers and other stakeholders. However, this raises an important question: how do we know we are doing it the right way?

Traditionally, the most common way to assess language knowledge and proficiency has been the use of tests. Although many alternative forms of assessment are gaining in popularity, most teachers still use tests to measure the achievement of their students. A lot of teachers perform wonderfully in the classroom, but when it comes to assessment they are often left alone with their own judgement. This raises another question: if language assessment plays such an important part, why is there a gap in the assessment knowledge of teachers? The factors that result in effective language learning, as well as its assessing techniques, are constantly evolving, making it extremely important for teacher preparation programs to better prepare preservice and in-service educators with the assessment literacy knowledge, skills, and expertise to effectively make the best use of robust data toward the ultimate goal of education.

A considerable amount of academic literature has been published on assessment in EFL classroom. A major contribution to the field was made by Bailey (1998), Law and Eckes (1995), Winking (1997), Reeves (2009), Brown (2004). A great number of researchers carried out studies to examine the use of different assessment methods among EFL teachers (Huseyin, 2014; Rezaee, 2013) and the impact of traditional and alternative assessment methods on students

performance (Kalra, Sundarajun and Komintaracat, 2017; Letina, 2014). The effect of self- and peer-assessment has been also broadly studied (Azarnoosh, 2013; Crosthwaite, Bailey and Meeker, 2015; Szénásiné, 2017). There has been also a number of research conducted on the effect of teaching to the test on learners' performance by researchers like Neil (2003), Newman, Bryk and Nagaoka (2001).

However, little is known about the views and beliefs of teachers about assessment and testing in Transcarpathia and more importantly, about how they use different assessment methods in the English lessons.

Accordingly, the object of the master thesis is assessment practices in Transcarpathian schools with Hungarian language of instruction.

The subject of the master thesis is the use of traditional and alternative assessment and feedback giving methods of English teachers and the applicaiton of "teaching to the test" in the English lessons in Transcarpathian schools with Hungarian language of instruction.

The aim of the master thesis is to study the use of traditional and alternative assessment methods among secondary school teachers in Transcarpathian Hungarian schools and find similarities and differences in the assessment practices of lower and upper secondary school teachers. A further aim is to highlight the areas in language testing and assessment that may need improvement and to put an emphasis on the importance of meaningful feedback. It also aims to explore the attitudes of teachers towards "teaching to the test" and to examine the way in which exam preparation for the external independent evaluation takes place in Form 11.

The task of the thesis are as follows:

1. analysis of academic literature;
2. developing a theoretical and conceptual framework of language assessment;
3. presenting the stages of test contruction;
4. studying English language teachers' beliefs and practices regarding language assessment;

5. analyzing the results and drawing conclusions;

A mixed methods study is carried out, the methodology used being both qualitative and quantitative. It involves an empirical investigation, using an online questionnaire as data gathering instrument.

The novelty of the research lies in the fact that a similar survey has not yet been conducted to assess the assessment practices of teachers in the area.

The practical value of the study consists in providing a useful insight into the assessment views and practices of teachers teaching in schools with Hungarian language of instruction in Transcarpathia, which can serve as a source for further research in the field of foreign language teacher training and English language teaching in the area. The results of the study may lead to the launch of more comprehensive programs and courses to develop teachers skills in assessment and to the improvement of the quality of language teaching.

The thesis is made up of an introduction, 3 parts, conclusions, references and resume. Part 1 provides a theoretical and conceptual framework for the study by reviewing literature on language assessment, involing the changes in the concept and practice of assessment and testing over the time and the characteristics and use of different assessment methods and instruments. The focus of part 2 is on the stages of test instruction, from test specification through pretesting to validation and post-test reports. Part 3 presents the procedure, results and discussion of an empirical investigation and its implications.

PART 1

Brief overview of language assessment

1.1 Clarifying the meaning of test and assessment

Test and assessment are two popular and sometimes misunderstood terms in educational practice. One might be tempted to think of testing and assessing as synonymous terms, but they are not. They appear to be similar in meaning and are often incorrectly used interchangeably. Therefore, it is important to distinguish between the terms, as well as the concept of them. What is a test?

According to Brown (2004), "test is a method of measuring a person's ability, knowledge, or performance in a given domain" (p.3). Tests are prepared procedures that take place at identifiable times in a curriculum when learners master all their faculties to offer peak performance, understanding that their answers will be assessed and evaluated. Assessment, on the other hand, refers to an ongoing process. When a student responds to a question, or tries out a new word or structure, the teachers assess the student's perfromance subconsciously. Hence, tests can be defined as a subset of assessment because they are only one among the many procedures that teachers can use to assess students (Brown, 2004). Observing, recording information, testing, scoring and interpreting results are all part of the assessment process. As Angelo and Cross (1993) state, teachers use assessment to obtain feedback on "what, how much and how well their student are learning" and to make decisions about how to "refocus their teaching to help students make their learning more efficient and more effective" (p.3).

1.2 The concept of assessment literacy

Language testing has become more and more important in recent years. Language has gained an influential place in education, politics and business. Educational boards are planning the implementation of standardised language tests and businessess make high-stakes decisions for employment purposes based on test results. That's why it is surprising, that language testing often coincides with a

serious lack of knowledge of teachers and teacher educators. Given the significance and growing relevance of language testing, it should play a major role in teacher education programmes. Due to the lack of appropriate educational background, teachers and other testers are left alone with their own judgements and testing practices they believe to be proper. This inconsistency in current demands in language testing and assessment, and the lack of expertise create a growing need for teacher education programmes to prepare teachers to their role as testers (Berger, 2012).

There is a recently coined term describing what teachers need to know about assessment and what basic testing knowledge teachers and other stakeholders should possess, it is called "assessment literacy". This term was made by several of writers like Boyles (2005), Malone (2008), Stiggins (1991), Stoynoff and Chapelle (2005). It was once thought to be the ability to select, design and evaluate tests and assessment procedures, as well as to score and grade them on the basis of theoratical knowledge. Taking into consideration the implications of assessment for teaching, more recent approaches embrace a broader understanding of the concept. Stiggins (1991) defines assessment literacy as the ability to differentiate between sound and unsound assessment. Assessment literates ask two main questions:

- What does an assessment tell students about the achievement outcomes we value?
- What is likely to be the effect of this assessment on students? (Stiggins, 1991)

It is considered crucial to know and understand the main concepts of sound assessment and to be able to translate them into quality information about students' achievements and into effective instruction. Boyles (2005) also explains assessment literacy as the understanding of the practices of testing and assessment. Language teachers need the necessary tools for analysing and reflecting upon test data in order to make informed decisions about instructional practise and programme design (Armstrong, 1994). Both include the notion of assessment

literacy to include not only technical knowledge about how to select and create appropriate assessment instrument for specific purposes, but also the ability to analyse empirical data to improve instruction. Thus, being literate in assessment implies a move away from a passive interpretation towards and active application of data that will affect and likely improve teaching.

1.3 Conceptual transformations in the past and present of language assessment

The notion of language-assessment literacy has changed a lot in its nature and underlying philosophies over the years. Spolsky's division of language testing into three eras could provide a useful historical perspective of these changes.

Spolsky's devision of language testing (1978):

- pre-scientific or traditional,
- psychometric-structuralist or modern,
- psyholinguistic or postmodern.

It should be noted by Spolsky's devision, that the trends he stated follow in order, but overlap in time and approach. The third picks up elements from the first and the second and third co-exist and compete. Each period has its own set of values and beliefs about the knowledge and skills that are dominant in the educationl assessment traditions of the time.

1.3.1 Pre-scientific or traditional period

Prior to the 1960s, the pre-scientific or traditional period may be characterized by a lack of concern for objectivity and reliability. The term "pre-scientific" indicates that science was not yet applied in this field of educational measurement. During this period, there was a strong reliance on the judgement of experienced teachers. According to this viewpoint "if a person knows how to teach, it is to be assumed that he can judge the proficiency of his students" (Spolsky, 1978, p. 5-6). Since there was no formal instruction for teachers on how to design assessments and score them, the required expertise can be seen as intuitive in the sense that the expert status of teachers was sufficient to legitimize their decision in the testing

and rating process (Berger, 2012). There were no oral examinations and language testing was assumed to be a matter of open-ended written examinations. Written examinations would tipically consist of passages for translation into or from the foreign language, free composition in it, and selected items of grammatical, textual or cultural interest (Spolsky, 1978).

1.3.2 Psychometric-structuralist or modern period

In contrast, the next period from the 1960s onward, sees the invasion of the field by experts. The psychometric-structuralist approach assumes an independently existing reality that can be discovered and measured using objective, scientific methods (Berger, 2012). This trend may be characterized by the interaction and dispute of two sets of experts, agreeing with each other primarly in their assumption that testing can be made precise, objective, reliable, and empirical.

The first of these groups of experts were the testers with their main concern to provide "objective" measures (Sharon & William, 2008). Validity and reliability became fundamental requirements during this period. As Ingram (1968) stated in his work "firstly the shape of all tests, whether predictive or non-predictive, language or non-language, is primarly determined by the need to tests for reliability and validity. That is why, for instance, the multiple choice technique of answering is so common" (p. 74).

Their first aim was to prove the unreliability of traditional examinations and to show how unreliable subjective scores can be. After that, they started to develop more reliable measures and working on test items that could be more amenable to control. As a consequence, open-ended questions gave way to discrete-point items testing structural aspects of the language such as grammar, vocabulary, spelling and pronunciation (Berger, 2012). The fruit of their work was the development of short item, multiple choice, objective tests. It had two results. Firstly, these objective tests required written response, and so were restricted to reading and listening. Secondly, the chosen items didn’t show advanced ideas about language teaching (Spolsky, 1978). The pedagogical emphasis of this measurement pradigm

is characterized as "assessment of learning" (Gipps, 1994), in which learning outcomes are assessed summatively at the end of a learning period, and the set of beliefs underlying it constitutes what has been called a "testing culture", in which formal tests dominated over other more informal and process-oriented forms of assessment (Wolf, 1991). Teaching and testing are seen as two distinct sets of activities in the time of "testing culture".

This kind of language testing left a number of deficiencies. According to Lado (1951) there have a great lag existed in measurement in English as a foreign language. Caroll (1954) confirmed Lado's judgement and added that "a great lag exists in all foreign language measurement".

Another major impulse in the scientific period was when a group of experts added notions from the science of language to those of the science of educational measurement. John B. Carroll is an academic who has spent most of his career straddling the two fields. Carroll's impact on the development of language tests has been important, mainly because of his unique role as being both a linguist and a psychologist. Carroll was the one who brought Lado's work to light, which marked the start of the second stage of the scientific period. The construction of English achievement tests for Latin-American students of the language was the subject of Lado's (1951) doctoral dissertation.

A few years later, he wrote the famous book *Language Testing* (1961) and according to the website of the International Language Testing Association (ILTA), Lado was considered the "founder of modern language testing research and development". Works in language testing since him are widely based on his work. His approach is modest in that he acknowledges the tester's right to establish types of tests and methods for judging validity and realiability, while still insisting on the linguist's responsibility to decide what to test (Spolsky, 1978).

1.3.3 Psycholinguistic or postmodern period

A new trend arouse in the late 1970s and early 1980s, in which the emphasis was not on the knowledge of structural elements of the language, but rather on the

appropriate use of language specific to the context and audience. This approach was called the structural-psychometric trend. As Spolsky (1978) suggests, this new trend has not entirely overcome the doubts of the traditionalists. They still claim that less specific measures are still of great value. Therefore, they played an important role in developing more reliable methods of assessing the more subjective kinds of performance.

Firstly, they were concerned with the judgement of written proficiency. Some linguists have showed that objective writing tests - usually involving multiple-choice items - correlate well with other measures. Also, some other scholars have pointed out kinds of techniques of shorter essays and scoring guides that add reliability to subjective marking. Traditional tests were proved to be able to improve.

The second effort focused on the assessment of oral proficiency, which is a skill that objective test do not adequately cover. Despite its significance, speech production remains the hardest to assess (Spolsky, 1978). As Perren (1968) states "the most difficult problems arise when trying to construct tests of ability to speak a language… Suffice is to say that although the ideal of a test based on free conversation is very attractive, the problems of sampling and reliable scoring are almost insoluble, unless a good deal of time and many standardized expret testers are avilable" (p.115). These tests can be made reliable and objective, but it costs a lot to do so. Thus, the problem is a practical issue, the question of affordability. The supporters of discrete item tests have many strong arguments on their side, but still there have been more and more attacks on their principles. These attacks can be associated with two trends in contemporary linguistics – "language competence trend" and "communicative competence trend".

The "language comptenece trend" is based on a belief in overall language proficiency, and a belief that knowledge of a language is more than just the sum of a set of discrete parts. This trend is connceted to psycholingustics. The "communicative competence trend" accepts the belief in integreative testing, but insists on the need to add a strong functional dimension to language testing

(Spolsky, 1978). Caroll (1961) stressed the need for an integrative approach, where one pays attention not to specific structural or lexical items, but to the total communicative effect of an utterance. This trend is connected with the views of modern sociolinguists.

The trend toward communicative forms of assessment is continuing today and new models of assessment are emerging in many countries. There is a strong emphasis on situationally and interactionally authentic performance-based assessment to learn what learners can do with language in non-test situations (Bachman, 1991). As well as, there is an increased emphasis on methods of collecting information from learners, such as portfolios, students self- and peer-assessment and authentic assessment of real-life tasks.

A wider notion of assessment has gained ground. In this new understanding, assessment is viewed as a means to promote learning and the concept of "assessment for learning" (Chappuis & Stiggins, 2002; Gipps, 1994) emerged. In this sense, information gained from broader forms of assessment is used in addition to testing, not to just monitor student outcomes and certify the end products of learning, but rather to improve instruction. Parallel to a "learning culture" (Shepard, 1998), an "assessment culture" (Inbar-Lourie, 2008) has appeared.

Another step in this development is "assessment as learning", which directs the focus of attention away from the teacher towards the learner. It focuses on the role of the learner as the crucial connector between assessment and learning. Assessment is used to develop and encourage metacognition for students so that they can use the knowledge gained from the assessments for new learning. In this understanding of "assessment as learning", students are enabled and encouraged to use feedback from assessment to monitor their learning autonomously and to reflect and analyze critically their own progress (Earl, 2003). As a result of these developments, the gap between assessment and teaching has narrowed. Assessment is regarded as an integral part of both teaching and learning.

It is obvious, that the notion of assessment literacy is a complex and dynamic one that has been transformed in response to social, political and epistemological changes to reflect current values, views and attitudes in language teaching and assessment. Although language-assessment literacy in the early days was not yet an epistemological category, in the psychometric period it coincided with expertise in scientific measurement. The most recent understanding of assessment literacy draws a significance correlation between language assessment, learning and teaching. The shift from a "testing culture " towards an "assessment culture" confirms that learning and assessment are interwined (Berger, 2012).

1.4 Approaches in language testing

The shifting sand of teaching methodology has affected language-testing trends. Testing focused on specific language elements such as phonological, grammatical and lexical contrasts between two languages in the era of behaviorism and contrastive analysis, in the 1950s. In the 1970s and 1980s, communicative theories of language brought with them a more integrative approach to testing in which experts claimed that the entire communicative event was considerably greater than the sum of its linguistic components (Clark, 1983). Today, test designers are still challenged in their quest for more authentic, valid instruments that stimulate real-world interaction (Brown, 2004).

1.4.1 Discrete-point and Integrative testing methods

In the 1970s and early 1980s, there were two major approaches to language testing that were debated and still prevail today: the choice between discrete-point· and integrative testing methods (Oller, 1979). Discrete-point tests are based on the assumption that language can be broken down into its constituent parts and that those parts can be successfully tested. These components include listening, speaking, reading, and writing skills, as well as various units of language.

New approaches were needed as the profession emerged into an era that prioritized communication, authenticity and context. According to Oller (1979),

language competence is a unified set of interacting abilities that cannot be tested separately. He believed that communicative competence is so global and requires such integration that it cannot be measured in additive tests of grammar, reading, vocabulary and other discrete points of language. Others, like Cziko (1982) and Savignon (1982), soon followed in their support for integrative testing.

The integrative approach involves the testing of a language in context and concentrating mainly on meaning and the communicative effect of a discourse. There are two types of tests that are typically regarded as being integrative tests: cloze tests and dictations.

1.4.1.1 Cloze test

A cloze test is a reading passage in which usually every sixth or seventh word has been removed and the test-taker is required to supply words that fit into those blanks (Brown, 2004). Cloze testing is based on Gestalt Psychology and the Information Processing Theory of "Closure" which refers to people's ability "to complete a pattern once they have grasped its overall significance" (Weir, 1998, p. 46).

Cloze tests assess the reader's ability to make the most acceptable substitutions from all the contextual clues available (Heaton, 1988). The ability to fill in the blanks with suitable words necessiates a variety of abilities that are at the core of language competence: knowledge of vocabulary, grammatical structure, discourse structure, reading skills and strategies, and an internalized "expectancy" grammar. An internalized "expextancy grammar" enables one to predict an item that will come next in a sequence (Brown, 2004). Oller (1979) claimed that cloze test results are good measures of overall proficiency.

1.4.1.2 Dictation

Dictation is a traditional language-teaching method that has evolved into a testing technique. During a dictation learners listen to a passage of 100 to 150 words read aloud by a teacher or recorded on an audiotape and write what they hear, using proper spelling. There are three stages of the listening: an oral reading without

pauses; an oral reading with long pauses between each phrase allowing the learner time to write down what is heard; and a third reading at normal speed to give test-takers a chance to check what they wrote. Supporters argue that dictation is an integrative test, since it integrates the grammatical and discourse skills required for other modes of language performance. Success on a dictation requires careful listening, reproduction in writing of what is heard, efficient short-term memory, and, to an extent, some expectancy rules to help short-term memory. Since large-scale dictation administration is impractical from a scoring perspective, most dictation takes place in the classroom.

1.4.1.3 Unitary trait hypothesis

The unitary trait hypothesis became the centre of the arguments of the proponents of integrative test methods. Unitary trait hypothesis proposed an "indivisible" view of language proficiency, implying that vocabulary, grammar, phonology, the "four skills", and other discrete points of language could not be separated in language performance. "The unitary trait hypothesis contended that there is a general factor of language proficiency such that all the discrete points do not add up to that whole" (Brown, 2004, p. 9). However, there were some experts who strongly argued against the unitary trait position. One of them was Farhady (1982), who discovered widely varying differences in the performance on an ESL proficiency test in a study of students in Brazil and the Philippines, depending on such factors as the subjects' native country, major field of study, and graduate versus undergraduate status. There were other studies that strongly challenged the unitary trait hypothesis and backed up Farhady's standpoint. Finally, in the face of mounting evidence, Oller (1983) acknowledged that "the unitary trait hypohthesis was wrong" (p. 352).

1.4.2 Communicative language testing

By the mid-1980s, the language-testing field had begun to focus on developing communicative language-testing tasks. The need for a correspondence between language test performance and language use was stated by Bachman and Palmer as

one of the "fundamental" principles of language testing. "In order for a particular language test to be useful for its intended purposes, test performance must correspond in demonstrable ways to language use in non-test situations" (Bachman & Palmer, 1996, p. 9). The problem was that the tasks appeared to be artificial, contrived, and unlikely to represent real-life lanaguage use. As a result, a quest for authenticity began with test designers focusing more on communicative performance.

Communicative tests are mainly concerned with the use of language in communication. It introduces the concept of qualitative models of assessment as opposed to quantitative models (Parviz, 2002).

Brown (2005) identifies five requirements that make up what is to be called a communicative test. The requirements in question are:

1. "meaningful communication, i.e. the test needs to be based on communication that is meaningful to students, that is, it should meet their personal needs. Making use of authentic situations can increase the likelihood that meaningful communication will be achieved.
2. authentic situation, i.e. communicative test offers students the opportunity to encounter and use the target language receptively and productively in real-life situations to show how strong their language ability is.
3. unpredictable language input, i.e. the fact that in reality it is usually impossible to predict what speakers will say; this natural way of communication should be replicated in a communicative test.
4. creative language output, i.e. the fact that in reality language input is largely dependent on language input to prepare for one's reply.
5. integrated language skills, i.e. a communicative test will elicit the learners' use of language skills integratively, as is the case in real life communication." (p. 21)

1.4.3 Performance based assessment

In recent years, performance-based assessment has made a comeback in education. Test designers are now tackling this new and more student-centered agenda in language courses and programs all around the world (Alderson, 2001, 2002).

The definition of performance-based assessment varries greately depending on author, discipline, publication and intended audience (Palm, 2008). A simple definition would be that a performance-based assessment assesses the students' ability to apply the skills and knowledge learned during a unit or units of study. The most important feature of performance-based assessment is that it should accurately measure one or more specific course standards. This kind of assessment is also complex, authentic, process/product oriented, open-ended and time-bound (Hilliard, 2015).

Performance-based assessment of language tipically involves oral production, written production, open-ended responses, integrated performance, group performance, and other interactive tasks, rather than merely providing paper-and-pencil tests of a number of separate items. In this way, higher content validity is achieved because learners are measured in the process of performing the targeted linguistic acts. The presence of interactive tasks is a characteristic of almost all performance-based assessments. In such cases, learners are assessed while performing the desired action, and test-takers are measured in the act of speaking, requestioning, responding, or in combining listening and speaking, and in integrating reading and writing. Paper and pencil tests do not elicit this kind of communicative performance (Brown, 2004).

An oral interview is one good example of an interactive language assessment procedure. During an oral interview the test-taker is required to listen accurately to someone else and to respond appropriately. The oral interview has three desirable traits: authenticity, communicativity and flexibility. Underhill (1987) defines an authentic task as "one which resemble very closely something which we actually

do in everyday life" (p. 8). Madsen (1983) says that an oral interview can be one of the most communicative of all language examinations.

1.5 Current trends in classroom testing

Current trends in classroom testing are affected by new theories of intelligence, the advent of alternative assessment and the increasing popularitiy of computer-based testing.

1.5.1 New perspectives on intelligence

Intelligence was once solely defined as the ability to perform linguistic and logical-mathematical problem solving. This "IQ" (intelligence quotient) concept of intelligence has pervaded the Western world and its way of testing for nearly a century. "Since "smartness" in general is measured by timed, discrete-point tests consisting of a hierarchy of separate items, why shouldn't every field of study be so measured?" (Brown, 2004, p. 11).

However, research on intelligence by psychologists like Howard Gardner, Robert Sternberg, and Daniel Goleman has started to transform the psychometric world (Brown, 2004). Howard Gardner (1983, 1999) is well known for rejecting a unitary explanation of intelligence and developing his theory of multiple intelligences (MI). He extended the traditional view of intelligence to seven different components. He accepted the traditional conceptualizations of linguistic intelligence and logical-mathematical intelligence, which are the foundations of standardized IQ tests, but he included five other "frames of mind". These are the followings:

- spatial intelligence (the ability to find your way around an environment, to form mental images of reality)
- musical intelligence (the ability to perceive and create pitch and rhythmic patterns)
- bodily-kinesthetic intelligence (fine motor movement, athletic prowess)
- interpersonal intelligence (the ability to understand others and how they feel, and to interact effectively with them)

- intrapersonal intelligence (the ability to understand oneself and to develop a sense of self-identity) (Gardner, 1983, 1999)

There has also been a great deal of effort put into understanding genius, giftedness and creativity in relation to normal intelligence. Robert Sternberg (1988, 1997) broke new ground in intelligence research in recognizing creative thinking and manipulative strategies as part of intelligence. Sternberg (2003, 2005) proposed a model of intelligence that involves synthesizing wisdom, intelligence and creativity (WICS). He deplored the fact that Western society is organized around a closed system that defines intelligence very narrowly. Not everyone who is believed to be "smart" is capable of fast, reactive thinking. There is still extensive and fascinating research in intelligence and its relation with creativity.

Daniel Goleman's (1995) introduction of the concept of EQ (emotional quotient) made us to emphasize the importance of emotions in our cognitive processing. He (Goleman, 1995) describes emotional intelligence as the ability to identify, assess, and control one's own emotions, the emotions of others, and the emotions of groups. According to Golman (1998), emotional competencies are learned skills, and not natural abilities. These skills need to be worked on and improved to achieve outstanding performance. Goleman believes that individuals are born with a general emotional intelligence that determines their ability to develop emotional competencies.

These modern conceptualizations of intelligence have not been widely accepted by the academic community. For example, the research literature has criticized Goleman's EI model as being merely "pop psychology" (Mayer, 2008). However, EI is still considered by many to be a valuable framework for businesses in particular.

1.5.2 Traditional and Alternative assessment

Recently, there has been a movement from traditional assassment toward alternative assessments. It is difficult to distinguish between them, since many forms of assessment fall in between the two, and some combine the best of both.

Table 1 highlights differences between the two approaches (Armstrong, 1994; Bailey, 1998).

Table 1

Traditional and alternative assessment (adapted from Brown, 2004, p. 13)

Traditional Assessment	Alternative Assessment
One-shot, standardized exams	Continuous long-term assessment
Timed, multiple-choice format	Untimed, free-response format
Decontextualized test items	Contextualized communicative tasks
Scores suffice for feedback	Individualized feedback and washback
Norm-referenced scores	Criterion-referenced scores
Focus on the "right" answer	Open-ended, creative answers
Summative	Formative
Oriented to product	Oriented to process
Non-interactive performance	Interactive performance
Fosters exstrinsic motivation	Fosters intrinsic motivation

When people became more aware of the impact of testing on curriculum, alternative assessment began to be used as a means for educational reform (Dietel, Herman & Knuth, 1991). Reeves (2000) remarked that "traditional assessment, which is generally called testing, is challenged by alternative assessment approaches" (p. 103).

According to Bailey (1998), traditional assessments are indirect, inauthentic and standardized and for that reason, they are one-shot, speed-based, and norm-referenced. Bailey also mentions that this type of assessment does not provide learners with any feedback. Law and Eckes (1995) point out that most standardized tests only measure the lower-order thinking skills of the learner. Alternative assessments, on the other hand, assess higher-order thinking skills, where students have the opportunity to show what they have learned. More authentic alternative assessment tools, such as portfolios, projects, journals, oral presentations, diaries

and writing folders, let learners express their knowledge on the material in their own ways.

Winking (1997) discusses the several advantages of alternative assessment:

- firstly, they tend to simulate real-life contexts;
- secondly, they encourage collaborative working;
- finally, alternative assessments assist instructors to have a better understanding of student learning.

1.5.2.1 Self- and peer-assessment

Some instances of alternative assessments include self- and peer-assessment. Self- and peer-assessment, in which learners assess each other and themselves, has the potentiality to encourage the learners to take greater responsibility for their own learning by getting engaged with assessment criteria and reflection of their own performance and that of their peers (Fathi, Mohammad & Sedighravesh, 2017).

Theoretically, self-assessment is justified by a number of well-established concepts of second language learning. The principle of autonomy is one of the cornerstones of successful language acquisition. Developing intrinsic motivation that comes from a self-propelled desire to succeed is at the top of the list of mastering any set of skills. Peer-assessment is based on similar concepts, the most apparent of which is cooperative learning (Brown, 2004).

According to Henner-Stanchina and Holec (1985), self-assessment is an assessment technique in which learners create and undergo the evaluation procedure at the same time, assessing their performance in relation to themselves against their own personal criteria, in accordance with their own goals. Topping (1998) believes that peer-assessment is an arrangement in which individuals evaluate the success of their peers of similar status, with regard to the amount, level, worth and quality of their achievement.

1.5.3 Computer-based testing

In recent years a new type of assessment has emerged in which the test-taker performs responses on a computer. In simple terms, computer-based exams or tests

are those administered through the computer instead of paper and pencil format. Some computer-based tests, which are also known as "computer-assisted" or "web-based" tests, are small-scale tests available on websites, but there are other standardized, large-scale tests in which thousands of test-takers are involved. Almost all computer-based test items· have fixed, closed-ended responses; however, some tests, like the Test of English as a Foreign Language (TOEFL) have a written essay section that must be scored by humans as opposed to automatic, electronic scoring (Brown, 2004).

1.5.3.1 Computer-adaptive test

A computer-adaptive test (CAT) is a specific type of computer-based tests, which adjusts to the ability level of the examinee. For this reason, it has also been called tailored testing. In other words, as the National Council on Measurement in Education's Glossary of Important Assessment and Measurement Terms (2017) defines it, "it is a form of computer-administered test in which the next item or set of items selected to be administered depends on the correctness of the test taker's responses to the most recent items administered".

When test takers answer a question, the computer scores it and uses that information, along with the responses to previous questions, to select which question will be presented next. As long as examinees respond correctly the computer typically chooses questions of greater or equal difficulty. Incorrect responses, on the other hand, typically bring questions of lesser or similar difficulty. The computer is programmed to fulfill the test specification as it continuously adjusts to find questions of appropriate difficulty for test-takers at all performance levels (Brown, 2004).

According to Brown (2004), computer-based testing offers the following advantages:

- "classroom-based testing;
- self-directed testing on various aspects of a language (vocabulary, grammar, discourse, one or all of the four skills, etc.);

- practice for upcoming high-stakes standardized tests;
- some individualization, in the case of CATs;
- large-scale standardized tests that can be administered easily to thousands of test-takers at many different stations, then scored electronically for rapid reporting of results. " (p. 14,15).

He (Brown, 2004) also emphasizes some disadvantages that are present in computerized testing. Among them:

- "in classroom-based, unsupervised computerized test, there is a lack of security and possibility of cheating
- occasional "home-grown" quizzes that appear on unofficial websites may be mistaken for validated assessments;
- the multiple-choice format preferred for most computer-based tests contains the usual potential for flawed item design;
- open-ended responses are less likely to appear because of the ·need for human scorers, with all the attendant issues of cost, reliability, and turn-around time;
- the human interactive element (especially in oral production) is absent. " (p. 15)

1.6 Types of assessment

1.6.1 Informal and formal assessment

Formal and informal assessments are two general types of assessements.

Informal assessment can take a number of forms, from incidental, unplanned remarks and replies to coaching and other improvizational feedback to students. As Brown (2004) pointed out it includes examples like saying "Nice job! " "Good work! " "Did you say can or can't? " "I think you meant to say you broke the glass, not you break the glass", or putting a ☺ on some homework. Most of the informal assessment of a teachers is based on classroom activities designed to elicit performance without recording results or making judgements about a student's competence. Examples of this include marginal comments on papers, responding

to a draft of an essay, advice about how to properly pronounce a word, a suggestion for a technique to compensate for a reading difficulty, and showing how to change a student's note-taking to better remember a lecture's material (Brown, 2004).

On the other hand, formal assessments, as Brown (2004) defines them, are "exercises or procedures specifically designed to tap into a storehouse of skills and knowledge" (p. 6). They are systematic, planned sampling techniques that are used to provide both the teacher and the students with an evaluation of student achievement. He (Brown, 2004) even uses a tennis analogy, comparing formal assessments to tournament games that occur systematically during the course of a regimen of practice.

Knowing all this, a question arises: is formal assessment the same as a test? While all tests can be said to be formal assessments, not all formal assessments can be classified as tests. For example, a teacher might use a student's journal or portfolio as a formal assessment of the fulfillment of certain goals, but it is questionable to call these two procedures "tests". A systematic collection of observations of a student's oral participation in class is unquestionably a formal assessment, but it is hardly a test (Brown, 2004).

1.6.2 Formative or summative assessment

Another useful distinction to bear in mind is the function of assessment. From this viewpoint assessment can be either formative or summative.

Formative assessment happens when we test students in order to help them to perform better next time. One might say that a lot of correction in oral or written form is a kind of mini formative assessment. Summative assessment, on the other hand, happens when we want to see how well students have done. It involves testing their knowledge at the end of a period of time, such as a semester or a year, or in some public exam (Harmer, 2007).

Most of our classroom assessment is formative, which involves assessing students in the process of "forming" their skills and competencies with the

intention of helping them to continue their imrovement. Practically, all kinds of informal assessment are formative. Their main focus is on the ongoing development of the language of the learner. So if one gives a students a remark or a suggestion, or call attention to a mistake, that feedback given is intended to enhance the language ability of the learner.

Brown defines summative assessment as something that is meant to evaluate, or summarize what a student has learned, and typically occurs at the end of a course or unit of instruction. A summation of what a student has learned accomplished means looking back and taking note of how well that student has met the set objectives, but not necessarily pointing the way forward. Final exams and general proficiency exams are examples of summative assessment (Brown, 2004).

The following table summerizes some key differences between formative and summative assessment.

Table 2

Key differences between formative and summative assessment (adapted from Assessment Types: Diagnostic, Formative ans Summative, in *Teaching and Learning in Higher Education*)

Formative	**Summative**
used during the leraning process	used at the end of the learning process
provides feedback on learning-in-process	evaluates student learning against some standard or benchmark
dialogue-based, ungraded	graded

Paul Black (1998), who is often considered to be the forefather of these concepts, used the analogy of cooking to describe the difference between the terms of formative and summative assessments. When a cook is making a soup, he or she tastes it every now and then to see if it needs more spices or ingredients. Every time the cook tastes the soup, the cook is assessing it and uses that feedback to change or improve it. To put it another way, the cook is engaging in formative

assessment. Then, when the soup is served to the customer, the customer tastes it and makes a final judgment about the quality of the soup – otherwise known as summative assessment.

1.7 Types of tests

There are many kinds of tests and each test has specific purpose and a particular criterion to be measured. The following types of tests are differentiated on he basis of their purposes: proficiency test, diagnostic test, placement test, achievement test and language aptitude test.

1.7.1 Proficiency tests

Hughes (1989) defines proficiency tests as tests that "are designed to measure people's ability in a language regardless of any training they may have had in that language" (p.9). Therefore, the content of a proficiency test is based on the specification of what candidates must be able to do in order to be considered proficient in the language, and not on the content or objectives of language courses. This raises the question: what we mean by the word "proficient". In the case of some proficiency tests, "proficient" means "having sufficient command of the language for a particular purpose" (Hughes, 1989, p. 9). A test to determine whether someone can be a successful United Nations interpreter, or a test to determine whether a student's English is sufficient enough for a British University study course, are both good examples of proficiency tests. Whatever the particular purpose, it will be reflected in the test content specification at an early stage of a test's development.

There are other kinds of proficiency tests in which the concept of proficiency is more general, like the Cambridge examinations or the Oxford EFL examinations. The purpose of these tests is to show whether candidates have reached a certain level in terms of certain specified abilities. Despite tha lack of a clear objective, these general proficiency assessments should provide a thorough specification of what successful candidates would need to demonstrate. Each test

should be based on these specifications in order for all test users to be able to determine whether the test is suitable for them.

Regardless of their differences, all proficiency tests have one thing in common - "they are not based on courses that candidates may have previously taken" (Hughes, 1989, p. 10).

1.7.2 Achievement tests

Unlike proficiency tests, achievement tests are directly linked to language courses, their purpose being to identify how successful individual students or groups of students have been in achieving objectives (Hughes, 1989). Brown (2004) states that "achievement tests can also serve the diagnostic role of indicating what a student need to continue to work on in the future, but the primary role being to determine whether course objectives have been met – and appropriate knowledge and skills acquired – by the end of a period of instruction" (p. 47-48).

Hughes distinguishes two kinds of achievement tests: final and progress. Final achievement tests are administered at the end of a course of study and the content of these tests must be related to the courses that they are dealing with. The nature of relationship between the test and the course is a matter of dispute between language testers. According to some testers, the content of a final achievement test should be based directly on a detailed course syllabus or/and on other books and materials that are used. This is offen referred to as "syllabus-content approach". It has an appearent appeal, since the test only covers what the students are supposed to have encountered and thus can be considered as a fair test in this way. The disadvantage is that if the syllabus or the book and other materials are poorly chosen, then the results of a test can be very deceptive. Successful test performance does not always imply successful achievement of the course objectives (Hughes, 1989).

According to Hughes (1989) an alternative approach would be to base the test content directly on the objectives of the course. This has a number of advantages:

- "first, it compels course designers to be explicit about objectives;
- secondly, it makes it possible for performance tests to show just how far students have achieved those objectives" (p. 11).

Progress achievement test is a type of achievement test, that is meant to assess the progress that students are making. As progress is towards the achievement of course objectives, these tests should also contribute to objectives. The question is: how? Hughes (1989) explains that the best way to measure process is to set a series of short-term objectives. He (Hughes, 1989) adds, that when "the syllabus and instruction are relevant to these objectives, progress tests based on short-term objectives will fit well with what has been taught" (p. 12).

1.7.3 Diagnostic tests

Joelle Brummitt-Yale (2017) defines diagnostic assessment as "a form of pre-assessment that helps a teacher to evaluate students' individual strengths, weaknesses, knowledge, and skills prior to instruction". Diagnostic tests are primarily used to diagnose student difficulties and to guide lesson and curriculum planning accordingly. Harmer (2007) likens a teacher giving diagnostic tests to students to a doctor who is diagnosing a patient's symptoms. Both the teacher and the students benefit from diagnostic assessment.

J. Brummitt-Yale identifies three major benefits of diagnostic tests (2017):

1. It allows teachers to plan efficient and meaningful instruction. When a teacher knows exactly what students know or don't know about a topic, he or she can tailor the lessons to the topics that students still need to learn rather than to what they already know. This reduces student frustration and boredom.
2. A diagnostic test provides information that can be used to individualize instruction. It may show a teacher that a small group of students needs additional instruction on a particular aspect of a unit or course of study. The teacher can then provide remediation for those students so that they can fully engage with new material. Similarly, if a teacher notices that a group of students has already mastered a significant portion of a unit of study, he or she

can design activities that allow the students to go beyond the standard curriculum.

3. It establishes a baseline for assessing future learning. It shows both the teacher and the students what is known prior to instruction. Thus, it creates a baseline on a topic. The students can see what they are learning or not leraning as they move through instuction and the teacher can provide remediation or enrichment as needed.

Unfortunately, very few tests are designed for sosely diagnostic purposes, since their size would make it impractical to administer on a regular basis. The lack of diagnostic tests is unfortunate, because they could be very useful for individualized or self-instruction. "Learners would be shown where gaps exist in their use of the language, and could be directed to sources of information, exemplification and practice" (Hughes, 1989, p. 14).

1.7.4 Placement tests

A placement test is a test that measures someone's ability in order to put that person in a particular class or group. Normally they are used to put students to classess of different levels (Hughes, 1989). A placement test usually includes a sample of material to be covered in the curriculum, and it should indicate the point at which the students will find a level to be neither too easy nor too difficult, but sufficiently challenging.

Placement tests come in many varieties, including those that assess comprehension and production, as well as those that assess written and oral performance, multiple choice, and gap filling formats. One of the examples of placement tests is the English as a Second Language Placement Test (ESLPT) at San Francisco State University (Brown, 2004).

Hughes (1989) suggests that placement tests may be purchased, but it is not to be recommended unless the institution in question is absolutely certain that the test suits its particular teaching programme. There is no one-size-fits-all placement test that will work with every institution. The most successfull placement tests are

those constructed for specific situations. They depend on identifying the key features at different levels of teaching in the institution. "They are tailor-made rather than bought off the peg" (Salim, 2001, p. 178).

1.7.5 Language aptitude test

The purpose of a language aptitude test is to measure a person's aptitude, thus the individual's ability to learn a foreign language. According to John Carrol and Stanley Sapon (1959), language aptitude tests are used to determine how well a person can learn a foreign language in a given amount of time and under given circumstances, rather than whether or not they can learn a foreign language in general.

In his review of early aptitude research, Caroll (1981, cited in Ellis, 1999) defines general aptitude as "capability of learning a task which depends on some combination of more or less enduring characteristics of the learner" (p. 490). Language aptitude is thus a kind of special gift for learning languages and is analogus with other special skills such as musical talent or chess mastery.

Carroll administered a large number of tests and through factor analysis was able to detect a relatively small number of factors which he interpreted as the abilities that underlie successful L2 acquisition. These abilities were later confirmed by subsequent studies. Caroll's (1959) reasearch led to the development of Modern Language Aptitude Test (MLAT). The primary purpose of MLAT was to assist the US Government in finding and training people who would be successful learners of a foreign language in an intensive program. The Modern Language Aptitude Test is now the property of the Second Language Testing Foundation, which is a non-profit organization.

According to Caroll's (1959) model of language aptitude there are four major abilities involved (cited in Ellis, 1999):

1. Phonemic coding ability – the ability to code foreign sounds in a way that can be remembered later.

2. Grammatical Sensitivity – the ability to recognize the grammatical functions of words in sentences.
3. Inductive Language Learning Ability – the ability to identify patterns of correspondence and relationships involving form and meaning.
4. Rote learning ability – the ability to form and remember associations between stimuli.

The MLAT consists of five sections. Each section measures a particular skill required to acquire a new language. The sections are the followings (Ellis, 1999):

1. Number Learning, where learners are asked to learn words for numbers in an artificial language;
2. Phonetic Script, where learners are asked to listen to sounds and learn the phonetic symbols for them;
3. Spelling Clues, when learners are required to decipher phonetically spelt English words, which they must then identify with words with the same meaning;
4. Words in Sentences, where learners have to recognize the syntatic functions of words and phrases in sentences;
5. Paired Associates, which is a test of learners' ability to learn and recall paired associates.

1.8 Types of testing

1.8.1 Direct versus indirect testing

We distinguish between two approaches to test construction: direct and indirect testing.

According to Hughes (1989), testing is said to be direct when it "requires the candidate to perform precisely the skill that we want to measure" (p.15). For example, if our objective is to assess how well candidates pronounce a language, then the direct testing method would be getting them to speak. When measuring the productive skills of speaking and writing, direct testing is easier to conduct, since the very acts of speaking and writing provide us with information about the

candidate's ability. However, when it comes to listening and reading, candidates must not only listen and read but they must also demonstrate that hey have done so successfully. That is why one may find it interesting, that the testing of productive skills are mainly presented as being most problematic, for reasons usually connected with reliability (Hughes, 1989).

Hughes (1989) lists three main advantages of direct testing:

- firstly, if we are clear about what abilities we want to assess, it is relatively simple to create the conditions that will elicit the behaviour on which our judgements will be based;
- secondly, assessing and interpreting the performance of the students performance is also quite straightforward, especially in the case of the productive skills;
- thirdly, since practicing for the test entails practicing the skills that we wish to improve, there is likely to be a beneficial backwash effect.

Indirect testing aims to assess the abilities that underline the skills in which we are interested. The primary appeal of indirect testing, according to Hughes (1989), is the possibility of testing a representative sample of a finite number of abilities. Direct testing, on the other hand, is typically limited to a small number of tasks including a limited range of grammatical structures.

The main issue with indirect testing is that the relationship between how well the candidates perform on them and the performance of the skills in which we are usually more interested is often rather weak and ambiguous in nature. Hughes (1989) suggests that focusing on direct testing is preferable because it allows for more precise estimates of the skills that really matter to us. In addition, the fact that direct tests are generally easier to construct simply reinforces this view, as does their greater potential for beneficial backwash.

1.8.2 Norm- referenced versus criterion-referenced testing

We distinguish between two types of testing based on the interpretation of test scores: criterion-referenced and norm-referenced testing. Robert Glaser (1963), an

American psychologist dedicated to the educational field, created the terminology for both of these methods of assessment.

A norm-referenced test interpretation defines the performance of test-takers in relation to one another (Underhill, 1987). In this case, candidates are compared with other, rather than applying a mark scheme that was predefined. "In the case of a norm-referenced test, we are not told directly what the student is capable of doing in the language" (Hughes, 1989, p. 17).

In contrast to norm-referenced tests, criterion-referenced tests define the performance of each test-taker without regard to the performance of others. Unlike the norm-referenced tests, criterion-referenced tests define success as being able to perform a specific task or set of competencies (Sharon & William, 2008). Caroll (1970) suggested that a CRT yields results which indicate as precisely as possible whether the pupil has achieved the specified goals of the learning task.

According to Hughes (1989), one of the most appealing aspects of criterion-referenced tests is that students are encouraged to measure their progress in relation to meaningful criteria, without feeling that they are destined to failure, because they are less able than their peers.

He (Hughes, 1989) defines two other positive virtues of criterion referenced tests:

- they establish meaningful standards in terms of what people can do, that are consistent across different groups of candidates, and
- they motivate students to attain those standards.

1.8.3 Objective testing versus subjective testing

The distinction here is between the methods of scoring. An objective test is on which equally competent scorers will obtain the same scores, whereas a subjective test is one where the scores are influenced by the opinion or judgement of the person doing the scoring.

Harmer (2007) defined a number of advantages and disadvantages of objective and subjective testing. These are concluded in Table 3.

Table 3

Advantages and disadvantages of objective and subjective testing

(adapted from Harmer, 2007)

Objective testing	Subjective testing
Advantages: - objective tests can be administered by one person - objectives tests are easy to correct - objective tests can be corrected by a machine - developed test items can be reused	Advantages: - subjective test items are easier to write - the cost of development is quite low - subjective questions are suitable for testing a broad range of learning tasks - subjective tests can be administered online to students being tested at a distance
Disadvantages: - good objective test items are difficult to write - objective test items are not appropriate for every learning objective	Disadvantages: - grading is time-consuming - only instructors and experts can grade them - grading costs may be higher - can be administered online, but it must be graded by humans with adequate expertise

According to Hughes (1989, 2003), the difference between these two types is the way of scoring and presence or absence of the examiner's judgement. If there is not any judgement, the test is objective. Many testers seek objectivity in scoring for the greater reliability — the less subjective the scoring, the greater agreement there will be between two different scorers. However, there are ways of obtaining reliable subjective scoring, even in the case of compositions.

1.9 Principles of language assessment

Many experts consider validity and reliability to be the most important criteria in judging the quality of a test (Bachman & Palmer, 1996; Davies, 1990). "How accurately we measure what we purpose to measure is based on estimates of reliability and validity" (Davies, 1990, p. 10). Others, including Weir (1990) would prefer to include practicality to this criteria. However, Coombe and Hubley (2007) argue that other factors come into play apart from validity, reliability and practicability, such as washback, authenticity, transparency, and security.

Brown (2004) defines the following five cardinal criteria for "testing a test":

- practicability
- reliability
- validity
- authenticity, and
- washback

These principles offer valuable guidance both in evaluating an existing assessment procedure and in developing one on your own. In the following, Brown's five cardinal criteria are explained in more detail.

1.9.1 Practicability

A test should be practical in terms of time, cost, and energy. An effective test is practical. As Brown (2004) defines, it means that it:

- "is not excessively expensive,
- stays within appropriate time constraints,
- is relatively easy to administer, and
- has a scoring/evaluation procedure that is specific and time-efficient" (p. 19)

A test that is overly expensive is impractical. A language proficiency test that takes a student long hours to complete is impractical, since it consumes more time and money than necessary to accomplish its objective. Similarly, a test that takes a few minutes for a students to complete and several hours for an examiner to assess, is also impractical. The same applies to a test that can only be scored by

a computer and it takes place a thousand miles away from the nearest computer (Brown, 2004).

1.9.2 Reliability

Reliability refers to consistency and dependability. It means that the results of the same test given to the same student on two separate occasions should produce similar results. The more similar the scores, the more reliable the test is said to be (Hughes, 1989). A number of factors can lead to unreliability. These factors may be functuations in the test iself, in the student, in test administration and scoring (Mousavi, 2002). Let us take a closer look at each of these factors (adapted from Mousavi, 2002):

- Student-related reliability

The most common learner-related reliability problem is caused by temporary illness, exhaustion, anxiety, and other physical or psychological factors, which can cause the "obeserved" score to deviate from the "true" score.

- Rater reliability

The scoring process may be influenced by human error, subjectivity and bias. Here we can make a distinction between inter-rater and intra-rater reliability. (See Part 2, 2.5.1).

- Test administration reliability

The conditions under which the test is administered may also cause unreliability. It involves a situation, for example, where students seated next to the windows could barely hear the tape during the listening comprehension test because of the noise ouside of the building.

- Test reliability

Sometimes, it is the nature of the test itself what is the source of problems. If a test is too long, test-takers may get tired by the time they reach the later items and respond incorrectly. Students who do not perform well under the pressure of a set

time limit may be discriminated against in timed tests. Poorly written test items can also be a cause of test unreliability.

1.9.3 Validity

The principle of validity is by far the most complicated and perhaps the most important criterion of an effective test. Gronlund (1998) defines validity as "the extent to which inferences drawn from assessment results are appropriate, meaningful and useful in terms of the purpose of the assessment" (p. 226). Heaton (1988) gives a simpler definition of validity describing it as "the extent to which it measures what it is supposed to measure" (p. 159).

How is the validity of a test established? Validity is a unitary concept (Bachman, 1999) and to gain valid inferences from test scores, a test schould have some kinds of evidence.

- Content-related evidence

A test can assert content-related evidence, if it samples the subject matter from which conclusions are to be drawn and requires the test-taker to perform behavior that is being tested (Hughes, 2003; Mislevy & Bock, Mousavi, 2002). As Hughes (1989) states "a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned" (p. 22).

Another way of understanding content validity is to consider the difference between direct and indirect testing. "Direct testing involves the test-taker in actually performing the task. In indirect test, learners are not performing the task itself but rather a task that is related in some way" (Brown, 2004, p. 23-24).

- Criterion-related evidence

Criterion-realted evidence is "the extent to which the criterion of the test has actually been reached" (Brown, 2003, p. 24). Criterion-related evidence is best demonstrated in the case of teacher-made classroom assessments by comparing the results of an assessment with the results of another measure of the same criterion.

For example, comparing the results of a teacher-made test about present continuous tense to the results of a test of the same topic in a textbook.

There are essentially two kinds of criterion-related validity: concurrent and predictive. "Concurrent validity is established when the test and the criterion are administered at about the same time" (Hughes, 1989, p. 23). Predictive validity, on the other hand, focuses on using test results to predict future performance. As Hughes (1989) defines it, "predictive validity refers to the degree to which a test can predict candidates' future performance" (p. 25).

- Construct-related evidence

According to Hughes (2003) a test, a part of a test, or a testing technique is said to have construct validity if it can be proven that it measures just the ability which it is supposed to measure. The word "construct" refers to any underlying ability (or trait) which is hypothesised in a theory of language ability. It does not play as large role for classroom teachers. "Constructs may or may not be directly or empirically measured – their verification often requires inferential data" (Brown, 2004, p. 25). In a sense, tests are operational definitions of constructs in that they operationalize the entity that is being measured (Davidson, Hudson & Lynch, 1985).

- Consequential validity

Consequential validity comprises all the consequences of a test, including its accuracy in measuring intended criteria, its impact on the training of test-takers, its influence on the learner and the social consequences of a test's interpretation and use (Brown, 2004). Messick (1989), Gronlund (1998), McNamara (2000) and Brindley (2001) among others, underscore the potential importance of the consequences of using an assessment.

As high-stakes assessment has gained popularity in the last two decades, particular attention has been drawn to one aspect of consequential validity: the effect of test preparation courses and manuals on performance. McNamara (2000) warns against test results that may indicate socieconomic factors such as coaching

opportunities. It is because thesse opportunities are differentially available to the students being assessed. For instance, because only some families can afford coaching, or because children with more highly educated parents get help from their parents.

Another important consequence of a test is washback. Weir (1990) calls this evidence "washback validity". Gronlund (1998) encourages teachers to consider the effect of assessments on students' motivation, subsequent performance in a course, independetn learning, study habits, and attitude toward school work.

- Face validity

Face validity can hardly be considered a scientific concept, yet it is very important. Hughes (1989) suggests that a test has face validity if "it looks as if it measures what it is supposed to measure" (p. 27). Gronlund thinks about validity as "the extent to which students view the assessment as fair, relevant and useful for improving learning" (1998, p. 210). According to Mousavi (2002), "face validity refers to the degree to which a test looks right and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgement of the examinees who take it, the administrative personnel who decide on its use and other psychometrically unsophisticated observers" (p. 244).

Brown (2014) notes that face validity is likely to be high if learners encounter:

1. "a well-constructed, expected format with familiar tasks,
2. a test that is clearly doable within the allotted time limit,
3. items that are clear and uncomplicated,
4. directions that are crystal clear,
5. tasks that relate to their course work (content validity), and
6. a difficulty level that presents a reasonable challenge. " (p. 27)

See Part 2, section 2. 7 for more details on validation procedures.

1.9.4 Authenticity

Bachman and Palmer (1996) define authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a target language task" (p. 23).

Many test items fail to simulate real-world tasks. They may be artificial in their attempt to target a grammatical form or lexical item. Brown (2004) notes that the authenticity of test tasks has improved noticeably in recent years. It was one assumed that largescale testing could not include performance of the productive skills, but now many such tests offer speaking and writing components. Reading passages are selected from real-life sources that test-takers are likely to have come across or may come across. Listening comprehesnion sections feature natural language with hesitations, white noise and interruptions.

Brown (2004) lists the ways in which authenticity may be present in a test:

- "the language in the test is as natural as possible,
- items are contextualized rather than isolated,
- topics are meaningful, relevant and interesting for the learner,
- some thematic organiztaion to items is provided such as through a storly line or episode,
- tasks represent, or closely approximate, real-world tasks. " (p. 28)

1.9.5 Washback

According to Hughes (2003), washback is "the effect of testing on teaching and learning" (p.1). Washback is a term used in large-scale assessment to describe the impact of tests on instruction in terms of how students prepare for the test (Brown, 2004).

Saehu (2012) mentions the negative and positive aspects of washback. Negative washback is not hard to find, such as focusing only on exam-related language skills and ignoring the rest. While language is a medium of communication, the majority of students and teachers in the language classes focus exclusively on language competencies in the test. On the other hand, if a test

encourages better teaching and learning, it produces positive washback. Washback improves a number of fundamental principles of language acquisition, like intrinsic motivation, autonomy, self-confidence, language ego, interlanguage, and strategic investment, among others (Brown, 2004). Washback can also be strong or weak (Saehu, 2012). A national examination is a good example of strong effect, meanwhile, the impact of a formative test is small.

The task for teachers is to create classroom tests that function as learning instruments by which washback is achieved.

1.10 Testing teachers - TETP

There are various examination boards around the world that administer exams that qualify applicants to teach English as a foreign language. Some of the most well-known are RSA/DTEFLA (Diploma in the Teaching of English as a Foreign Language to Adults), COTE (Certificate of Overseas Teachers of English), or DOTE (Diploma for Overseas Teachers of English). These examinations assess not just the candidates' level of language proficiency. They include a range of exams from methodology to practice teaching.

I.Hock in her book of *Test Construction and Validation* (2003) proposes a test that was designed to assess future English teachers. The test is called TETP – Test of English for Teaching Purposes. The prospective English language teachers who are regularly tested on their knowledge of pedagogical content in traditional subject matter tests during their training years, are the target population of TETP.

Since the content of the test and the methods of testing are derived from a study of language use in a particular context, TETP is a Language for Specific Purposes (LSP) test, thus it replicates the tasks that frequently occur in target language situations. LSP tests involve activities that not only engage test takers' communicative language abilities, but also their knowledge of the field in which they are being assessed. According to Douglas (2000), the material on which a specific purpose test is based must engage test takers in activities in which both language skill and field awareness communicate with the best content in a manner

that is close to how the target language is used. TETP, on the other hand, claims not to test pedagogical material, such as knowledge of communicative language teaching's basic principles. Pedagogy-specific content knowledge is not part of the construct.

The TETP construct is characterized by an association between language knowledge and language-related pedagogical skills that can be identified and are thought to be essential for judging language performance in the classroom. The tests tasks include the measurement of specific language-related pedagogical skills which may affect the success of language performance during classroom language teaching. I.Hock [54] defines TETP as "a criterion-referenced, performance-based test of productive communicative skills to be used in simulated authentic classroom teaching situtations".

The TETP consists of two papers: an Oral Examination and a Use of English Paper.

The Oral Section is claimed to be communicative since it is based on a construct of contextualized language proficiency. The error correction part of the oral exam lasts for 15 minutes and is done in writing. The rest of the test, which involves four performance-based tasks, takes 35 minutes for a pair of candidates to accomplish. Candidates must complete the following tasks during the Oral Examination: note-taking, error correction and summary, role-playing and problem-solving, providing instructions and explanations.

The Use of English Paper lasts for 100 minutes and involves eleven tasks of language knowledge and use. The Use of English Paper includes task types such as multiple choice, sentence tranformation, sentence and text-level error identification and correction, matching, gap-filling, word formation, phonetic transcription of words, providing hungarian versions of word, providing synonyms in authentic texts, and parapharisng words occuring in authentic texts.

The papers are reported on a three-point scale: fail, pass, good pass. If a candidate fails one of the two papers, they would only need to retake the paper they have failed.

After the development of construct, specifications, item writing and moderation stages the TETP was pretested in a main trial. The final draft version of TETP was administered to four groups of candidates, that comprise 65 students altogether, at the University of Veszprém, in 1999. All the candidates were 4th and 5th year university students training to become secondary school EFL teachers.

From the results and analysis it can be concluded that the content of the TETP complies reasonably well with its stated purposes, and the test is fairly consistent in its construction, administration, and scoring. However, there were certain disrepancies.

One of them is that the administration of the test does not represent genuiune performance testing, since there are no actual students in the classroom. Nevertheless, administering the exam with real students would be time-consuming and could lead to uncontrolled speech patterns and threaten the reliability of the scoring and therefore it could make the language assessment invalid.

Another problem is that due to limitations in participants, the measurement of using appropriate functional language to encourage, praise and give feedback on student performance is not applicable in the testing situation.

Analyses of item level statistical figures have addressed issues of consistency of test taker perfromance across all the items. For reasons of poor discrimination, approximately 20% of all the items in the Use of English Section of TETP were changed or replaced.

During the trial of the examination, low scores occured because the lack of preparation, time and concern interfered with test takers' demonstration of competence.

Other results gained from analyses of group differences in performances on both Oral and The Use of English Section provide evidence for the external aspect of construct validity. In both sections the groups had been anticipated to gain the highest and the lowest scores did indeed do so (Hock, 2003).

This chapter of the master thesis provided us with an overview of how the concept of assessment has changed over time. First, we distinguished between testing and assessment, and then the concept of assessment literacy was also discussed. This chapter gave a report on how alternative methodologies have gained ground and also on the characteristics of different types of tests and testing methods. It also gave a brief introduction to the principles of language assessment and the last section briefly introduced a test designed to assess future English teachers in Hungary.

PART 2

Stages of test construction

Language testing is crucial to language teaching (Davies, 1990). A test is an important instrument used to find out how well students are learning and how effectively teachers are teaching. Since tests play such an important role in the learning-teaching process, we need to understand how tests are and should be constructed, in order better to understand the assessment process and to select from a range of available tests one instrument that is suitable for their own contexts.

Alderson, Clapham & Wall (Alderson, Clapham & Wall, 1995) define the following stages of test construction:

1. test specification
2. item writing and moderation
3. pretesting and analysis
4. training examiners and administrators
5. monitoring examiner reliability
6. reporting scores and setting pass marks
7. validation
8. post-test reports

2.1 Test specification

The construction of a successful test, according to language testing experts, begins with specification (Alderson, 2000; Hughes, 2003; Messick, 1989). Tests that are constructed without specification are more likely to fail in practice. Brown (1994) refers to test specifications as "practical outlines of your test" (p. 387). Bachman and Palmer (1996) liken test specification to a road map which describes how actual test tasks are to be constructed, and how these tasks are to be arranged to form the test. Alderson, Clapham & Wall (1995) define specifications as "the official statement about what the test tests and how it tests it" (p. 9). McNamara (2000) defines specification as "a set of instructions for creating a test" with the aim of making design decisions explicit and allowing new versions to be written in

the future by someone other than the creator of the original test (p. 31). Davidson and Lynch (2002) added that test specification can also indicate a test's purpose, its motivation and context.

According to Hughes (2003), even if the test specification is very precise, "it is not to be expected that everything in the specification will always appear in the test, there may simply be too many things for all of them to appear in a simple test" (p. 27).

Douglas (2000) suggests that test specification should at least contain the following minimum elements:

- "a description of the test content and organization of the tasks,
- a description of the number and type of test tasks, time allotment for each task and specification for each test task,
- the criteria for item correctness, and
- sample task items" (p. 110-113).

In their *Language Test Construction and Validation*, Aldersen, Clapham & Wall (1995) define test specifications as answers to the following questions:

1. "What is the purpose of the test? Tests tend to fall into one of the following broad categories: placement, progress, achievement, proficiency and diagnostic.
2. What sort of learner will be taking the test – age, sex, level of proficieny/stage of learning, first language, cultural background, country of origin, level and nature of education, reason for taking the test, likely personal and, if applicable, professional interests, likely levels of background knowledge?
3. How many sections/Papers should the test have, how long should they be and how will they be differentiated?
4. What target language situation is envisaged for the test, and is this to be simulated in some way in the test content and method?
5. What text types should be chosen – written and/or spoken? What should be the sources of these, the supposed audience, the topics, the degree of authenticity? How difficult or long should they be? What functions should be embodied in the texts? How complex should the language be?

6. What language skills should be tested? Are eanbling/micro skills specified, and should items be designed to test these individually or in some integrated fashion? Are distinctions made beteen items testing main idea, specific detail, inference?
7. What language elements should be tested? Is there a list of grammatical structures/features to be included? Is the lexis specified in some way – frequency lists etc.?
8. What sort of tasks are required – discrete point, integrative, simulated „authentic", objectively assessable?
9. How many items are required for each section? What is the relative weight for each item – equal weighting, extra weighting for more difficult items?
10. What test methods are to be used – multiple choice, gap filling, matching, transformation, short answer question, picture description, role play with cue cards, essay, structured writing?
11. What rubrics are to be used as instructions for candidates? Will examples be required to help candidates know what is expected? Should the criteria by which candidates will be assessed be included in the rubric?
12. Which criteria will be used for assessment by makeres? How important is accuracy, appropriacy, spelling, length of utterane/script, etc.?" (p. 11-13)

2.2 Item writing and moderation

The next stage of test construction is item writing and moderation. Aldersen, Clapham & Wall (1995) formulate the following questions that need to be answered in this stage of test construction:

- "Where do you start when writing an item?
- What methods are most suitable for testing particular abilities?
- When people disagree about the quality of a test item, how can we resolve the disagreement?
- What principles and guidelines should we follow when writing test items?

- What is the role of the moderating committee, and how do such committees best work?" (p. 40)

The first step when writing a test item is to look at the test specifications. After consulting the test specification, what to do next depends on what kind of test is being designed. If the test is one of the language elements, the next step would probably be to consult past papers or some inventory of the content of previous tests to avoid the possibility of too much repetition of content across tests. For many tests, the next task of the item writer is to find appropriate texts. Before moving to create items or activities based on the chosen text, it is often a good idea to get the approval of the editing or moderating committee. It is both wasteful and depressing to spend time creating items in texts that would finally be rejected.

When choosing which method to use to test a particular ability, it is important to be attentive to the so-called *method effect*, because the method used for testing a language ability may itself affect the student's score. During testing, we are not curious about how good a test writer is at certain types of tasks, but whether he or she has the necessary grammatical knowledge, speaking or reading skills.

It is being investigated by more and more research today how students actually respond to particular test methods. There has been considerable reaserch done into the Cloze technique and C-test. Different cloze tests measure various skills, thus a test created by applying the technique to a text may or may not measure the same thing on the same text as a different cloze test. In brief, without validating the test in the usual manner, one cannot know in advance what a given cloze test will measure. This implies that the method effect of the cloze techhnique is quite complex.

Nevertheless, there is some proof that many candidates read in a different manner than normal when they take cloze tests. They read the short amount of context just before the blank, but often neglect to read the context after the blank. Alderson, Chapman & Wall (1995) suggests that the possible explanation behind the phenomenon may be that the presence of blanks at regular intervals tends to

induce a form of "short text" reading, and many close test takers show a lack of attention to the meaning of the wider context that is not shown by their normal reading, when they are indeed context-sensitive.

Students taking multiple-choice tests have also been shown to develop techniques that "artificially" inflate their scores. These strategies include techniques for guessing the correct answer, for eliminating implausible distractors, for avoiding two options with very similar meaning, for choosing an option that is notably longer than the other distractors, etc.

Hughes (2003) experienced that multiple choice tests that are produced for use within institutions are often shot through with faults like: there is more than one correct answer, there is no correct answer, there are clues in the options as to which is correct, and the presence of ineffective distractors. Hughes (2003) also stresses the harmfulness of backwash. When a test that is important to students is multiple choice in nature, there is a chance that practice for the test will adversely affect learning and teaching. Multiple choice item practice is not the best way for students to develop their language skills.

Unfortunately, since our understanding of the test method affect is still rudimentary, it is difficult to recommend specific methods for assessing specific language skills. Even though the effects of the various test methods are not known, item writers need to be aware of the drawbacks of specific test methods and learn to avoid the most common errors in designing certain types of test items.

Alderson, Clapman & Wall (1995) highlight that there are certain problems that apply to all types of tests and the issue of what an item is actually measuring is probably the most fundamental one. It is very easy with many kinds of test items to test something that is not intended. It is very common for intelligence to be measured as well as or instead of language, especially in high-level proficiency tests. Similarly, instead of reading or listening comprehension, background knowledge is frequently tested.

Another important problem may be discovered in marking as well. If each item worths one point, then each item should be independent of the others.

Success on one item should not depend on succes on another. If it is only possible to answer the second item after correctly answering the first, then a candidate who fails Item 1 will also fail Item 2, and will lose two points instead of one.

Instructions for all items also must be clear. It is common that students fail a test item because they do not understand what they are meant to do and not because their language is poor. If possible, the language used in the instruction should be simpler that in the item itself. There are also situations when the instructions should be written in the first language of the candidates. Giving an example of what is expected of them is also a good idea.

- Test edition and moderating committees

No one can produce a good test or a good test item without guidance. The item writer, as the designer of an item, knows what the item is intended to test. Knowing the correct answer means that the item writer has a quite different viewpoint on how students can or should process the item. Therefore, in all test development, it is absolutely vital that individuals other than the actual item writer examine each item carefully and respond to them as a student would. These outside observers need to consider the skills that are needed to complete the item successfully and compare what he or she thinks the item is testing with what the item writer believes it tests (Alderson, Clapham & Wall, 1995).

Alderson, Clapham & Wall (1995) suggests that this form of item review should ideally take place at an early stage in the construction process and need not to be a formal matter involving a whole committee. Once items have been edited into its draft stage, they should be assembled into a draft test paper for the consideration of a formal committee. This committee should include experienced item writers, teachers who are experienced in teaching for the test or in teaching the target group of learners, and possibly test experts, or even subject experts, when some form of specific purpose test is being prepared. The role of this committee is to evaluate each item and the test as a whole in terms of degree of compliance with the test specifications, level of difficulty, possible unforeseen

problems, ambiguities in the wording of items and instructions, layout issues, match between texts and questions and the overall test balance.

It is essential that the member of the moderation committee do not simply read the test and its items, they must look at each item as if they were students. This implies that the members of the committee would have to devote ample time to taking the test in advance of the editing meeting. An effective editing committee should have a firm chairperson who shall ensure that enough time is allocated to the meeting, that no more time than necessary is spent on each item, that the opinions of each member are heard and considered and that a clear decision is made by the committee and recorded by the secretary or institutional official. The recommendations of the committee need to be acted upon and incorporated in a revised test, which is then subjected to some sort of confirmatory vetting before pretesting the test.

2.3 Pretesting and analysis

2.3.1 Pretesting

Now matter how well designed an examination might be, and no matter how carefully it has been edited, it is not possible to know how it will work until it has been tried out on students. Alderson, Clapham & Wall (1995) define pretesting as "all trials of examination that take place before it is launched, or become operational or live" (p. 74).

Most of the pretesting takes place during the "main trials" but these should be preceded by less formal pretesting which may be called pilot testing. Pilot testing may vary in scope from trying out a test on a small group of colleagues to running a trial on a hundred students, but in all cases the aim is to resolve the main problems before the major trials. Pilots can be run relatively quickly and cheaply, and will provide useful information about the ease of administring the test, the time students need to complete the test, the consistency of the instructions, the type of language required for open-ended questions, the accuracy and comprehensiveness of any answer keys and the usability of marking scales. The results of the pilots

will reveal several unanticipated flaws in the test and will save time and effort when the main trial are run.

The scope of the main trialling and the kinds of analysis required depend on factors such as the importance and purpose of the exam and the degree of objectivity of the marking. The most objectively marked tests are those where the answer does not have to be provided by the candidate, but is chosen from a number of possible alternatives, and can be marked as accurately by a clerk or computer as by a trained teacher or tester. A good example of an objectively marked test may be a multiple-choice test. The most subjectively marked tests are those, such as oral interviews and essays, where the marker has only a marking scheme for guidance. Of course, there is a range of item types between these two extremes which demand a greater or lesser degree of subjectivity in marking.

One of the key questions is the number of students who should be tested in the trials. It is impossible to lay down a rule for this, as it depends on a number of factors. However, it is commonly agreed that "the more the better". Henning (1987) recommended 1000 students for trial multiple-choice tests. It is also crucial that the students should take the trial test seriously and perform on it as well as possible. Otherwise, the findings could cast doubt on the entire trial procedure. The test have to be administered in precisely the same way as the final test, ensuring that not only the administration instruction, but also the test items are presented under the same circumstances as in the live exam. The only aspect that would need to be different is the timing of the test. If the examiners want to estimate the reliability of the test, students should be allowed to take as long as they wish to complete the test (Alderson, Clapham & Wall, 1995).

2.3.2 Analysis

2.3.2.1 Correlation

It is crucial to clarify the concept of correlation before assessing individual test items. The "extent to which two sets of results agree with each other" is referred to as correlation (Alderson, Clapham & Wall, 1995, p. 77). Let us take a look at how

it works with hypothetical results on a very small number of students adapted from *Language test construction and evaluation* by Alderson, Clapham & Wall (1995).

Figure 1

Correlation = +1.0. (adapted from Alderson, Clapham & Wall, 1995, p.77)

	Test 1 (Ranks)	Test 2 (Ranks)
A	1	1
B	2	2
C	3	3
D	4	4
E	5	5
F	6	6
G	7	7
H	8	8

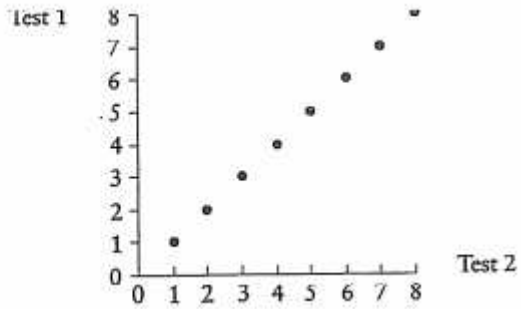
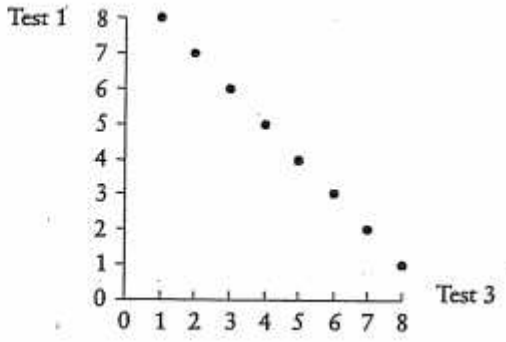


Figure 1 above gives the ranks of 8 students - students A-H - on two tests. In each case the studens were ranked identically on the two tests, so that A came first each time, B came second and so on. This is shown graphically in the scattergram. Each dot on the graph stand for student's rank on both Test 1 and Test 2. In this case, as all the ranks were the same for both tests, the dots progress diagnolly up the graph from bottom left to top right. This scattergram shows a perfect correlation between the two sets of scores. This is described as a perfect positive correlation, or a correlation of +1.0.

Figure 2

Correlation = - 1.0 (adapted from Alderson, Clapham & Wall, 1995, p. 78)

	Test 1 (Ranks)	Test 3 (Ranks)
A	1	8
B	2	7
C	3	6
D	4	5
E	5	4
F	6	3
G	7	2
H	8	1



In Figure 2 we can see what happens when the two sets of ranks are as different from each other as possible. In this case, the student who came first in Test 1 came last in Test 3 and so on. The scattergram again shows a diagonal line,

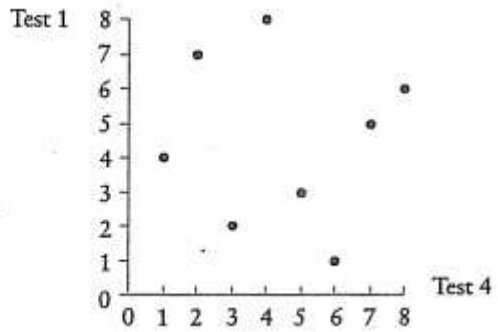
except this time it falls from top left to bottom right. This is described as a perfect negative correlation, or a correlation of -1.0. It is unlikely that there will be a strong negative correlations in the results of two language tests.

Figure 3 below shows the results of Test 1 and 4. Here, we can see, that there is no obvious relationship between the two sets of results. The dots appear all over the graph. The correlation index is for this set of results is +0.5, which is so near to .00 that we can say that there is no correlation between the two sets of results.

Figure 3

Correlation = + 0.5 (adapted from Alderson, Clapham & Wall, 1995, p. 79)

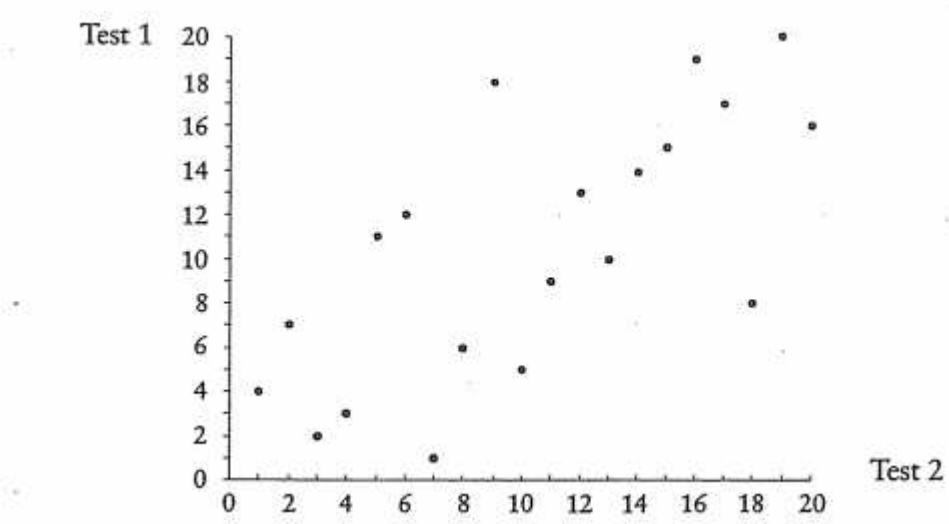
	Test 1 (Ranks)	Test 4 (Ranks)
A	1	6
B	2	3
C	3	5
D	4	1
E	5	7
F	6	8
G	7	2
H	8	4



It is not very common for there to be no correlation between two language tests. A more likely correlation between two language tests is shown in Figure 4.

Figure 4

Correlation = +.70 (adapted from Alderson, Clapham & Wall, 1995, p. 79)



Here the ranks of the students show that there was some similarity between the two sets of results, since the dots tend to progress up the graph from the bottom left to top right. However, it is not possible to connect all the dots with a single straight line. The correlation here is +7.0 which means there is a quite strong agreement between the two sets of scores.

2.3.2.2 Classical Item analysis

Traditionally there are two measures which are calculated for each objective test item – the facility value and the discrimination index. The *facility value* (F.V.) measures the level of difficulty of an item and the *discrimination index* (D.I.) measures the extent to which the results of an individual item correlate with results from the whole test (Alderson, Clapham & Wall, 1995).

- Facility value

An item's facility value is the percentage of students to answer it correctly. If there are 300 students and 150 of them get the item right, the F.V. of the item is 150/300, which is 50% (often shown as proportion:.5). If only 6/300 people get an item right, the F.V. is 2%, and is clear that the item is very difficult. Similarly if the F.V. is 95% (285/300) the item is very easy. If an item is too easy or too difficult, then it is not very informative since they tell us little about the varying levels of ability of the trial group. If examiners want a wide spread of scores from an exam, then they would prefer items which are as near to an F.V. of 50% as possible.

However, if the test contructors are more interested in ensruing that a test is at a particular level of difficulty, they have to select items with the appropriate F.V. so that the test population achieves the required mean score. The mean — also commonly referred to as the avarage — is the sum of all the students scores divided by the number of students. For example, of students get a mean score of 70% on a trial test, the mean F.V. of all the items is 70% and the test must herefore have many items with a F.V.s of over 70%.

- Discrimination index

As well as knowing how difficult and item is, it is important to know how it discriminates, that is how well it distinguishes between students at different levels of ability. If an item discriminates well, then we should expect more top-scoring students to know the answer than the low-scoring ones. If the strongest students get an item wrong, while the weakest students get it right, there is clearly a problem with the item, and it needs investigating.

The easiest way to calculate discrimination index involves ranking students according to their total scores on the test, and comparing the proportion of correct answers in the top third of the sample with those of the bottom third. For example, if the top group consists of 10 students, and 7 of them get and item right (.7), whereas only 2 out of 10 in the bottom group (.2) do, then the D.I. is .7 - .2 = +.5. An item with a D.I. of +.5 is usually considered to be discriminating well since the high scoring students have answered it better than low scoring ones.

The highest discrimination possible is +1.00, which is achieved if all the students in the top group get an answer right and none of the students in the bottom group does. Such items are very unusual. Often item writers are content with D.I.s of +.4 or above, but there are no rules as to what D.I.s are acceptable, since the possibility of getting high D.I.s varies depending on the type of test and range of ability of the examinees.

Sometimes, however, an item has a negative D.I., which means that more students in the bottom group were correct than in the top group. There is obviously something very wrong with such an item and it should be revised. The removal of these low discriminating test items may have a significant impact on test validity.

According to Mehrens and Lehman (1991), there are a number of reasons items may have low discriminating power:

1. "the more difficult or easy the item, the lower its discriminating power – but we often need such items to have adequate and representative sampling of the course content and objective; and

2. the purpose of the item in relation to the total test will influence the magnitude of its discriminating power." (p. 888)

Although item analysis is not suitable for subjectively marked tests such as essays and oral interviews, these tests still need to be pretested to see whether the items elicit the intended sample of language; whether the marking system, which should have been drafted during the writing stage, is usable and whether the examiners are able to mark consistenly. It is usually impossible to try out such tests on large numbers because of the time needed to make the script or run the interviews.

2.3.2.3 Item response theory

The above-mentioned analyses have one major drawback – the examinees' characteritics and the test characteristics cannot be separated, so that the results of the analyses are only true for the actual sample on which the trials are carried out.

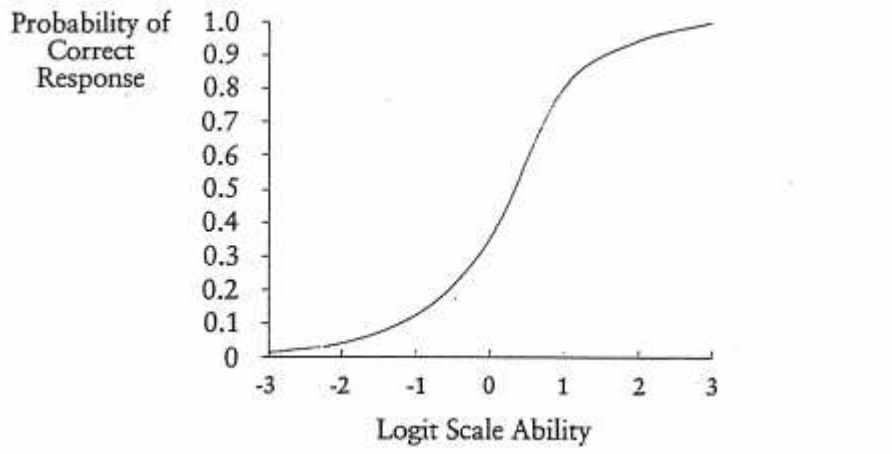
Measurement using Item Response Theory (IRT) is designed to cope with this problem. In development since the 1950s, with several seminal publications emerging in the 1960s (Birnbaum, 1958; Lord & Novick, 1968; Rasch, 1960) it was not until the 1980s that IRT realized its full potential, when computers became powerful enough to execute the complex calculations (Millman & Greene, 1993). To this day, IRT models are the preferred choice for large-scale, high stakes test administrations due to their strong theoretical underpinnings and practical benefits, including sample-free item calibration, item-free person measurement, misfitting item and person identification, and test equating and linking (Henning, 1987).

We can use IRT to establish an item difficulty scale that is independent of the sample used to test the items, allowing us to compare the performance of examinees who have taken different tests. IRT is based on probability theory and represents the probability of a given person getting a particular item right. Students' scores and item totals are transformed onto one single scale so that they can be compared. If a person's ability is equal to the difficulty level of the item, the person has a 50/50 chance of getting that item right.

An item characteristic curve (ICC) represents the relationship between the item performance of the examinees and the abilities that underpin item performance. As the level of students' ability increases, so does the probability of a correct response. A simple example of ICC can be seen in Figure 5, which shows us a logit scale. The probability of an examinee answering the item correctly is shown on the left side of the graph, and students' levels of ability are shown across the bottom. The ability levels here range from -3 to +3. It can be seen that students with an ability level of 0 have a 0.3 (or 30%) probability of answering the item correctly.

Figure 5

ICC (adapted from Alderson, Clapham & Wall, 1995, p. 90)



IRT is suitable for those who want to store items in item banks. Pretested items or sets of items can be "calibrated" according to characteristics like the ability of the individual, item difficulty and discrimination power, and store them in a bank to be used when needed. Then, when test developers are designing a new version of a test, they can select from the bank items which will not only be of a suitable level for the test population, but will also combine to form a test that is equivalent in difficulty and discriminatin to previous tests in the series (Alderson, Clapham & Wall, 1995).

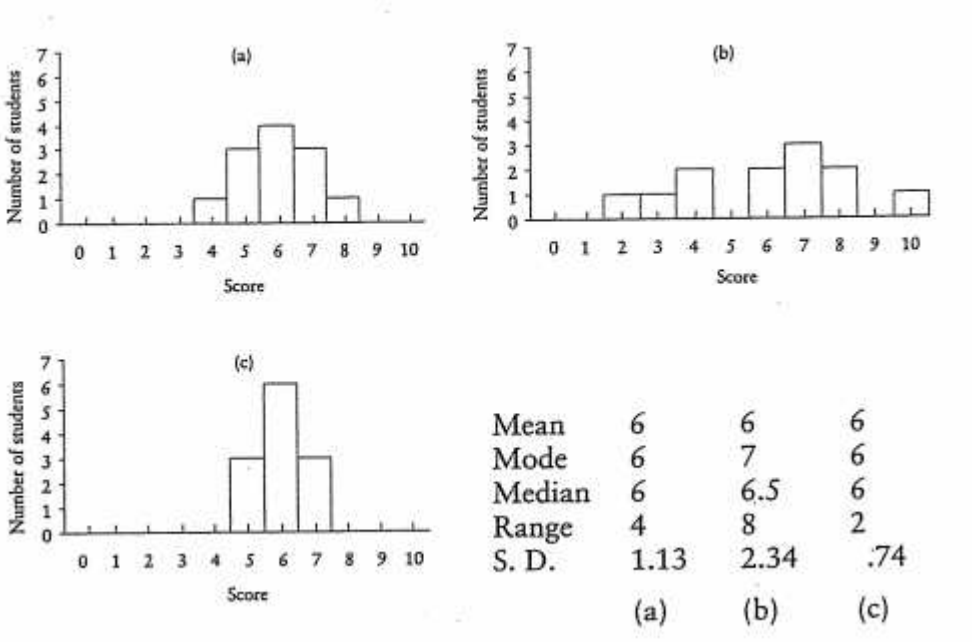
2.3.2.4 Descriptive statistics

Alderson, Clapham & Wall (1995) also draw attention to the usefulness of reporting on the overall performance of a test paper, or on the performance of sections within a test. The most important statistics to be reported are the *mean*, the *mode* and the *median*, which show how the score cluster together, and the *range* and *standard deviation*, which show how widely the scores spread out.

The following figure shows three different distributions of scores which can be described using the five statistics mentioned from from *Language test construction and evaluation* by Alderson, Clapham & Wall (1995). In all three cases, 12 students have taken a test of 10 items. The histograms show that although the mean is 6 each time, the overall test results are different.

Figure 6

Mean, mode, median, range, standard deviation (adapted from Alderson, Clapham & Wall, 1995, p. 93)



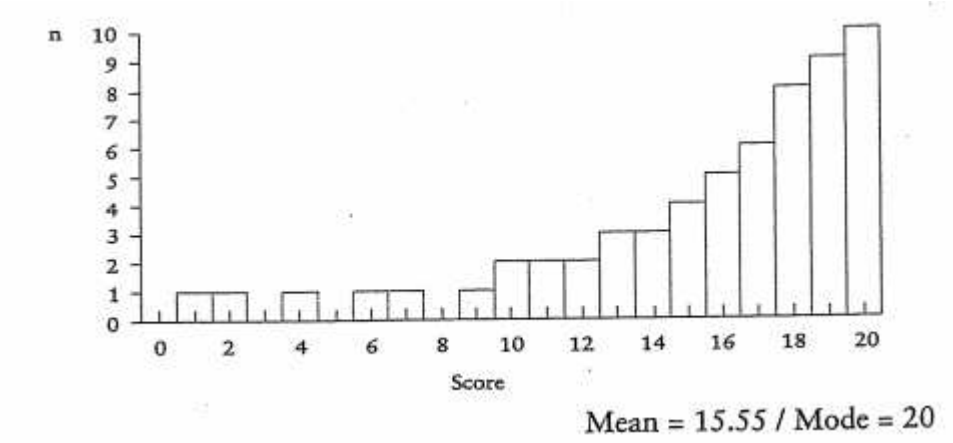
Mean is the term we use for average score. To calculate the mean we simply add up all of the individual results, get the total, and then divide it by the number of students taking the test. In (a) and (b), for example, more students got the mean score than any other score – 4 students scored 6 in (a) and 6 students scored 6 in (c). However, in (b) more students got a 7 than a 6. The score gained by the largest

number of students is called the *mode*, so in (b) mode is 7. It is useful to report both the mode and the mean, particularly when the test is very easy or very difficult, or when it appears that students of different levels of ability have taken the test.

The results of a test that was very easy for the students are seen in the next figure. The mean is 15.55, whereas the mode is 20. Such a distribution of scores is described as being "negatively skewed", because the scores tail off towards the left end of the graph. If a test is very difficult, and the scores tail off towards the right end of the graph, then the results are "positively skewed".

Figure 7

Negatively skewed distribution of scores (adapted from Alderson, Clapham & Wall, 1995, p.93)



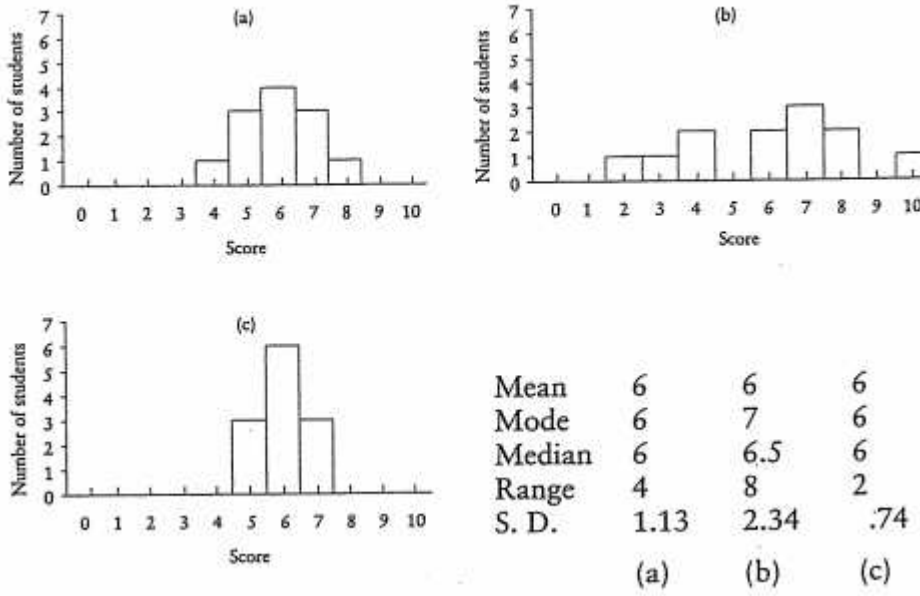
The third measure is the median. A *median* is simply the score that falls exactly in the middle, thus the score obtained by the student who is in the middle of the student rankings. To find the median, you don't have to do any actual maths. For example, if five students took the test and had score of 10,8,7,3 and 2, then the median would be 7. The median is especially informative when the tester thinks that the mean is in some way not representative of the level of ability of the whole group.

Once the mean, mode and median are reported, we get a good understanding of the differences in the distribution of scores. However, none of this measures accounts for the differences in the spread of the scores. For this reason we have to calculate *range* and *standard deviation*. For example, in Figure 6 (a) and (c) have

identical means, modes and medians, but it can be seen that (a) has a much wider spread of scores than (c). Reporting the range of each distribution is the most simple way to report this difference. The *range* refers to the difference between the highest and the lowest scores. So the range in (a) is 8-4 which is 4, and in (c) it is 2. It is clear that (c), with its range of 2, has a rather narrow spread of scores, whereas (b), with a range of 8, has a wide one.

Figure 6

Mean, mode, median, range, standard deviation (adapted from Alderson, Clapham & Wall, 1995, p. 94)



The range is a very useful measure of "dispersion", but the problem is that it does not take account of any gaps in the distribution. These gaps are scores which were achieved by no one. So in (b), where no students got a score of 5 or 9, the range is perhaps an overestimate of the spread of scores. The measure of dispersion which takes account of every single score is the standard deviation. The *standard deviation* (S.D.) is, approximately, the average deviation of each student's score from the mean. If a student has a score of 4, and the mean score is 6, than the student deviates -2 from the mean. Similarly a student with a score of 10 will deviate +4 from the mean.

Other statistics and graphs are used to explain the distribution of scores, but the five measures described above are sufficient for most of the purposes. With these statistics, it is possible to compare the difficulty level and the distribution of scores of different section of a test, or of different tests with each other.

2.4 Training of examiners and administrators

2.4.1 Training examiners

Examiners need to become familiar with the marking systems, thus the schemes and scales, that they are expected to use and they must learn how to apply them consistently. Even experienced examiners need constant updating and retraining. The term examiner indicates the person who is responsible for judging a candidate's performance in a test or examination. In the testing of speaking we distinguish between examiner and interlocutor. The former refers to the person who assesses the candidates, while the latter refers to a separate person who interacts with the candidate while the examiner assesses the performance of the candidate.

The training of examiners is a vital component of any testing programme because if the marking of a test is not valid and realiable then all the other work undertaken earlier to construct the test will have been a waste of time. Measurement, according to Mathews (1985) "implies a standardised instrument of assessment and an operative who can consistently apply it" (p. 90).

We distinguish between two types of marking: objective and subjective.

In objective marking the examiner compares the candidate's response to the response or responses the item writer has determined to be correct. The full set of acceptable answers may be called a "key" or a "mark scheme", depending on how much need there is for examiners to exercise their discretion in marking (Mathews, 1985, p. 101). The term *key* is usually used when there is only one correct answer for each item, while the term *mark scheme* is used when there is more than one potential answer for an item or when candidates are allowed to use their own wording to express the required concept. The main problem that arises

in some forms of objective marking is that item writers cannot foresee all of the responses that candidates might come up with to answer their items correctly. The goal of the training is to expand the mark scheme so that the examiners, who often do their marking alone and from their homes, will not have doubts about whether responses should be considered correct or not. In addition, examiners need to know what to do when they faced with an unpredected situation, in order to ensure that they do not act independently of each other and reach contradictory decisions.

Subjective marking is usually used for marking tests of writing or speaking. In the case of subjective marking, examiners are required to make more difficult judgements than a "rigth-wrong" decisions. Their job is to assess how well a candidate completes a given task and for this they need a rating scale. There are two types of scales. Examiners may be asked to rate a candidate's performance as a whole, in which case a *holistic scale* can be used. This type of scale is sometimes also called an impression scale, because examiners are asked no to pay too much attention to any particular aspect of the candidate's production, but rather to make quick judgement of its overall effectiveness. When examiners are asked to judge several components of a performance separately, this type of marking requires an *analytic scale*. In this case, descriptors are given for each component which are statements of the kind of behavior that each point on the scale refers to. In analytic marking a candidate's performance may be rated higher on one component than on another.

The choice between holistic and analytic scoring partly depends on the purpose of the testing. If diagnostic information is required, then analytic scoring is essential. The choice is also often affected by the circumstances of scoring. If scoring is carried out by a small group at a single site, then holistic scoring might be the best option, which is likely to be more time efficient. But if scoring is being conducted by a heterogeneous, possibly less trained group, or in a number of different places, analytic scoring is probably called for (Hughes, 1989). The main purpose of training programmes for examiners is to understand the principles

behind the particular rating scales they must work with and to be able to interpret their descriptors.

2.4.2 Training administrators

It is also of significant importance to train the administrators of a test. As Alderson, Clapham and Wall (1995) define it, administrators are those people who deliver the test to the candidates, and they are responsible for seeing that the circumstances under which the test is administered provide all candidates with the best chance possible to display the abilities being tested. Though the training of administrators need not to be as complex as that provided for examiners, it is still important that administrators understand the nature of the test they will be conducting, the importance of their own role and the potential consequences for candidates if the administration is carried out inadequatly.

The role of the administrator in speaking tests is especially important, as it is often necessary for at least on person to elicit language from the candidate and to respond in an encouraging manner in order to keep the language flowing. The role of the administrator becomes more complicated if two or more candidates are tested together. They have to make sure that everyone understands the task, to keep track of the number and types of contributions offered by each person, and to think about ways of getting into the conversation candidates who have not been able to speak earlier.

For certain speaking tests it may be necessary to employ another administrator to give instructions to the candidates and to provide them with the materials they need to study before entering the testing room. This person is sometimes reffered to as the "usher".

Another important task for the administrators of speaking tests is to create an environemnt that will help candidates to feel at ease. In the case of listening tests the choice of room is especially important, as is the decision about how many candidates should take the test at the same time. It is important for administrators to conduct trial listening tests to check whether the person who is speaking is

visible and audible from all parts of the room or whether recordings can be heard equally from all parts of the room, no matter where the candidate sits. It is also important to learn how equipment is to be set up, when and how it is to be used, and what to do if there is a malfunction.

However, there are other administrators, whose job is not so specialised: those responsible for distributing and collecting test papers, keeping track of timing and ensuring that the candidates cannot help one another during the test. They are often referred to as "inviligators" or "proctors". It is usually not necessary for them to undergo special training, but it is crucial that they understand all of their duties and what to do when unexpected problems occur.

2.5 Monitoring examiner reliability

There are many factors that can affect the ability of an examiner to give sound and consistent judgments, but an examiner should strive to be always reliable whether we talk about objective tests or about the marking of subjective skills, such as speaking and writing. In the case of objective tests, monitoring the marking simply means ensuring that the examiners have properly applied the marking key or mark scheme. The marking of subjective tests are more complicated than that.

2.5.1 Intra- and inter-rater reliability

There are two concepts we need to become familiar with: intra-rater and inter-rater reliability.

- Intra-rater reliability

An examiner is said to have intra-rater reliability "if he or she gives the same marks to the same set of scripts or oral performances on two different occasions" (Alderson, Clapham & Wall, 1995, p. 129). Even if some of the marks are different, the examiner can still be considered reliable, but not much variation can be allowed before the reliability becomes questionable. Intra-rater reliability of examiners are commonly measured by correlation coefficient. Examiner should agree with themselves marking the same performance on a different occasion. Intra-rater reliability can be established by getting examiners to re-mark scripts

they have already marked. In this case, examiners should not be told that they are re-marking scripts. It is then possible to check the correlation between first and second marks, their respective means and standard deviations and to take appropriate action if intra-rater reliability proves to be poor. Similarly, if the oral performances have been taped, procedures like this can be conducted for oral examining.

- Inter-rater reliability

The degree of similarity between different examiners is referred to as inter-rater reliability. It would not be realistic to expect all examiners to match one another all the time, however, it is important for each examiner to always try to match the standard. There are several operational definitions of inter-rater reliability representing different perspectives on what constitutes a reliable agreement between raters. There are three operational definitions of agreement (Saal, Downey & Lahey, 1980):

1. "Reliable raters agree with the "official" rating of a performance.
2. Reliable raters agree with each other about the exact ratings to be awarded.
3. Reliable raters agree about which performance is better and which is worse." (p. 413)

These combine with two operational definitions of behavior (Page & Petersen, 1995):

1. "Reliable raters are automatons, behaving like "rating machines". This category includes rating of essays by computer. This behavior can be evaluated by generalizability theory.
2. Reliable raters behave like independent witnesses. They demonstrate their independence by disagreeing slightly. This behavior can be evaluated by the Rasch model." (p. 561)

2.5.2 Monitoring techniques

There are many ways in which the marking of examiners can be monitored. The chosen method depends on a number factors, the most significant of which being

whether the marking is done centrally or elsewhere, and whether the marking is of written scripts or oral performances. Alderson, Clapham and Wall (1995) provide us with useful techniques to use in all of the mentioned cases.

- Central marking

If it is done centrally thare are at least three ways of monitoring.

The first method involves sampling by the Chief Examiner or Team Leader. Examiners are normally divided into teams when marking takes place centrally. There may only be one team if the test or examination is small, coordinated by the Chief Examiner. There may be several teams, each one coordinated by a Chief Leader, if there are more candidates. The Chief Examiner will standardise all of the team leaders and they will standardise the members of their teams. Each team will mark in its own area of the marking hall or in separate rooms allowing the Team Leader to monitor marking efficiently making it easier for the team members to discuss problems as they arise.

The second type of monitoring involves the use of realiability scripts. In this type each examiner marks the same packet of reliability scripts individually, chosen by the Chief Examiner and they will be marked by the Chief Examiner and the standardising committee.

This third type of monitoring involves routine double-marking for every part of the exam which requires a subjective judgement. This implies that two examiners mark a piece of writing each working separately. In this way, the mark that the candidate receives is the mean of the marks given by the two examiners.

- Marking carried out elsewhere

If the marking does not take place centrally, but rather in the examiners' homes or in an eximining centre, then we may have to modify the monitoring procedures.

Firstly, lets take the case of examiners marking at home. They may not be in a position to guarantee to mark a certain number of scripts a day and therefore it is not practical to expect them to be able to send in a sample of each day's marking. However, asking them to submit a sample from each batch of marking they are

requested to do would be realistic and practical. This enables the Team Leader to access scripts marked by the examiners under different conditions at different times during the day. It will be more representative if the Team Leader chooses the sample the examiners should send back, since if the examiners would choose they may send scripts which they have marked when their judgement was fresher or on which they spent more time. The Team Leader is responsible for interacting with the examiners as soon as possible, telling them that it is all right to go ahead or advising them about any issues they have.

Another procedure for monitoring examiner marking at home would involve all the examiners marking the same packet of relibaility scripts. This exercise might prove to be useful in uncovering those examiners, if any, who are having problems even when they know they need to mark carefully.

The third technique would be the routine of double-marking. The main difficulty in this kind of procedure is that it may be quite hard for separate examiners to discuss their differences of opinion in the cases where they are important enough to require attention. Although, in this case the Team Leader should be asked to read the scripts and make the final decision.

Another type of non-central marking takes place in individual testing centres and it mostly involves oral tests. This type of marking is very difficult since examiners have only a limited time in which to make their decisions and there is usually no way of reviewing a candidate's performance after the test to confirm or change their decison about the mark.

Nevertheless, there are a few monitoring procedures for oral tests. The most common one is sampling. Sampling is normally done by the Team Leader, who sits in on oral tests being conducted by the examiner. The Team Leader observes the performence and marks the candidate individually. After the test is finished, the Team Leader and the examiner compare their mark and discuss the differences of opinion that may occur. The taping of candidate performances would be a good possibility for institutions which test many candidate on different sites, because in this way they can be sampled or even double-marked by a Team Leader.

2.6 Reporting scores and setting pass marks

Once tests have been marked, it will be time to calculate some sort of score for each candidate. It is important to determine whether simply to add marks up for a total score for the test, or whether to give more importance to certain items than to others.

If the test comprises of a number of objective subtests, then each correct item may score 1 and each incorrect may score 0. Then these marks can be added together to arrive at a total for the whole test. In the case of subjectively marked tests, holistic or analytic ratings may be given for performance on the subtests or the whole test.

Test designers often assume that some items are more important than others and therefore those items should bear more weight. They believe that some aspects of language proficiency are more important than others in a given context. Giving extra value to certain items is known as *weighting*. An explanation for giving more value to certain components might be to emphasise to students the importance of particular parts of the curriculum. Items are typically weighted because they are considered to require more advanced proficiency or knowledge to accomplish, or to take more time to complete. It might be also important, to indicate the weighting of the components to help the candidates to allocate their time appropriately while taking the test. Equal weighting is the simplest method of weighting, it means giving the same mark to each item.

If each subtest is considered to be equally significant, despite their differences in length, then the subtest forms would need to be transformed before adding or comparing them. The most common method of transformation is converting subtest scores into percentage score. It involves dividing each subtest score by the number of items and multiplying it by 100.

The *reported score* is what is always of utmost significance in interpreting test scores. The reported score "is the score that is reported to candidates or employers or schools" (Alderson, Clapham & Wall, 1995, p. 152). In principle, after weigting and transforming the subtest scores it is then possible to report each

of the subtest scores separately, or to combine them in some way for decision and reporting purposes. The simplest procedure is to combine the scores by addition, and to decide a "pass" mark for the examination. In this approach, a candidate's good performance in one subtest can compensate for poor performance in another subtest. Several distinct cut-off points can be used to refine this pass/fail approach. For example, one score is the border between Pass and Fail, a second and higher score is the border between Pass and Credit, and the border between Credit and Distinction is a third score. In the case of UCLES' FCE and CPE candidates pass with a grade A, B or C and fail with a grade D or E.

However, it is also argued that reporting a single letter grade can be unfair to certain applicants because it does not offer proper recognition to their expertise in the component parts of the test or examination. Treating each component separately and reporting scores or grades on a profile is an alternative solution. The issue with this approach is that it lacks the real world requirement. Decision-makers commonly need only a single piece of data, not multiple pieces that require more complicated consideration (Alderson, Clapham & Wall, 1995).

Another approach would be to simply report scores, but not to stipulate a "pass" mark. In such cases, the repsonsibility for deciding whether a score is adequate or inadequate rests with the test score user.

It is important to make an overall decision about a candidate based on subtest scores. One could decide that a candidate would have to pass each subtest in order to pass the exam as a whole. One might allow a failure on one paper out of four or five or one might decide that if a candidate fails one paper, he or she would have to reach a compensatingly high mark to pass the exam. It should be remembered, however, that the idea of passing the test as a whole poses possible conceptual issues and can contribute to a great deal of arbitrariness. In a number of different ways, individuals can achieve the same overall score and thus be awarded a pass, although their profiles are different. Another problem is that one perfromance or score may have different values depending upon the purposes for which it is being used: what is sufficient for one purpose or for one population of

candidaes my be insufficient for another purpose or population. This is the reason why many test and exam results are reported on a scale, not as a pass or fail decision (Alderson, Clapham & Wall, 1995).

2.7 Validation

Probably the most important question in testing is whether the test tests what it is supposed to test. Since if a test does not measure what it is intended to do, then the scores are not reliable either - they do not mean what they believed to mean. The question is how can one know if a test is valid?

According to Henning (1987), validitiy refers to "the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure" (p. 89). Henning (1987, p. 89) also adds, that if the term "valid" is used to describe a test, it should be followed by the preposition "for", since any test may be valid for some purposes, but nor for others. One of the most common concerns with test use is test missuse, which is when a test is used for a reason for which it was not planned and therefore its validity is uncertain. It does not mean that a test cannot be valid for more than one purpose, but the validity of use for certain purposes need to be preestablished and demonstarated. Henning (1987) suggests that there are different degrees of validitiy, so tests can be more or less valid for their purposes.

Bachman and Palmer (1996) advise, that it is best to validate a test in as many ways as possible – the more different "types" of validity that can be established, the more evidence that can be gathered for any type of validity. Bachman (1990) emphasises that those different "types" of validity are different "methods" of assessing validity.

Most testers identify three types of validity: rational, empirical and construct. However, studies have shown that the rational/ empirical distinction is no longer valid. It is because both rational and empirical validation might include empirical data and content analyses of tests may include systematic studies of test content. Therefore, we shall use the terms "internal" and "external" validity, which is also referred to as "criterion validity". Internal validity relates to studies of the

percieved content of the test and its perceived impact, whereas external validity refers to studies comparing the test score of students with measures of their skills and ability gleaned from outside the test. Now lets take a closer look at both of them.

2.7.1 Internal validation

We distinguish between three common ways of assessing the internal validity of a test. These are: face validation, content validation and response validation.

- Face validation

According to Ingram (1977), face validity refers to the test's "surface credibility or public acceptability" (p. 18). Stevenson (1985) mentions that face validity is frequently dismissed by testers as being unscientific and irrelevant. Face validity requires and intuitive judgement on the content of the test by individuals whose judgement is not necessarily expert, like the judgement of administrators, non-expert test users or students. Typically, the judgement is holistic, relating to the test as a whole. Although attention may also be focused upon particular items, vague instructions and unrealisticly set time limits.

Since the advent of communicative language testing (CLT), there has been an increasing focus on face validity. Many advocates of CLT argue that it is important for a communicative language test to look like something you can do with language in the real world. Face validity is believed to be important in testing for two main reasons. One of them is that tests that do not seem to be valid to users may not be taken seriously for their given purpose. The other is that if a test is believed to be valid by the test takers they are more likely to perform well on that test.

- Content validation

According to Kerlinger (1973), "content validity is the representativeness or sampling adequacy of the content – the substance, the matter, the topics - of a measuring instrument" (p. 458). In contrast to face validation, content validation

involves the judgement of experts. In content validation one have to gather the judgement from people one is ready to believe.

One of the most common ways for content validation is to analyse the content of a test – which might be the test's specifications – and compare it with a statement of what the content ought to be. Another procedure for content validation would involve the creation of some data collection instrument. For example, a scale might be developed on which the test could be rated by the experts to the degree to which it met certain criteria.

Another alternative suggested by Alderson and Lukamani (1989) would be to provide the experts a list of skills that are supposedly being tested by certain set of test items and ask them to indicate the skill or skills that the items test. These decision could then be combined to get a sense of the level of agreement among the judges. Low content validity would be assigned to items on which there was no or little agreement.

- Response validation

Response validation involves the gathering of information on how individuals respond to test items. Studies have revealed interesing insights into test performance through learner-centred accounts. For example, introspection on a cloze task may reveal whether students have to respond to an item using the reading skills intended by the designer of the test, or whether only some knowledge of the grammatical structure of the phrase in which the item appears is required.

A similar example would be a reading comprehension task, where introspection may identify weak test items, that may produce cases where students choose the wrong answer although they understand the text, or get the answer right although they do not understand the text (Alderson, 1990).

The question is how should be this kind of introspective data gathered. The answer is: retrospectively. It means, that after the testees have taken a test, they can be interviewed about the reasons behind the answers they gave. The downside is

that candidates might be unable to recall why they answered in a particular way. This problem can be overcome by concurrent introspection, which means that candidates respond while taking the test, during period indicated by a silent observer. However, it is important that individuals who do not really take the test should be informants for this kind of validation, since if the test has important consequences for the testees it would be quite unreasonable to subject them to such an investigation.

2.7.2 External validation

We distinguish between tyo types of external validity: concurrent and predictive.

- Concurent validation

Concurrent validation involves the comparison of the test scores with some other measure for the same candidates, taken at about the same time as the test. The other measure may be scores from a parallel version of the same test or from another test, it can be the candidates' self-assessments, or teacher evaulations of the candidates.

The important thing with concurrent validation is that the external measure should be reliable and valid. This seems to be quite obvious and logical, but in actual pratice it is often hard to gather believable data. It is a common problem that no test which is known as valid and realibale is avaliable for concurrent validation. Yet if one still wishes to know how the test compares with other tests that are known and used in that particular context one should treat the results very cautiously.

Besides comparing test results, it proved to be useful to compare them with other measures of the students ability, such as the teachers' rankings. If the teachers have all worked with the same groups of students for a long time, they may have a clear understanding of their proficiency levels and will be able to rate them in order based on some aspect of their language ability. Self-assessment could be also used as the other measure to compare the test against. However, it

should be remembered that students are often not really accustomed to rating their own language ability.

- Predictive validation

Unlike concurrent validity, predective validation involves gatherering the external measures only some time after the test has been given and not at the same time as the administration of the experimental test. Predictive validation is most commonly used with proficiency tests. The most basic method of predictive validation is to give students a test and then give them another test of the ability the first test was supposed to predict at a later time in the future (Alderson, Clapham & Wall, 1995).

2.7.3 Construct validation

Ebel and Frisbie (1991), define construct validation as "the process of gathering evidence to support the contention that a given test indeed measures the psychological contruct the makers intend it to measure" (p. 108). The goal is to assure that the scores mean what we expect them to mean. Gronlund (1985) describes construct validation as measuring "how well test performance can be interpreted as a meaningful measure of some characteristic or quality" (p. 58).

Correlating the various test components with one another is one way to determine a test's construct validity. We should expect the correlations to be fairly low because the different test components assess something different and therefore add to the overall picture of language skill attempted by the test. If two components have a high correlation with each other, we might wonder if the two subtest are really testing different aspects or skills, or whether they are testing the same thing. The correlation between each subtest and the whole test could be expected to be higher, at least according to classical test theory, since the overall score is considered to be a more general measure of language skill than each individual score. Making theoretical predictions about the relationships among the compontents of the test in light of the underlying theory, and then comparing these

prediction with the correlation coefficients, is a slightly more refined version of constuct validation.

Another popular methods of constuct validation is to compare test results to biodata characteristics such as gender, age, number of years learning the language, first language etc. and to gather other information from the students at the time of the test. The aim of this kind of validation is to detect bias in a test for or against students identified by particular biodata characteristics.

A factor analysis approach is another choice for construct validation. Factor analysis reduces the complexity of a matrix of correlation coefficients, which is usually too complicated to grasp by a cursory study, and reduces the complexity of such a matrix to more understandable and managable proportions using statistical means. There are two main types of factor analysis: explaratory and confirmatory. In exploratory factor analysis, one actually examines the data to try to make sense of the factors that emerge, which is typically accomplished by determining which tests are most closely related to which factors and labelling the factors accordingly. In confirmatory factor analysis, the researcher predicts which tests or components would be related to which other and in what ways, and then tests the prediction's "goodness of fit" with the data (Alderson, Clapham & Wall, 1995).

2.8 Post-test reports

It is of great importance to write reports after a test has been administered. Test reports can have a number of different audiences and the features of the reports depend on these audiences. It's crucial for institutions to keep records of their decisions, procedures, test results analyses and the feedback they receive, as well as to pass information on to the audiences they consider relevant.

People employed within the institution, like those in charge of planning future versions of the test and coordinating related activities such as pretesting, administration, marking etc., are the most obvious audience. Teachers who trained this year's candidates and will be training other students to take the test in the future are another obvious audience. These individuals may need technical details

as well as summaries of how well their students performed and guidance on how to better prepare their next group of candidates. Other audiences that may need test details include administrators in other institutions who want to know whether to use the test or accept its results, as well as experts in language testing and other related fields who want to learn how the different testing bodies address the challenges of validity, reliability and practibality.

Hereinafter, we examine why post-test reports are of great importance for the two largest audiences, the institution itself and the teachers.

2.8.1 Post-test reports for the institution

Post-test reports written for the institution itself serve two functions. Firstly, it can be seen as a historical record of the test, showing how the test worked in practice. Secondly, it give guidance for future test development. It is not possible to collect all the information from a test especially if there is a large number of candidates. The results of each objective component, as well as the marks given for each subjective section, are the most important data to collect. It's also vital to gather the scores of all the markers who took part in the inter-rater reliability study.

Alderson, Clapham & Wall (1995), define the following analyses that should be reported:

1. "Descriptive statistics for the entire test as well as its individual components: mean, mode, median, range, and standard deviation.
2. Item analysis: facility value and discrimination index for each objective item.
3. Correlations among the components and between each component and the total minus that component.
4. Reliability of each objective section.
5. Reliability of marking of each subject section. " (p. 199)

Observations should be made during the administration of the test, as well as during the examiner preparation program and the marking sessions. When observing each form of operation, the observers should be given a list of features to look for, which should be written clearly on an observation instrument.

Feedback from administrators, applicants, and examiners should be obtained on a regular basis using questionnaires that inquire about particular aspects of the test, such as the consistency of the instructions. The report should provide summaries of this feedback as well as suggestions on how to develop the processes in the future.

The aim of reviewing candidates' scripts is to figure out what kinds of problems they had with particular items or tasks. The explanation for this is that it is not uncommon for an examination of candidate scripts to discover flaws in items or tasks that could have gone unnoticed by both test designers and moderators.

The institution might also be interested in gathering background information from all of the candidates in order to assess the performance of various groups of people. This form of comparison, which can be done by gender, religion, first language, age, and other factors, can sometimes show bias in specific test items or tasks.

2.8.2 Post-test reports for teachers

The teacher who trained students for the current version of the test and who will be preparing new students for future versions are the second obvious audience for a post-test report. These teachers are looking for summaries of the types of problems that candidates encountered, as well as suggestions on how to better prepare candidates in the future.

Statistical information about the test population and its performance on the test as a whole is not vital for teachers but it may be useful for the students to see how they fit in with the rest of the candidates and how their performance compares with the others'. These information include the number of the candidates taking the test, their characteristics, grade distribution, mean scores and standard deviations. The most common way of giving this type of information for the students is using tables with brief comments to help them to interpret what they are reading.

The marking key for objective items and especially the rating scales for subjective items are essential for teachers. It is not always clear to teachers how certain test questions should be answered and how should writing and speaking

tasks be marked. Since using past papers is a popular method of preparing students for exams, it is crucial for teachers to know if the answers suggested by their students would have been considered right by the testing body.

The testers' discussion of each component of the test should be one of the most important parts of the report. Firstly, testers should explain what was being tested in each segment. Then, they should note how candidates performed on each component and the types of problems that the population or certain part of the population found the most difficult. The testers should also give advice on the language and skills candidates should work on in the future, as well as on useful strategies they should learn to improve their performance.

Teachers should also be instructed about how to administer tests. When it comes to conducting tests, one of the most common problems is with listening tests. Bad sound equipment placement, poor acoustics in the testing room, and interference caused by noise in corridors or nearby rooms are all examples of these problems. Speaking tests can also be troublesome, particularly if more than one applicant is being evaluated at the same time. Some concerns occur as a result of inadequate teacher preparation: they may not have properly told the applicants about the protocols to be followed, or they may not have put them in groups or pairs that are compatible. The post-test report can be an important way for teachers to be reminded of these issues and what they need to do to avoid them in the future (Alderson, Clapham & Wall, 1995).

The second chapter of the master thesis presented the stages of test construction, relying mostly on the work of Alderson, Clapham and Wall (1995) on the topic. In this chapter, the steps of test creation was discussed, including test specification, item writing and moderation, pretesting and analysis, the training of examiners and administrators. It also provided information on how to monitor examiner reliability, how the scores are or should be reported and the past marks set and what to include in post-test reports depending on the audience.

PART 3

Research on the assessment practices of teachers in the English classroom in Transcarpathian Hungarian schools

In this part of the thesis a research is presented on the assessment practices of teachers in the English classroom in Transcarpathian Hungarian schools.

3.1. Research design and methodology

3.1.1 Planning the study

Language testing and assessment is a relatively new, but a rather significant field that is still being explored. It is therefore clear that it is not possible to examine all its aspects within one research.

The aim of the present study is to examine the assessment methods that are often and less frequently used in the English classroom and to explore the possible reasons behind them.

A further aim of the study is to identify possible deficiencies in language testing and assessment, and to draw attention to the importance of being assessment literate.

The research also aims to identify differences in the assessment practices of teachers teaching in lower and upper secondary forms and to collect data on exam preparation views and practices of the latter.

A mixed method study was carried out, the methodology used being both qualitative and quantitative.

The study involved an empirical investigation, the data collecting instrument of which was an online questionnaire.

The hypotheses being preoffered are that:

- teachers do not receive adequate training in assessment during their in-service years;
- teachers apply traditional assessment more frequently than alternative assessment;

- upper secondary school teachers apply alternative assessment methods more frequently than lower secondary school teachers;
- teachers consider "teaching to the test" bad for the students' overall language performance, nevertheless consciously prepare graduates for taking exams, mostly by practising common task types included in the exam.

3.1.2 Participants

3.1.2.1 Sampling

The target population of the research carried out were teachers and students who teach and study in schools with Hungarian language of instruction. Furthermore, the research focused on lower and upper secondary forms. The simple reason behind leaving out the elementary forms, which are forms 1-4, lies in the difficulty of collecting useful and reliable data from students of that young age.

However, only a part of the total population was approached for information on the topic. The sample was selected using *non-probability sampling*. A sample frame was created from all the Hungarian schools in Trancarpathia whose English teachers may have been potential participants of the research, so those who teach in forms 5-9 and 10-11.

Convenience sampling was used to select the teachers who would participate in the research from the created list which means that the participants were consecutively selected according to their convenient accessibility.

Snowball sampling was used in selection of student participants. Initially, the teachers were contacted, then they were asked to involve the students they teach in the part of research concerned with them.

3.1.2.2 Ethical considerations

The research was carried out completely anonymously, without the name of the respondent and the name of the institution where the respondent teaches or studies. Whether the subjects contacted participate in the research or not was entirely voluntary.

3.1.2.3 Participant information

The total number of teachers and students participating in the research is 348.

The number of teachers participating in the research is 30, of which 21 teach in lower secondary forms that are forms 5-9, and 9 teach in upper secondary forms that are forms 10-11. Table 4 presents the distribution of them by age and gender.

Table 4

Distribution of the teacher participants by age and gender

Broad age group, in years	Numbers		
	Total	Male	Female
less than 25	7	3	4
25-29	3	0	3
30-34	6	2	4
35-39	6	1	5
40-44	4	0	4
45-49	1	0	1
50-54	2	0	2
55-60	1	0	1

The total number of student participants is 318, 176 girls and 142 boys, aged 11-17. Table 5 shows the distribution of them by form and gender.

Table 5

Distribution of the student participants by form and gender

Forms	Numbers		
	Total	Male	Female
5.	86	39	47
6.	46	23	23
7.	30	13	17
8.	52	20	32
9.	48	22	26
10.	26	11	15
11.	31	15	16

Some other general information about the teacher participants:

- six of the responding teachers have BA (Bachelor of Arts) degree,
- 24 of the responding teachers have MA (Master of Arts) degree,
- the teaching experience of the respondents ranges from half a year to 35 years,
- one third of teachers have less than 5 years of teaching experience,
- only one teacher of them has a teaching experience of more than 30 years.

3.1.3 Research instrument

During the research, four online questionnaires were used: one for teachers teaching in lower secondary forms, one for teachers teaching in upper secondary forms, and two separate questionnaires for the students they teach accordingly.

The teacher questionnaires were in English, while the students completed the questionnaires designed for them in their native language to avoid any misunderstandings and reach a greater degree of reliability in responses.

- Content of the questionnaire of lower secondary school teachers

The questionnaire for teachers teaching in forms 5-9 conisted of two sections, containing a total of 13 questions.

The first section required general information from the respondents such as: gender, age, highest educational attainment, teaching experience, number of lessons, number of students in the grades in questions, and the training they got in assessing student learning.

In the second section, data on assessment beliefs and practices were collected, as: the opinion of the respondent teachers on what is the main purpose of assessment; the frequency of use of the different listed traditional and alternative assessment methods; the use self- and peer assessment; the use of norm- or criterion-referenced assessment, and the kind of feedback given to students.

- Content of the questionnaire of upper secondary school teachers

The questionnaire for upper secondary school teachers consisted of three sections, containing 16 questions.

The first two sections of the questionnaire corresponded to the content of the questionnaire for lower secondary school teachers.

The third section addressed the issue of "teaching to the test", asking respondents for their opinion on whether it may negatively affect students' general langugae performance. This section of the questionnaire paid particular attention to whether teachers consciously prepare students, within the framework of lessons, for the external independent testing, known as ZNO, which the students might choose to take at the end of form 11. Teachers were also asked about how regularly and in what way exam preparation takes place.

- Content of student questionnaires

The questionnaires for students contained fewer questions and was not divided into sections.

At the beginning of the questionnaire there were items that required some general information from the respondent students, such as: age, form, number of students in class and number of English lessons per week.

Then students also saw lists of traditional and alternative assessment methods, where they had to indicate how often they are used by their teachers during the English lesson and also the frequency of use of written and oral feedback they experience in the English lesson. Students were also asked about the kind of feedback they find most motivating and from whom.

Students in form 10-11 had to answer the same questions, with the difference that they also had to answer questions about ZNO preparation, as was done by the teachers who teach them.

3.1.4 Procedure

The research was planned, research questions were asked and hypotheses were set up in January, 2021. This was followed by the creation of the research instrument and considerations in how the collected data will by analyzed further on. The process of approaching potential participants and data collection took place in February and March, 2021. Subjects approached were informed of the purpose of

the research and were assured of the anonymity of their participation. Furthermore, they were asked to forward the questionnaires designed for students to the students they teach. The success of this process was facilitated by the platforms used during distance learning, where students had easy access to the research instrument. Data collection was completed by the end of March and its analysis and discussion began.

3.1.5 Data analysis methods

After the completion of the questionnaires, the preparation of the collected data for analysis began.

Online questionnaires reduce the time involved in administering and analysing data, since most of today's online questionnaire sites constantly analyze the responses received without our supervision. Nevertheless, the analysis of the responses to open-ended questions must be done like in the case of paper form questionnaires, so they were read through and catagorised.

The questionnaire used both open and closed questions. This is beneficial as it means both quantitative and qualitative data could be obtained. For several questions, in addition to the given choices, respondents had the opportunity to express their own opinions and ideas, or to justify their own responses in the "other" section. It proved to be a good solution to avoid unanswered questions, which is a quite common phenomenon when it comes to open-ended questions. Those who could not identify with any choice given were provided with the opportunity to write their own, while those who did not intend to do so could find an answer that suited them best.

Descriptive statistics, such as the mode was used as a measure of central tendency to indicate the most frequent responses, and bar charts were used to display the distribution of responses.

For close-ended questions, where respondents had the opportunity to check more than one option, frequency distribution tables were created to get an overview of the data that tell us how many times each response was selected.

In questions regarding the frequency of use of different assessment and feedback giving methods Likert scales were used to collect ordinal data. The frequency of use was classified into the following categories: never, rarely, sometimes, often, very often. Since these values have a natural order, they were coded into numerical values: 0 = Never, 1 = Rarely, 2 = Sometimes, 3 = Often, and 4 = Always. This method allowed us to establish an order in the frequency of use of the different assessment methods by the teachers.

Inferential statistics were used to investigate any association between the two different sets of data gathered from lower and upper secondary school teachers to highlight differences and draw inferences from the data in the case of particular questions.

3.2 Findings

3.2.1 Findings of teacher questionnaires

In the followings the data gathered from the questionnaires completed by teachers of lower and upper secondary forms are going to be described and analyzed.

- Training in assessment

To the question of what kind of training the respondent teachers received in the field of assessment, tests and measurement, 76.6% of the participants, which means 23 teachers, answered that they received training during their pre-service training at the teacher-training institutions where they studied. In addition, 16.6% – thus five of them, indicated that besides it they took a course in which assessment was also included among other topics. Four other teachers – 13.3%, said that they only took part in the aforementioned type of training. Only 10% of the respondents took part in a training specifically dedicated to assessment and testing. Furthermore, two teachers – 6.6%, indicated that they received no training in assessment at all.

Table 6 shows the frequency distribution of the responses of the thirty teachers participating in the research expressed in numbers.

Table 6

Frequency distribution of responses to the kind of training teachers recieved

Statements	Frequency
I received no training in assessment, tests, and measurement of student learning.	2
I received training in assessment, tests, and measurement of student learning during my pre-sevice training (at teacher-training colleges and/or universities)	23
Assessment, tests, and measurement were included in a course I took covering other topics.	10
I took a course dedicated to assessment, tests, and measurement of student learning.	3

- The main purpose of assessment

The next question in the questionnaire asked what teachers believe to be the main purpose of assessment. The table below shows the frequency distribution of the responses of the thirty teachers participating in the research expressed in numbers.

Table 7

Frequency distribution of responses to what is the main purpose of assessment

Statements	Frequency
to determine whether students have mastered the learning objectives	19
to determine student grades	10
to determine the effectiveness of my instruction	15
to make students accountable for their learning	11
to monitor students' learning progress	22
to motivate students	16

The most frequent response, indicated by 73.3% of the teachers, was that the main purpose of assessment is to monitor students' learning progress. Also a significant percentage of teachers – 63.3%, thought that the main objective of using assessment is to determine whether students have mastered the learning objectives. 53.3% of the participants found motivating students to be the main purpose of assessment. Half of the teachers responded that one of the main purposes of assessment is to determine the effectiveness of our instructions as teachers. 36.6% thought the main goal is to make students accountable for their learning and only 30% that assessment is to be used to determine student grades.

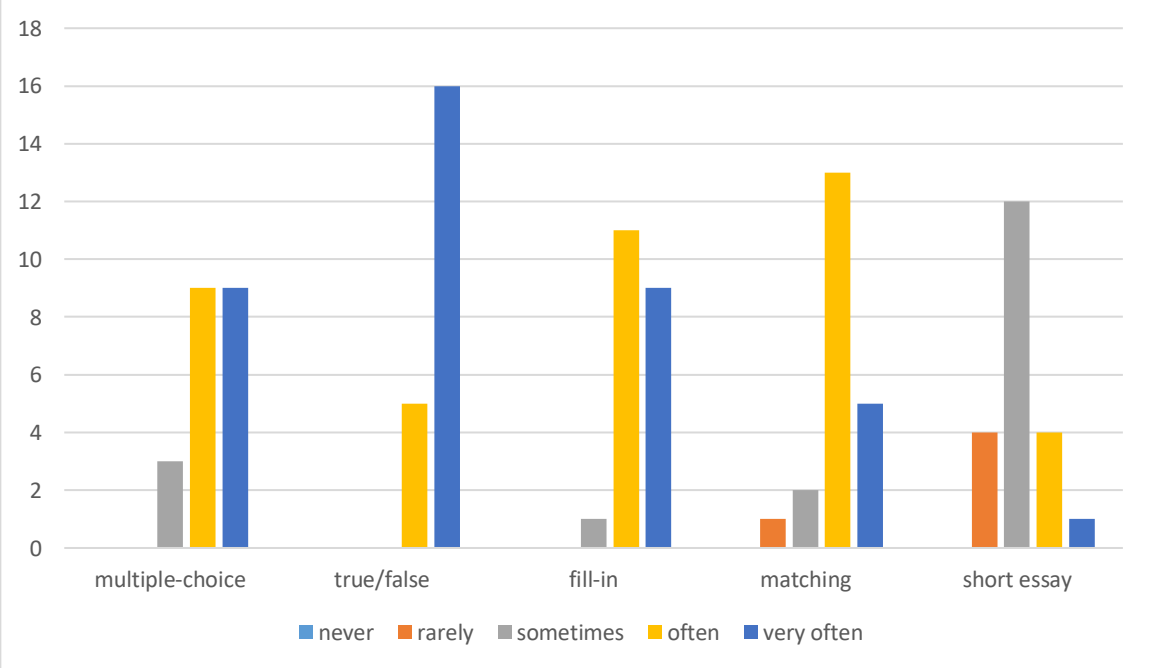
- Traditional assessment methods

In the following sections of the questionnaire, respondents had to indicate the frequency of use of different assessment methods and techniques. It is worth analyzing the responses of lower and upper secondary form teachers separately in order to be able to make correlations and draw conclusions from the findings later.

In the first item of this kind, the frequency of use of traditional testing techniques was indicated by the teachers.

Figure 8

The frequency of use of traditional assessment methods among teachers of lower secondary forms



The most frequent response received in this question was "often" – 40%, closely followed in numbers with "very often" – 38.1%, responses. The percentage of "sometimes" responses was 17.1 and of "rarely" was only 4.8. The fact that no teacher marked "never" for the frequency of use of any of the listed methods can not be overlooked.

It can be clearly seen from Figure 8, that the most commonly used traditional assessment method is the use of true/false questions, which are used very often by 76.2% of the teachers and often by the remaining 23.8%. The second most commonly used method among the respondents is the use of fill-in tasks, with 42.9% of the respondent teachers indicating the very common and 52.3% the common use of it, with only 4.8%, thus only one teacher indicating "sometimes".

According to the data collected, the frequency of using multiple-choice and matching tasks is almost the same. The same number of teachers, 42.8 – 42.8%, claimed to use multiple choice tasks often and very often during the lessons, and 14.4 % of them use them sometimes. Although only 23.8% of the respondents indicated very frequent use of matching tasks, it is not negligible that 61.9% use them often and just a few – 9.5% – sometimes, an only one teacher – 4.8% – rarely.

Assigning short essays proved to be the least commonly used method, with only 4.8% of the teachers, thus only one teacher indicating that he or she uses short essays very often and 19% often. Also, this is the assessment method that received the most "sometimes" – 57.2%, and "rarely" – 19%, responses.

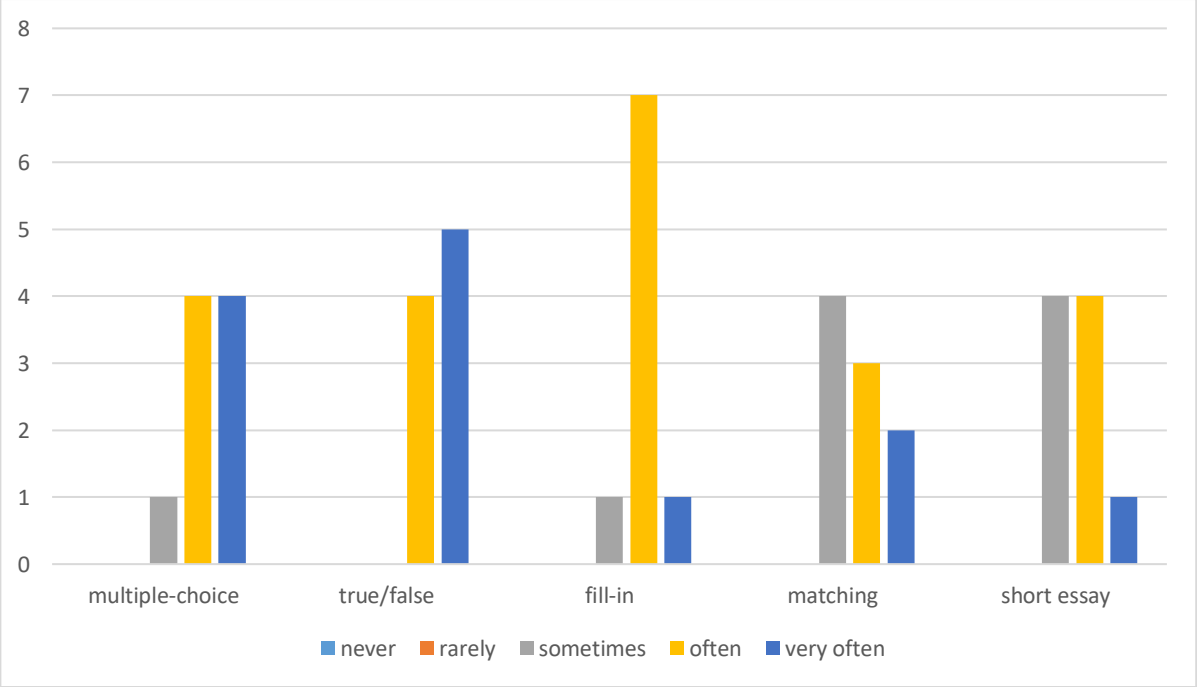
In the following, the responses of upper secondary English teachers will be described and analyzed.

The most frequent response received in the same question from upper secondary school teachers was "often", which add up to 48.9% of all the responses. It was followed by a markedly lower number of "very often" responses, with 28.9%. The percentage of "sometimes" responses was 22.2.

What is obvious from Figure 9, since we do not see either the light blue or the orange columns, is that the respondent teachers use all the traditional methods listed at least sometimes during the English lessons, as neither method received "never" or "rarely" responses.

Figure 9

The frequency of use of traditional assessment methods among teachers of the upper secondary forms



True/false questions proved to be the most commonly used items, just like among the lower secondary teachers, with 55.6% of the teachers indicating the very frequent use of this kind of assessment method and 44.4% the frequent use of it. The use of multiple-choice questions seems to be also a very popular form of assessment, as 44.4 – 44.4% of the respondents indicated that they use them often and very often and only 11.2% sometimes. However, also a significant percentage of the teachers – 77.9%, indicated that they use fill-in exercises often, 11.1% very often and 11.1% sometimes, to assess student learning.

There were quite varying responses to the frequency of use of matching tasks. 44.44% of the respondents marked "very often", 33.33% marked "often", and 22.22% "sometimes".

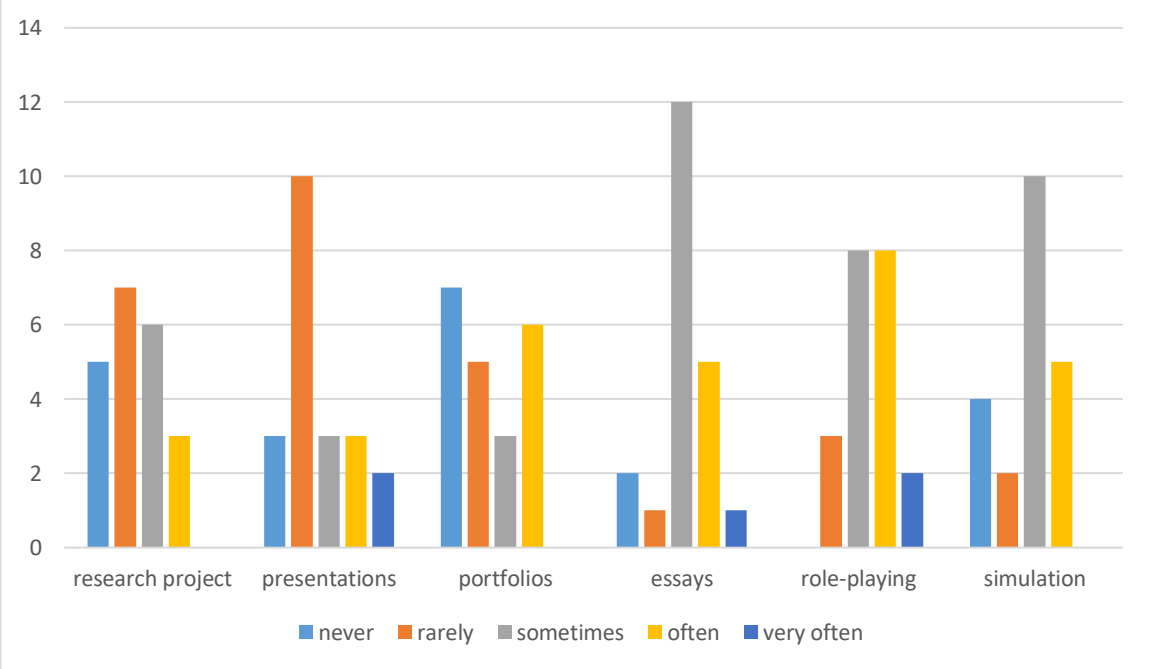
Just as in the case of lower secondary school teachers, here too the use of short essays appeared to be the least used method of assessment, although not with a big difference in numbers. The same number of teachers – 44.4 – 44.4% – indicated that they use them sometimes and often, and the remaining 11.2% indicated the very frequent use of them.

- Alternative assessment methods

In the following, the data gathered about the frequency of use of alternative assessment methods will be analyzed. As with the traditional assessment methods, data collected from lower and upper secondary school teachers are treated separately.

Figure 10

The frequency of use of alternative assessment methods among teachers of the lower secondary forms



The most frequent response received in this question was "sometimes" – 33.3%. This was followed in numbers by "often" responses – 23.8 %, then "rarely" – 22.2% and "never" – 16.7%. The least was the "very often" responses – only 4%. "Never" responses were present in a much greater extent than "very often".

The frequency of use of alternative methods by teachers can be seen to be more varied than in the case of traditional methods. The most popular alternative

method, to the use of which no one has indicated "never" and 38.1-38.1% of the respondents indicated that they use them sometimes and often, 9.5% very often and 14.3% rarely, is role-playing.

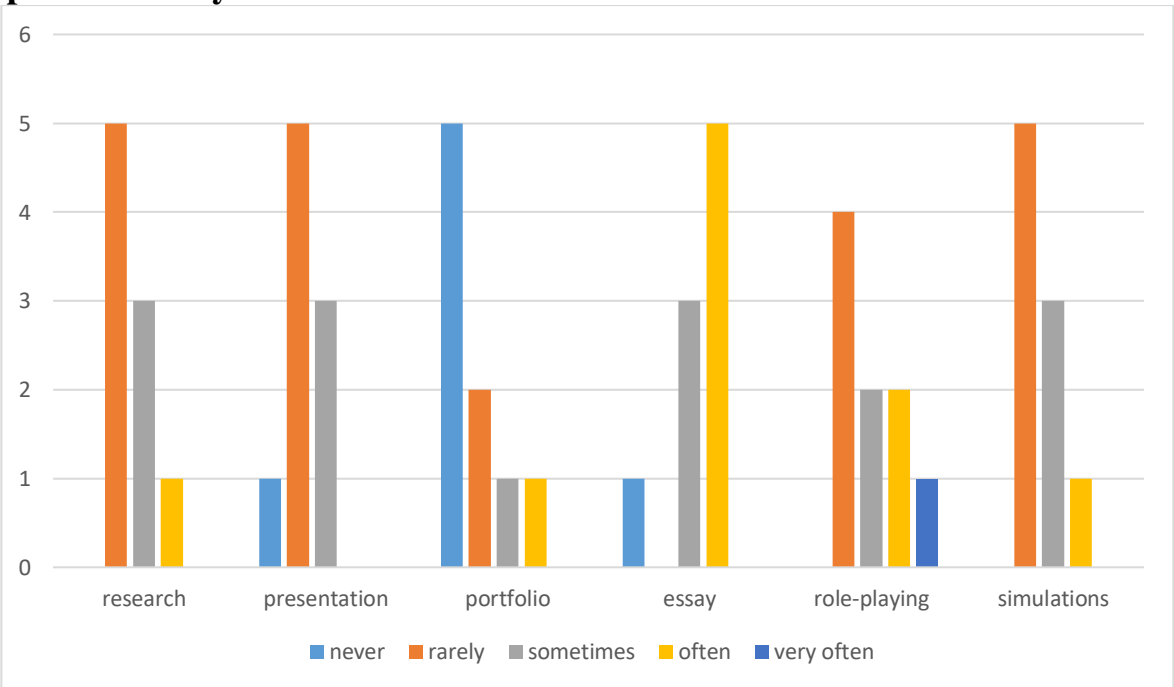
The next widely used alternative assessment method seems to be the use of longer essay questions. However, this is the item that received the most "sometimes" responses – 57%, therefore it is questionable whether this should be treated as a positive data or whether these answers can be interpreted as the respondent teachers' reluctance to give a specific positive or negative response, thus placing it in a neutral zone. The use of simulation in English lessons also received a large number of "sometimes" responses – 47.6%.

The frequency of use of portfolios to assess students produced quite varying results, with almost the same number of teachers indicating that they never use them – 33.3%, as those who indicated that they use them often – 28.6%.The least used method proved to be the use of research projects and presentations for assessment, which more than half of the respondents either never or rarely use.

Now let us examine how often these methods are used among the teachers of upper secondary forms.

Figure 11

The frequency of use of alternative assessment methods among teachers of the upper secondary forms



In Figure 11 we can see that the trend of less "very often" and more "never" and "rarely" responses continues here. "Rarely" responses make up 38.9% and "never" responses 13% of all the responses given. A significant percantage – 27.7% – of the responses is "sometimes", 18.5% is "often", while only 1.9% is "very often".

The most commonly used method here was the use of longer essays, with 55.5% of the respondent teachers indicating that they use them often, 33.3% that they use them sometimes and only 11.2% that he or she never assigns essays to students.

Only one alternative method of assessment that is the use of role-playing, was indicated to be used very often in the classroom, and only by one respondent. However, just like in the case of lower secondary school teachers, here it also proved to be a relatively widely used method, with 11.2% "very often", 22.2% "often", 22.2% "sometimes" and 44.4% "rarely" responses.

Simulations, research tasks and presentations are only rarely used by more than half of the respondent teachers – 55,5%.

Undoubtedly, the use of portfolios to assess students proved to be the least used method in the English classroom, with 55,5% "never" responses.

- Self- and peer assessment

In addition to the use of the traditional and alternative methods analyzed so far, teachers were also asked how often they use such alternative assessment practices as self- and peer-assessment.

Figure 12 shows the aggregate responses given by lower and uper secondary school teachers.

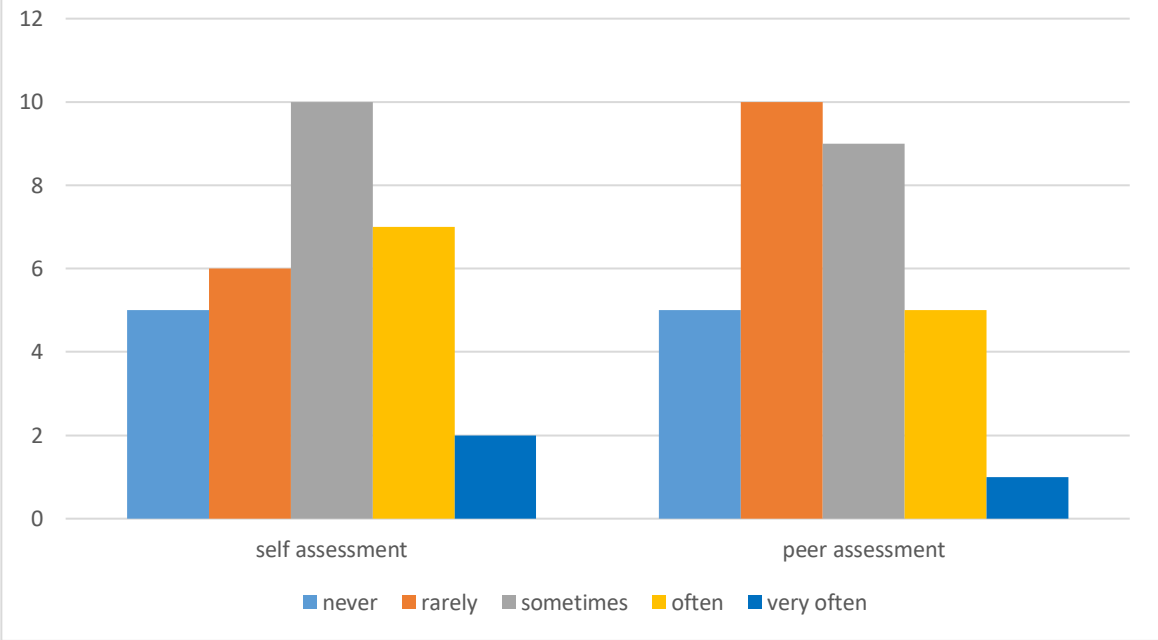
One can see that self-assessment gets more emphasis in the lessons than peer-assessment. Only one respondent indicated the very frequent and five the frequent use of peer assessment, while two teachers indicated that they use self-assessment very often and seven that they use it often.

Although it is immediately apparent that none of them are very common. The most frequent response was "sometimes", with a percentage of 31.7%. "Often"

and "very often" responses are present in a much lesser extent – 25%, than "rarely" or "never" – 43.3%.

Figure 12

The frequency of use of self- and peer-assessment



- Individual participation in whole class lessons

In addition, the respondent teachers were also asked about the frequency with which individual students participation is assessed during the lesson. Based on the responses, teachers commonly assess individual participation in the classroom. 20% of the respondents indicated that they use this kind of assessment practice very often and 43.3% that they use them often. Also, a significant number of teachers – 26.7%, responded that they sometimes assess individual performance in the classroom and only 6.7% indicated that rarely and 3.3% that they never apply this type of assessment.

- Feedback techniques

Besides the various methods and practices, feedback also plays a major role in assessment. For this reason, the questionnaires also inquired about how often the teachers use oral and written feedback, as well as about the use of a special kind of

feedback called "sandwich" feedback. "Sandwich" feedback means that between two praises the students also receive a constructive critique.

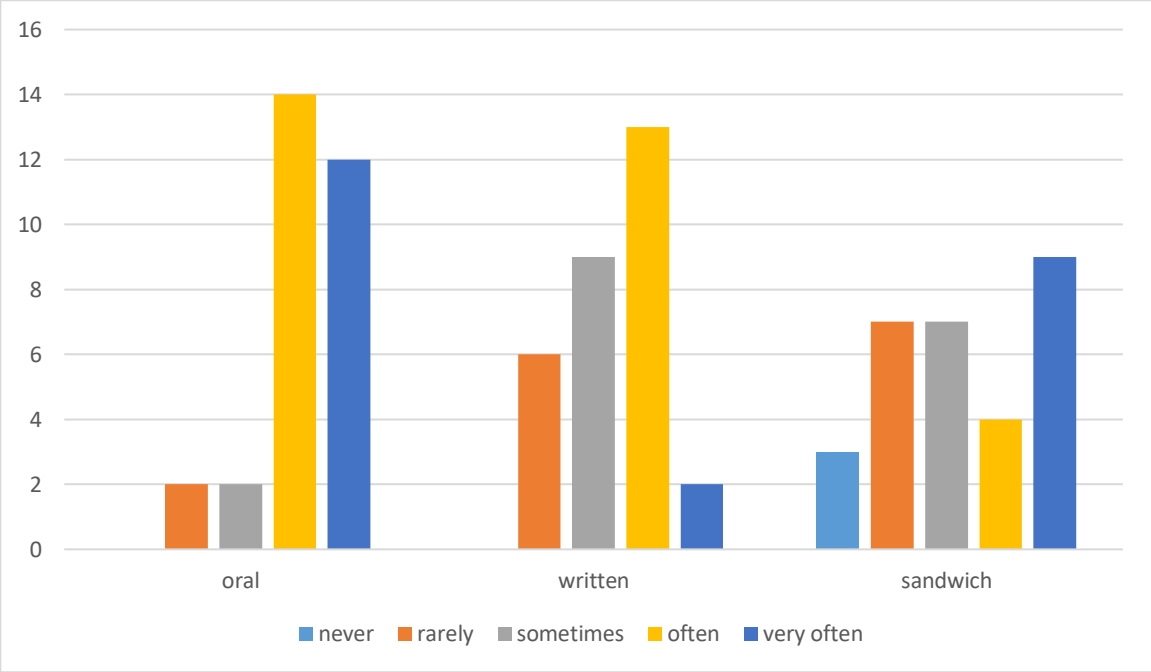
The most frequent response received to the frequency of use of oral and written feedback was "often" with 45%. The number of "very often" responses is also remarkable – 23.3%, while the number of "sometimes" and "rarely" responses is clearly less – 18.4% and 13.3%. This shows us that most of the teachers use feedback on a regular basis during the instruction.

There were quite varied responses to the frequeny of using sandwich feedback. 30% of the respondents claimed to use this technique very often, 13,3% often, 23.3-23.3% sometimes and rarely, and 10% never.

The next figure shows us the aggregate responses received from lower and upper secondary school teachers.

Figure 13

Frequency of use of feedback techniques



- Criterion- and norm-referenced assessment

The last question in the second section of the questionnaire was to find out how teacher interpret test scores, thus whether they use criterion-referenced assessment — assessing student progress compared to some predetemined standard, or norm-

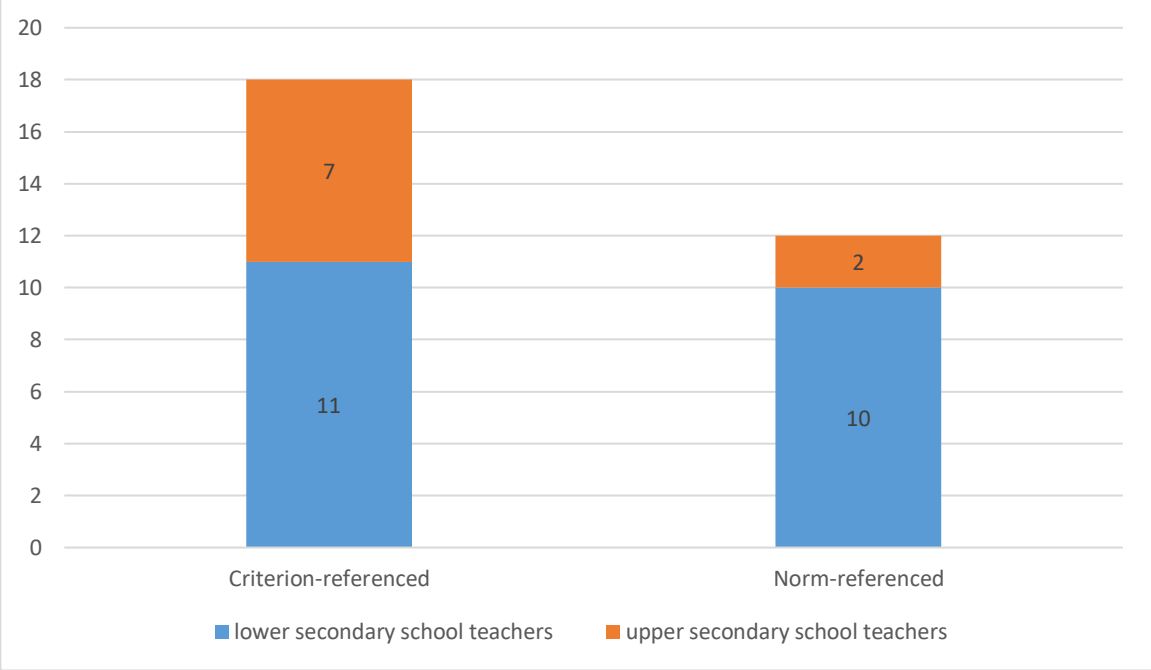
referenced assessment — comparing a student's results to someone else in their peer group.

60% of the respondents use criterion-referenced assessment, while 40% use norm-referenced assessment. It can be clearly observed from Figure 14 that the teachers of lower secondary forms split into two groups almost fifty-fifty, while the vast majority of the teachers of upper secondary forms prefer to use criterion-referenced assessment.

The aggregate responses received from all the teachers participating in the research can be seen in the following chart, indicating the number of responses given by lower and upper secondary school teachers.

Figure 14

Use of criterion- or norm-referenced assessment



- Teaching to the test

The data described and analyzed next only concern upper secondary teachers, as the third section of the questionnaire addressed to them dealt with the effect of "teaching to the test" and the way and methods of ZNO preparation taking place in the classroom.

The first question that the participants encountered was whether "teaching to the test" is bad for the students' overall language performance. They were also asked to justify their response in the other section. The teachers were quite divided on the issue, 44.4% saying that "teaching to the test" has a negative impact on students' language performance and 55.6% that it hasn't. Three out of the nine participants justified their opinion – two who claimed that the aforementioned practice has a negative effect and one who claimed it hasn't.

Justifications against "teaching to the test":

"When we teach for the test we focus only on topics needed for the test. Thus there is no emphasis on communication because the tests normally don't measure communicative competence. Teaching for test is like preparing learners to choose from answers, basically to push the right button. But unfortunately the main objective in 11th form is to pass ZNO. " (Teacher 6)

"In my opinion, it encourages students to learn only 'the most important' parts of the material which will ensure that they can pass their tests, after which they tend to forget most of what they've learnt. " (Teacher 2)

Justification in favor of "teaching to the test":

"Our responsibility as teachers is to prepare students for any hardship ahead whether it is a language exam or ZNO. " (Teacher 3)

- Integration of exam preparation into the English lesson

The next question in this section asked how and with what regularity teachers integrate ZNO preparation into their lessons.

Two of respondents plan the whole teaching in such a way as to prepare students for the exam. Similarly, two teachers incorporate exam preparation into the teaching by dealing with it one or two lessons a month. Three among the respondent teachers claim to prepare one lesson each week exclusively for the exam. Two of the participants added their own responses:

"I apply "teaching to the test" almost every lesson in the 11th form and a few times a week in the 10 th form." (Teacher 9)

"We prepare two lessons per week, especially during the second semester." (Teacher 6)

- Exam preparation practices

The last question focused on the practices used by teachers during exam preparation, where participants could choose more than one option and also provide their own answers.

It is apparent from Table 8 that 100% of upper secondary teachers prepare students for ZNO by practising task types that are included in the exam. Another widely used practice seems to be the practice of writing formal- and informal letters, and opinion essays, which we can really braket with the common task types that appear in ZNO. It can be presumed from the responses that preparation for the grammatical items the students may encounter while taking the exam, also plays and important role, while vocabulary seems to be the least important among the listed practices. Nevertheless, teaching of the required vocabulary is also practiced by two third of the respondents.

Table 8

Distribution of responses to what practices are used during exam preparation

Statements	Frequency
teaching grammar that is required by the exam	7
teaching vocabulary that is required by the exam	6
practising task types that are included in the exam	9
practising formal- and informal letter writing and opinion essay writing	8

3.2.2 Findings of student questionnaires

In the following the gathered data from the questionnaires completed by students of lower and upper secondary forms is going to be described and analyzed.

- Traditional assessment methods

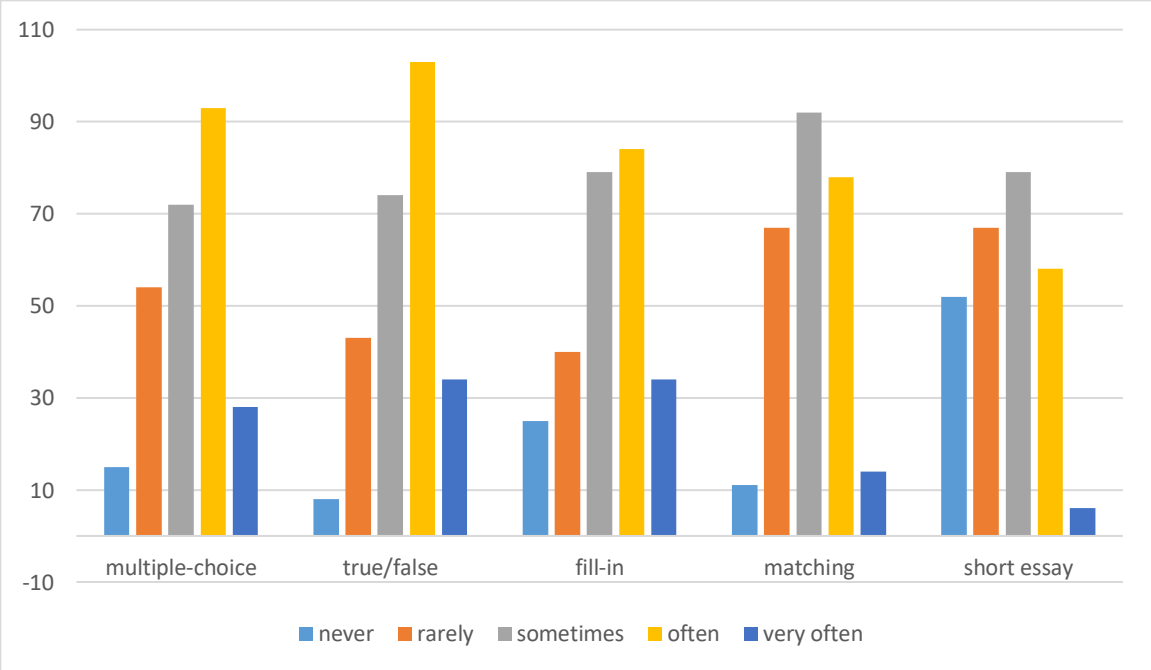
Once the respondent students answered the general background questions, the results of which were already presented in section 3.1.2, the students had to answer the questions regarding the frequency of use of different assessment methods during the English lesson applied by their teachers.

The first question of this kind asked students to indicate how often their teachers use the listed traditional assessment techniques. The overall response to this question was positive. The majority of the responses to this question was "often" – 31.8%. It was closely followed by "sometimes" responses with 30.2%. "Rarely" answers made up 20.7% of all the answers. The percentage of "never" and "very often" responses were roughly the same – 8.5% and 8.8%.

It is apparent from Figure 15 that true/false questions are widely used by the teachers according to their students, 39.3% of them indicating the frequent and 13% the very frequent use of this kind of question for assessment. 28.2% of the respondents also indicated that they are used sometimes in the classroom and only 19.5% that their teacher only rarely or never use true/false questions.

Figure 15

The frequency of use of traditional assessment methods by teachers based on responses from students of lower secondary forms



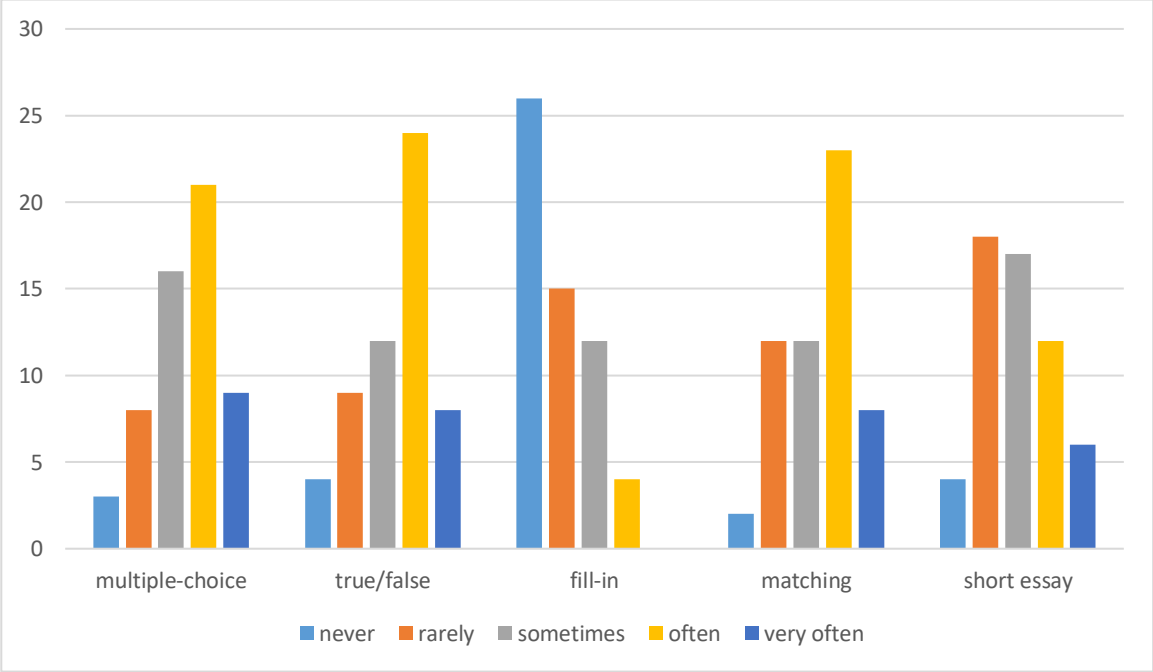
From the graph we can also see that multiple-choice and fill-in questions are also quite popular. Almost half of the students, multiple choice – 46.2% and fill-in – 45%, suggested that these kind of questions are often or very often used for assessment. Matching tasks are seem to be less commonly used, receiving the most "sometimes" – 3.51% and "rarely" – 25.6% responses.

From the presented data we can see that the least commonly used assessment method is the use of short essays, with the most "never" responses – 20.2%, with a significant percentage of "sometimes" – 30.1% and "rarely" – 25.6% responses and the least "very often" – 2.3% responses, among the listed methods. However, it should be noted that a considerable percentage of student respondents – 22.1%, indicated that they are often used by their teachers.

Figure 16 presents the responses of students of upper secondary forms to the same question. The figure indicates that the majority of responses to this question was "often", with 29.5% percentage of all the responses. "Sometimes" and "rarely" responses were present in almost the same extent – 24.2% and 21.75%. Only a small percentage of responses was "very often" – 10.9%, and there were slightly more, 13.7%, "never" responses.

Figure 16

The frequency of use of traditional assessment methods by teachers based on responses from students of upper secondary forms



The results show that multiple-choice, true/false and matching tasks occur with almost the same frequency. Just like in the case of lower secondary forms, students of upper secondary forms also suggest that true/false questions are commonly used. Over half of the participants indicated the frequent – 42.1% and very frequent – 14% use of them.

In descending order of frequency, the next is the use of multiple-choice tasks with 36.8% of the students indicating that they are used often and 15.8% that they are used very often to assess their knowledge. The percentage of "often" and "very often" answers are almost the same for matching tasks as for the multiple-choice ones, although, for matching tasks the number of "sometimes" and "rarely" responses is equal – 21% - 21% –, while for multiple-choice tasks the number of "sometimes" responses is more than that of "rarely" – 28% -14%.

Unlike in the case of lower forms, here the frequency of using short essays did not come last. However, the results suggest that in most classes it is either never or just rarely used – 35%, than frequently – 31.6%. Also, a significant percentage of the responses fall into the "sometimes" category, which is not very representative.

The least frequently used assessment method in the upper primary forms, according to the respondents, is the use of fill-in tasks, with almost half of the students – 45.6%, indicating that they are never used in the classroom. The number of "often" responses is also negligible – 7%, and no one indicated that they are used very often.

- Alternative assessment methods

In the following, the data collected from students about the frequency of use of alternative assessment methods will be analyzed. As with the traditional assessment methods, data collected from lower and upper secondary school students are treated separately.

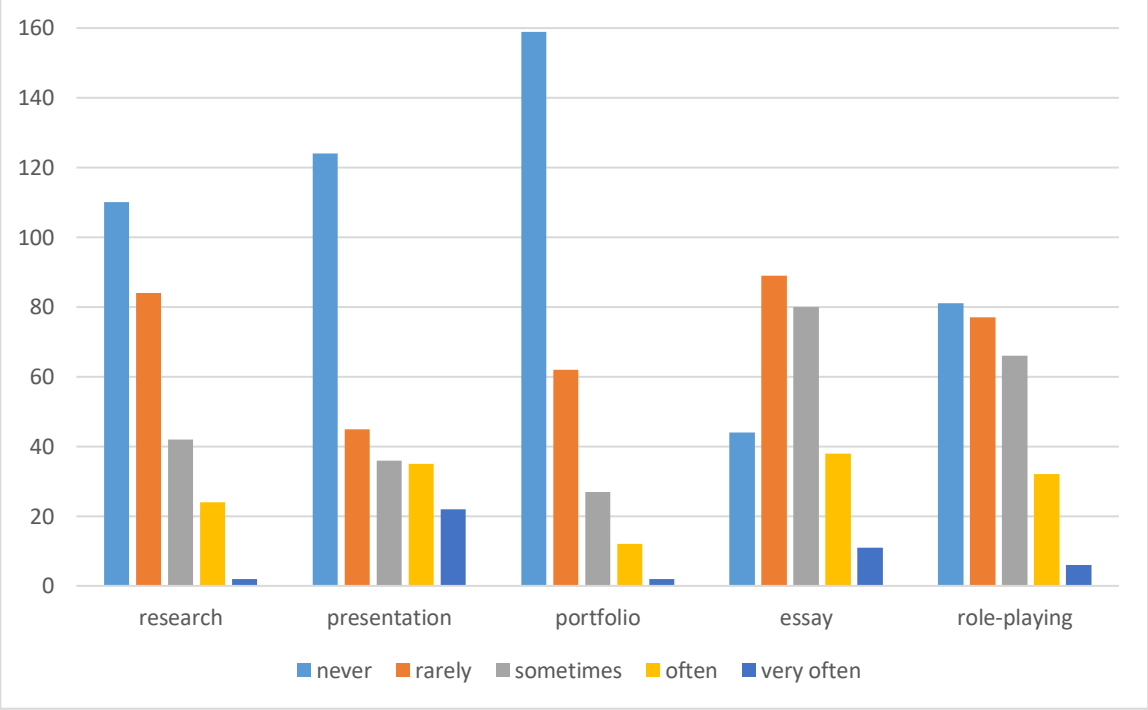
Figure 17 provides the responses of students of lower secondary forms. The overall response to this question was poor. The number of responses increases in a

negative way from "very often" to "never". All the responses received to this question are distributed as follows: never – 39.5%, rarely – 27.3%, sometimes – 19.1%, often – 10.8%, very often – 3.3%. It can be observed that for each kind of task, except for essays, most of the responses received was "never".

Based on the answers given by students, the most commonly used alternative method of assessment is the writing of longer essays. It is the method that received the most "often" and "very often" – 18.7%, and the fewest "never" responses – 16.8% among the listed alternative methods. In frequency of use, essays were followed by the use of role-playing. It was indicated to be frequently used by 14.5% of the respondents, sometimes by 25.2%, rarely by 29.4% and never by 30.9%.

Figure 17

The frequency of use of alternative assessment methods by teachers based on responses from students of lower secondary forms



Research projects and providing presentation are seem to be rarely applied by teachers based on the responses of the students, since nearly half of the respondents indicated that none of the aforementioned methods are ever used.

Portfolios proved to be the least used to assess students. The majority of the respondents, 60.7%, suggested that their teachers never use portfolios. 23.7%

indicated that they are rarely, 10.3% that sometimes, 4.6% that often and only 0.8%, which means two students, that portfolios are used very often.

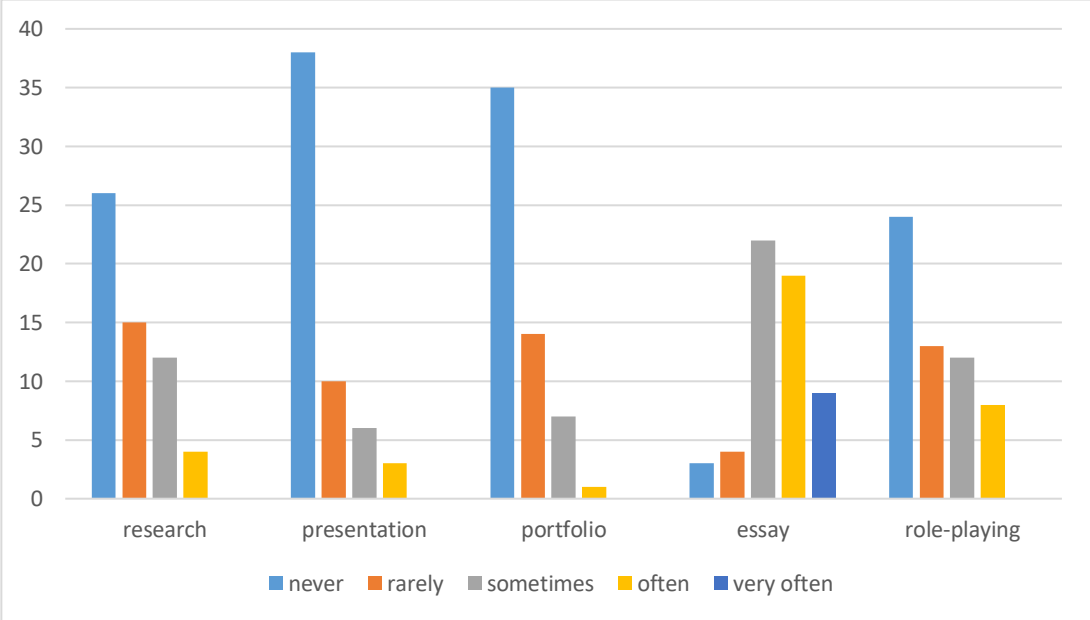
In the following, the responses of students of upper secondary forms to the same question are analyzed.

The responses presented in Figure 18 are quite similar to those of recieved from students of lower secondary forms. The trend of more "never" responses and less "often " and "very often " responses continues. Almost half of all the responses, 48%, to this question was "never".

However, there is a slight change compared to the data gathered from lower forms, as the number of responses does not entirely increase negatively from "never" to "very often". The percentage of "rarely" responses is 19.8%, of "sometimes" is 21.8%, of "often" 12.6% and of "very often" is 3.4%.

Figure 18

The frequency of use of traditional assessment methods by teachers based on responses from students of upper secondary forms



As can be seen from the chart above, the only kind of assessment method that received "very often" responses – 15.8%, and which received the most "often" responses – 33.3%, among the listed alternative methods, is essay writing. "Sometimes" responses made up the most of the responses to this type of task – 38.6%, while "rarely" and "never" responses the remaining 12.3%.

The next more commonly used alternative assessment method, according to the students, appears to be role-playing, with 14% of the respondents indicating the common use of it, 21% of the them suggesting that it is sometimes used to assess their knowledge and 22.8% of the students indicating that they are rarely used. After the essay, this kind of task received the least, though quite considerable number of "never" responses – 42.1%.

Research projects proved to be the third in the order of frequency, although only 7% of the respondents indicated that they are often and 21% that they are sometimes used. Presentation received the most "never" responses among the listed methods, with 66.6% of the respondents suggesting their teachers never ask them to provide any presentations. Even so, just as in the case of lower form students, the use of portfolios proved to be the least used based on student responses, with 61.4% of them indicating that they are never and 24.6% that portfolios are only rarely used.

- Self- and peer assessment

The next item of the student questionnaires aimed to inquire about the use of self- and peer-assessment in the classroom.

Figure 19

The frequency of use of self- and peer-assessment by teachers based on responses from students of lower and upper secondary forms

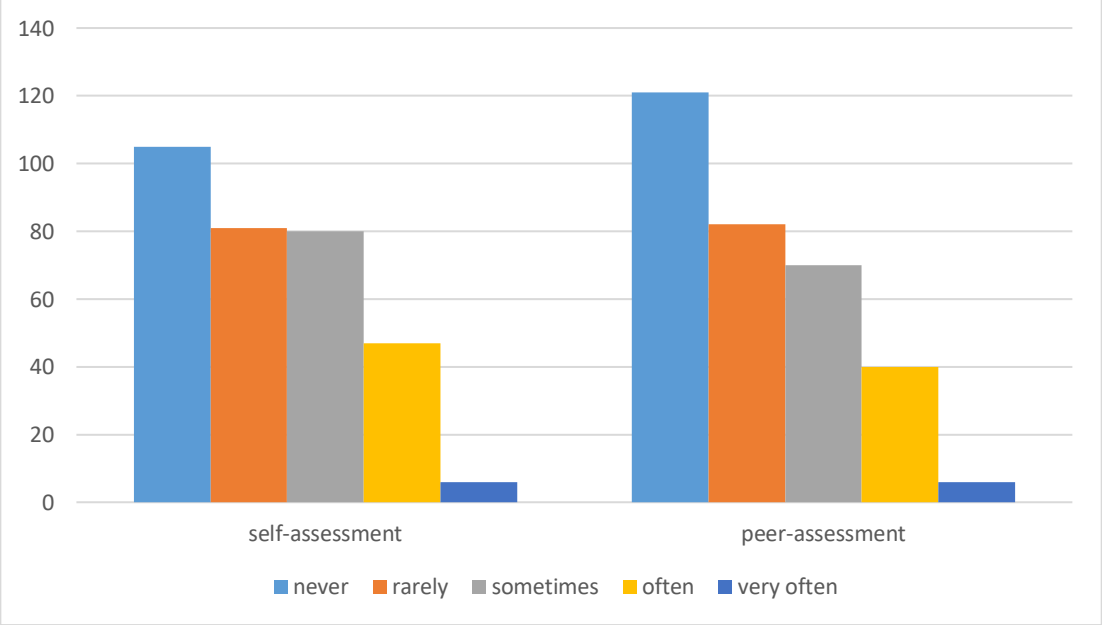


Figure 19 shows the aggregate responses given by students of both lower and upper secondary forms. As can be seen from the figure, self-assessment is more commonly used by teachers than peer-assessment, according to the responses recieved from students, although there is not much difference. What is apparent, is that "never" responses are present in a significant percentage in the responses recieved to both types of assessment.

"Never" repsonses made up 32.9% of all the responses to the frequency of use of self-assessment and 37.9% to peer-assessment. However, a not negligible number of students indicated that the mentioned practices are used by their teachers rarely, sometimes or often. In the case of self-assessment the percentage distribution of the remaining responses is the following: rarely – 25.4%, sometimes – 25.1%, often – 14.7%, never – 1.9 %. The percentage distribution of the remaining responses to peer-assessment: rarely – 25.7%, sometimes – 30 %, often – 12.5%, very often – 1.9 %.

- Individual participation in whole class lessons

In addition, the respondent students were also asked about the frequency with which individual student participation is assessed during the lesson.

Based on the responses received from them, teachers commonly assess individual participation during the English lesson. Nearly one-third, 32%, of the students indicated that their teachers often use this kind of assessment practice and 13.5% that it is used very often. Also, a signifcant number of students, 26.9%, suggested that assessment of individual participation is sometimes applied by the English teacher. Only a small number of students, 8.5%, indicated that their teacher never assess individual participation in the classroom, and a little more, though still a small number of them – 19.1%, that assessment of individual participation in whole class lessons is rarely applied.

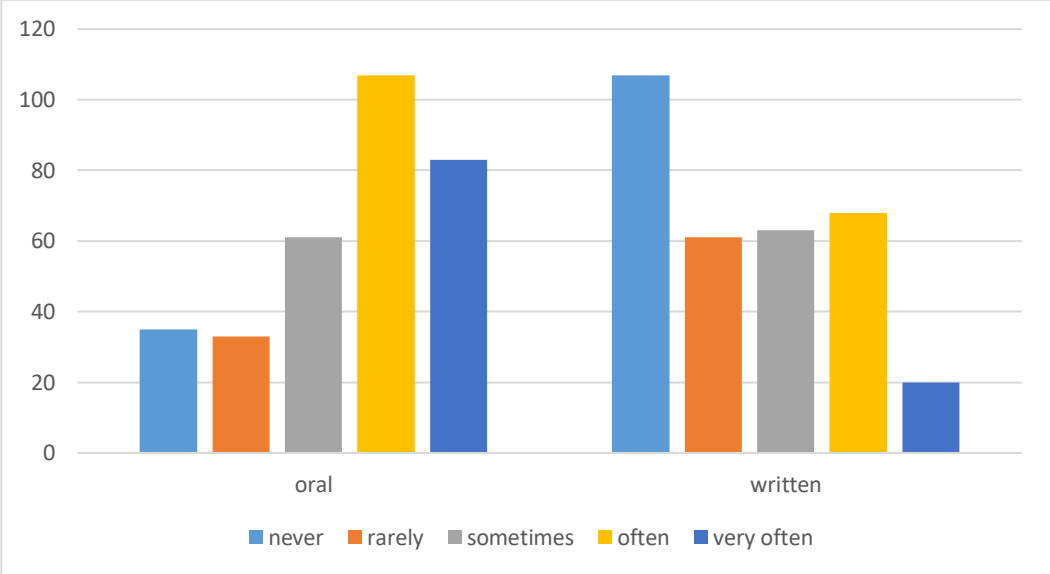
- Feedback techniques

In the next question of the student questionnaires the students were asked about how often the teachers use oral and written feedback.

From the data collected, it is clear that oral feedback is very common in the English classes. 33.5% of the respondents indicated the frequent and 26% the very frequent use of it, while only 11% of them suggested that their teachers never give them oral feedback. Also a rather low percentage of students, 10.4%, indicated the rare use of them and a slightly higher percentage, 19.1%, that they are sometimes given oral feedback from their teachers.

Figure 20

Frequency of use of feedback techniques by teachers based on responses from students of lower and upper secondary forms



However, written feedback is less common among teachers, according to the responses received from the students. As it can be seen in Figure 20, one third of the responses, 33.5%, to the frequency of use of written feedback was "never". While the percentage of "often" and "very often" responses is much less compared to oral feedback – 21.3% and 6.3%. The percentage of "rarely" and "sometimes" responses is almost the same – 19.1% and 19.8%.

- The most motivating feedback

The last question in the questionnaires that concerned all student participants asked them about the kind of feedback they find most motivating.

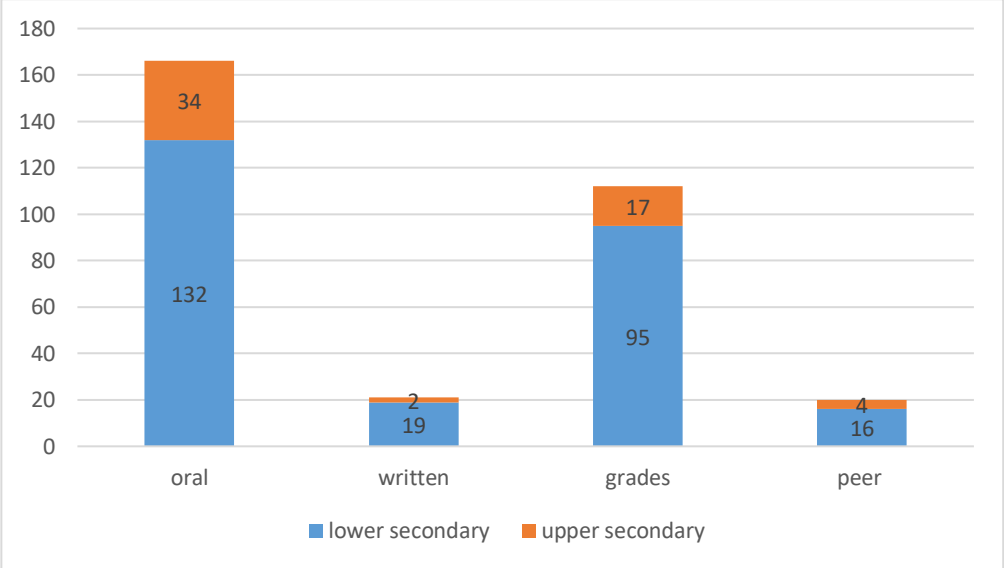
More than half of all the respondents, 52%, found oral feedback to be the most motivating. Significantly fewer, though a not negligible number of students

indicated that grades motivate them best in their learning – 35.1%. The remaining options to choose from were written feedback and feedback received from classmates that both received quite few responses – 6.6% and 6.3%.

The aggregate responses received from all the tudents participating in the research can be seen in the following figure, indicating the number of responses given by lower and upper secondary form students. No significant differences were found between the responses of the two sets of students.

Figure 21

The most motivating kind of feedback



- Teaching to the test

The data described and analyzed next only concern 11 form students, since the following questions dealt with exam preparation, more precisely preparation for ZNO.

The first question that the participants encountered asked them with what frequency they prepare for the examination. The students were provided with options to choose from, but they were also allowed to add their own answers in the "other" section, though no one did so.

The majority of the respondents, 58%, indicated that they prepare one lesson per week exclusively for the exam. Also, a significant percentage of respondents, 32.3%, indicated that they prepare for the ZNO during every English lesson. Only

6.5% suggested that preparation takes place only once a month and only 3.2%, thus one respondent, that there is no preparation for the exam at all.

- Exam preparation practices

The next and final question was concerned with the practices used during exam preparation, where participants could choose more than one option and also provide their own answers. Table 9 shows the frequency distribution of the responses.

Table 9

Distribution of responses from students to what practices are used during exam preparation

Statements	Frequency
teaching grammar that is required by the exam	17
teaching vocabulary that is required by the exam	15
practising task types that are included in the exam	26
practising formal- and informal letter writing and opinion essay writing	18

As it can be seen from the table, the most common exam preparation practice applied seems to be the practising of common task types included in the exam, since the vast majority of the students checked it – 83.9%. Learning the required grammar and vocabulary, as well as writing informal and formal letters and opinion essays received almost the same number of responses. It is also worth mentioning that a quarter of the responding graduate students, 25.8%, checked all the listed statements.

3.3 Discussion and interpretation of results of the research

After the description of the gathered data, this section will deal with the discussion and interpretation of the results.

- Training in assessment

One of the main objectives of the research conducted was to explore what training teachers received in assessment. The results show us that although the majority of the participants teachers received training in assessment during their pre-service years at colleges and universities, only a small number of them attended trainings where assessment was also included and even fewer in trainings that were specifically focused on assessment practices.

It is of utmost importance that teachers should regularly be trained in assessment methods. As it was stated at the beginning, the factors affecting successful language teaching and learning are constantly evolving, so it is important for teachers to update their knowledge in the field of assessment as well.

There are two possible explanations for the obtained results, one of which may be that teachers do not feel the need or are reluctant to participate in in-service training dedicated to assessment, while the other is that such trainings do not take place or just very rarely.

Therefore, it would be important to draw teachers' attention to the issue of testing and assessment, indicating that the incorrect and inadequate use of them may adversely affect the process of foreign language acquisition. Another important measure would be for the competent authorities and other stakeholders to see the need for teachers to be trained specifically in assessment and launch more comprehensive or assessment-focused trainings for in-service teachers.

- Purpose of assessment

On the question of what is seen by teachers as the main purpose of assessment, this study found that most teachers broadly agree on monitoring the learning process of students to be the main purpose of it. Also, a significant number of the teachers see determining whether students have mastered the learning objectives to be the main purpose of assessing them.

Only one-third of the teachers indicated determining students' grades to be one of the goals of assessment. However, it is interesting to draw a parallel here

with the question when students were asked about the kind of feedback they find most motivating. It is somewhat surprising that a fairly large number of students considered grades to be the most motivitaing kind of feedback, the only feedback outrunning it being oral feedback. It can therefore be assumed that, although determining grades may not be the main task of assessment, teachers must perform grading with sufficient objectivity and consistency, as we can see that they have a rather big impact on students' motivation.

- Frequency of use of traditional and alternative assessment methods

There is significant correlation between the frequency of use of traditional assessmet methods by lower and upper secondary teachers based on the responses recieved from teachers. For both sets of the respondent teachers, the most commonly used task type was true/false questions. Furthermore, the frequent use of true/false questions was also supported by the responses received from lower and upper secondary school students.

The responses of lower secondary school teachers and students are almost identical. However, in the responses received from upper secondary teachers and their students there is some discrepancy. While fill-in tasks seem to be relatively common based on the responses of the teachers, students placed it the last in the order of frequency established from their responses, suggesting that even short essays are more commonly used than fill-in tasks.

Table 10

Order of freqency of use of traditional assessment methods established from the responses of upper secondary school teachers and students

Upper secondary scool tachers	Upper secondary school students
1. True/false	1. True/false
2. Multiple-choice	2. Multiple-choice
3. Fill-in	3. Matching
4. Matching	4. Short essay
5. Short essay	5. Fill-in

The hypothesis preoffered that teachers tend to use traditional assessment more frequently than alternative was partially confirmed. This finding of the current study is consistent with the findings of Huseyin (2014). In his study he investigated Turkish teachers' preferences of assessment methods in the English as a foreign language (EFL) classroom. 120 EFL teachers completed the online self-report and the findings revealed that most of them rely on conventional methods of assessment. The findings of another study by Rezaee (2013) support this tendency. In his research he used a questionnaire to collect data on the views of 153 Iranian EFL teacher. His findigs revealed that traditional testing still seems to be the more commonly practiced approach, despite the reported advantages of alternative assessment and stigmatization of the traditional testing format by the teachers.

Based on the responses from both lower and upper secondary school teachers and their students, some types of alternative assessment methods are never or only rarely used. This result may be explained by a number of different factors. The use of traditional assessment is overall more simple, time-saving and straightforward. By using traditional assessment methods teachers are able to deal with more students in less amount of time and this kind of assessment is also more reliable and fixed. Another explanation for its popularity is that it allows teachers to compare the results of various students.

The less frequent use of alternative techniques may also be explained by a number of reasons, including that they are harder to evaluate for teachers. Teachers must put in more efforts to understand students work that also demand more time, thus this method is highly time-consuming. Furthermore, the use of alternative assessment can often lead to misundertanding and unfairness in evaluation.

In the case of alternative assessmet methods, the frequency of use order established from the responses of lower secondary teachers is almost completely different from that of upper secondary teachers. There is some similarity in that the two most commonly used tasks are role-playing and longer essay writing in both cases, although while role-playing is in the first place among lower secondary teachers, among upper secondary school teachers it is longer essay writing. The

rest of the tasks are used with different regularity by the two sets of teachers. The next table presents the alternative assessment methods in descending order of frequency.

Table 11

Order of freqency of use of alternative assessment methods established from the responses of lower and upper secondary school teachers

Lower secondary	Upper secondary
1. Role-playing	1. Longer essays
2. Longer essays	2. Role-playing
3. Portfolios	3. Reserach projects
4. Research projects	4. Presentations
5. Presentations	5. Portfolios

The responses given by the students of lower secondary forms do not support the order established from the responses of their teachers. What is noteworthy here is that a considerably large number of the respondent teachers indicated that they sometimes or even often use portfolios, while it was the least used kind of assessment on the list established from their students' responses.

On the other hand, the frequeny of use order established from upper secondary school teacher responses is completely in sync with that of their students. In the frequency order of use longer essays are the first followed by role-playing, then research projects and presentations and the least used method proved to be the use of portfolios.

There might be a number of reasons why portfolios are so underused, one of the most obvious of which is that they are time-consuming, since teachers have to organize and evaluate its contents and doing it all besides traditional testing makes it even harder. Moreover, since portfolios provide qualitative data, they can be difficult to anaylize and can be seen less reliable or fair compared to quantitative evaluations such as test scores. Unfortunately, these disadvantages are true for almost all alternative assessment methods, as mentioned earlier. However, a number of studies carried out on the impact of porfolios suggest that they have

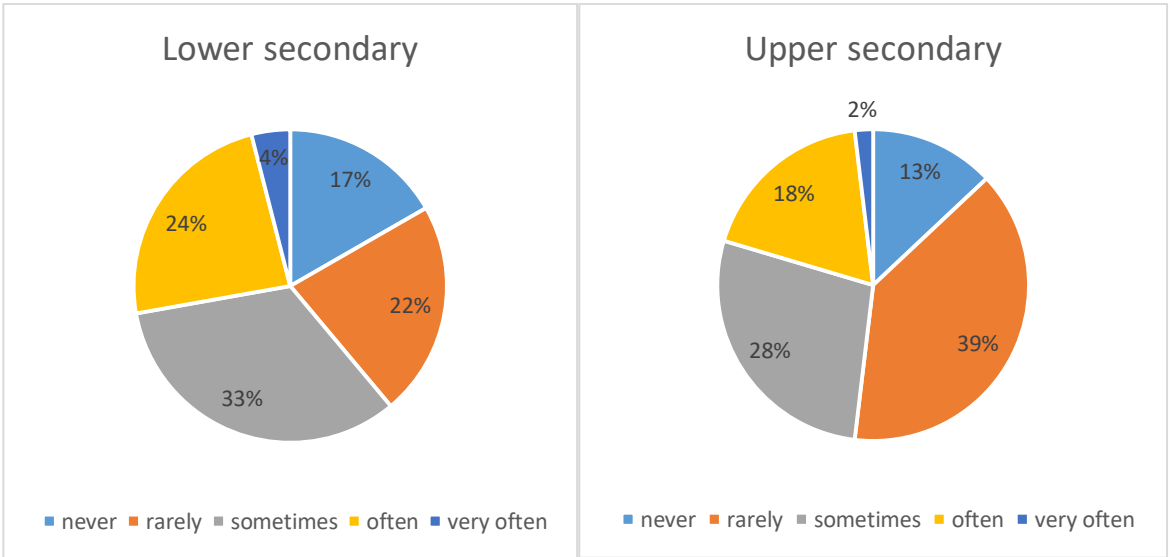
significant effects on the students' performance in writing skill (Kalra, Sundarajun & Komintaracat, 2017), therefore their use is widely suggested.

Wali Khan Monib (2020) carried out a study to highlight the definition, characteristics and effects of alternative assessment in EFL context by reviewing current research on assessment. In his study he concluded that the majority of the articles reported positive attitudes and effects of applying alternative assessment. All of the methods of alternative assessment reviewed, the majority of which focused on portfolios and self/peer assessment, proved to be helpful in enhancing students' development in foregin language learning. In another study that was carried out by Letina (2014), she concluded that though teachers find alternative assessment useful and they see the limitation of traditional methods, there is a contradiction between the teachers' positive opinion and the rare frequency of its application in teaching.

Contrary to expectations, the present study did not find a significant difference between the frequency of use of alternative assessment methods among lower and upper secondary school teachers. The following pie charts show the distribution of responses to the frequency of use of alternative assessment methods among lower and upper secondary school teachers.

Figure 22

Distribution of responses to the frequency of use of alternative assessment methods among lower and upper secondary school teachers



One difference to mention might be that while portfolios seem to be more widely used among lower secondary school teachers, at least according to their responses, it is not as widely used in upper secondary forms. The same is true for the use of longer essays - while the teachers of upper secondary forms use them more often, the teachers of lower forms use them with less frequency.

These results and the preasumed reasons behind them make it important for teachers, despite all the disadvantages, to learn about the alternative assessment methods and their positive effects that outweight the mentioned disadvantages and for teacher educating institutions and other competent authorities to encourage teachers to the use of them. The use of alternative assessment methods help in recognizing students' unique abilites and help in real-life application of their knowledge. Furthermore, they develop extensive cognitive skills, while traditional assessment tend to be more theory-based and often to promote an unhealthy learning atmosphere.

- Self- and peer-assessment

The research further inquired about the use of alternative assessment methods such as self- and peer-assessment. We got a more negative picture of their use from students' responses than we do from teachers' responses. Just a few of the participant teachers indicated that they never use either method for assessment, while in the responses of the students for both mentioned methods we notice that the majority of students indicated that neither of them are ever used.

This result is also consistent with that of Huseyin (2014) who found that self- and peer-assessment are the least used assessment methods in Iranian EFL classrooms. Another finding is that the responses of both teachers and students indicate that self-assessment is used more often than peer assessment. The more frequent use of self-assessment than peer-assessment is also supported by the findings of Huseyin (2014).

It is difficult to explain this result, but the less frequent use of peer-assessment may have something to do with the fact that peer pressure, friendship

or just the opposite - hateful feelings, may affect the reliability of the grades given by the students. Self-assessment carries slightly fewer pitfalls, although it can be also subjective since students might not be honest and may over-evaluate their perfromance or just the opposite, they might also under-evaluate themselves.

Another important finding was that individual participation is often assessed during whole class lessons. One of the issues that emerges from this is that whether this is fair to students with different learning styles. In EFL context, there is less research on the relationship between learners' learning styles and their level and form of classroom participation and whether the assessment of individual participation is fair to students of various learning styles. This gap was attempted to be filled by a study conducted by Crosthwaite, Bailey and Meeker (2015). The result of their study suggested that a wide range of learning styles may be found even in mono-cultural classes. Learners participating in their research with individualistic learning styles generally achieved lower proficiency test and participation scores than learners with styles that suit in-class interaction, thus those who are happy to speak and pro-actively participate in classroom discussions. Their findings suggest that assessing individual participation in whole class lessons may be both ineffective and unfair for those with certain learning styles, as it is is often paried with the ideas of requiring students to speak in class, answer questions, make comments and join in discussion. That is the reason why this type of assessment is often discriminatory against students who are reluctant to speak in the classroom and their learning style involves more active listening.

The results of a study (Azarnoosh, 2013) carried out on attitudes and bias in peer assessment in EFL context revealed no significant difference between learners' peer assessment and teachers' assessment. Azarnoosh did not find firendship bias during the practice of peer assessment and experienced a positive change in the attitudes of students toward peer assessment.

The importance of self-assessment is presented to us in a report on a specific program designed to develop students' independence in foreign language learning. This program is called "Depends on Me" and was designed by a Hungarian teacher

Szénásiné (2017) and other fellow teachers in 2002. "Depends on Me" is a learning methodology program designed to teach students how to learn a foreign language, to promote their independence in order to increase the effectiveness of language learning and to prepare them for lifelong learning within the framework of school language teaching. As Szénásiné (2017) stated, their main goal was to develop a program which is feasible within school settings, without organizing separate courses, to help the students. Their methods included group discussions at the end of lessons, before and after assessments, written self-assessment of students and keeping student diary. Their results achieved in the field of self-assessment and self-monitoring showed that students became more realistic in their assessment of their own oral and written performance.

- Feedback

Feedback is an important part of the assessment process. Therefore, the research paid special attention to the frequency of using different feedback techniques and the kind of feedback students find the most motivating.

Responses from both teachers and students show that oral feedback is used much more often than written feedback. This finding is not surprising, since oral feedback is an important part of verbal interaction between teachers and students. This is well supported by the fact that the vast majority of students consider oral feedback to be the most motivating kind of feedback.

Although oral feedback is mainly considered to happen between teachers and students, a great amount of oral feedback also comes from peers, thus classmates. However, feedback from peers was considered by only a small number of students to be the most motivating. The same is true for written feedback, students don't really find them motivating, but the received grades are all the more so. One of the issues that emerge from this finding is that students can easily lose sight of the purpose of education in their pursuit of good grades. They might become obsessed with their grades and they might think of them as a measure of their ability or even self-worth. This is why it is important that teachers

should assign grades along with or followed by some positive or constructive kind of oral or written feedback, so students can see their strengths and also identify weaknesses without the feeling that their worth and intellect is defined by a single number.

Another possible solution would be the combination of both positive and contructive feedback, for which we have got the "sandwich" feedback, also known as balanced feedback, to use. The use of this feedback giving technique also varies greatly among respondents. There might be several possible explanations for this result. One of the reason for not using this technique might be that they simply have not heard of it so far. Another possible explanation is that teachers want their feedback to be clear and draw attention to the areas for further development without hiding it between praises, in order to prevent students from only hearing the good. The exact opposite of the aforementioned reason might be also possible. Teachers may choose not to use this technique to avoid positive feedback losing its power after the "buts" and "howevers" we use to introduce the criticism.

Although one can find many publications on the characteritics of "sandwich" feedback with its preassumed negative and positive properties, empirical researches on the topic is not so common, espicially not in EFL context.

An article published by Parkes, Abercrombie and McCarty (2013) describes studies conducted by them on feedback sandwich. Their studies meant to inquire about the opinion of students about feedback sandwich and to find out whether the applied technique has led to improved performance. Although the students reported back the positive impact of the received feedback sandwich on their next assignment, their performance did not mirror it. This is an important issue for further research.

- Criterion- and norm-referenced assessment

The current study also found that more teachers tend to use criterion-referenced assessment than norm-referenced. However, contrary to expectations, there is no

big difference in the number of teachers indicating the use of criterion- and norm-referenced assessment.

On the other hand, there is quite a difference between lower and upper secondary school teachers on this question. Among the teachers of lower secondary forms, the use of both assessment methods were indicated by approximately the same number of teachers, while the vast majority of teachers of upper secondary forms clearly use criterion-referenced assessment.

There are several possible explanations why criterion-referenced assessment is more widely used. One of the most obvious of which is that in criterion-referenced assessment students are only compared to themselves, it doesn't pay attention to the performance of the other students and thus students have a better chance of scoring high and it can help students to improve their self-esteem and feel better about themselves. However, criterion-referenced assessment has also got its drawbacks, as norm-referenced assessment has also got its advantages.

We get mix opinions from literature dealing with the use of both of the assmessment forms. Some scholars suggest that the grades given to students in language tests should be based on a mix of the two assessment forms, indicating that they are complemetary to each other. However, some of the scholars suggest that criterion-referenced assessment should be seen as the primary and dominant principle. Others argue that norm-referenced assessment would not be appropriate for classroom evaluations and criterion-referenced assessment would be insufficient for outside the classroom evaluations. To sum up, both have their own places in education and one cannot decide which one is the better, since their purpose is different and they are complementary to each other.

- Teaching to the test

The study also aimed to examine the attitudes of teachers towards "teaching to the test" and the practices they apply for exam preparation.

Teachers seem to be quite divided on the topic. Although only by a little, a larger number of teachers indicated that teaching to the test has no negative impact

on education. Those who though that it had a detrimental effect on students' language acquisition supported their views with reasons such as those one might often find in the acedemic literature dealing with the topic. Among the issues emerging from "teaching to the test" are, that it reduces the depth of instruction and test skills don't help the students after the secondary school who did not develop critical thinking. Teachers also expressed their doubts about the lack of emphasis on areas that are not found in testing. Since ZNO is a written examination, "teaching to the test" means neglecting the entire communicative part of language learning.

Researches on the issue also suggest that while students' test scores may rise when teachers teach to the test, learning often does not improve. Moreover, even the exact opposite might be true. The findings of Neil (2003) support this idea, since in his research particular schools in New York and Boston have shown great impovements in students learning while their standardized test scores did not show significant gains.

Nevertheless, contrary to expectations, all teachers apply exam preparation in some way within the framework of English lessons, despite the fact that four out of nine teachers indicated its negative effect. The preassumed reason behind this is well supported by the remark of one of the responding teachers, who stated that teachers are expected to do so, since it is their responsibility as teachers to prepare students for whatever challenge they might face, taking ZNO being one of them.

The research also found out what is the most commonly used method when preparing to ZNO, which is not surprisingly the pracitce of common task types included in the exam. Unfortunately, this often implies that students do not develop their language knowledge, but rather their exam-taking strategies. Neil (2003) also discussed how test-taking techniques can degrade genuine reading comprehension abilities. In standardized tests students are often presented with a long text accompanied by several multiple choice questions for which one of the most popular approaches of studtens is to read the questions first, then the text. Even if

they don’t read the passage, the questions give students hints that suggest which is the correct answer.

Another research conducted by Newman, Bryk, and Nagaoka in the 1990s, investigated the assumption that "teaching to the test" force teachers to avoid the use of more communicative and authentic assignments in favor of strategies focused on memorization and repetitive practice. The results of their research suggest that the use of authentic material and assignemnts engaging critical thinking actually improves the scores of the students on standardized tests.

This finding has got important implications for developing a kind of exam preparation practice, that help students perform better on standardized tests and also develop advanced problem-solving and communication skills they will need later in the future.

Further research should be done to investigate the effect of exam preparation on teaching and learning. Research questions that should be asked include the impact of "teaching to the test" on communicative skills and the impact of exam preparation on the curriculum. Whether it is really the teachers’ responsibility to prepare students for the exam in such a direct way. Whether the curriculum is sufficiently aligned with the requirements of the exam or viceversa, could also make a good research question. Because if so, why do students need lessons specifically designed for exam preparation?

The third chapter of the master thesis presented a research on assessment practices of English teachers in the English classroom in Transcarpathian schools with Hungarian language of instruction. Firstly, the chapter reported on the methodology used for the research, its procedure, participants and the research instrument used. The data collected by the online questionnaires used as data collection instruments were described and analyzed, then in the last section these data were interpreted in relation to the pre-defined research hypotheses and previous research on the topic. Explanations were attempted to be found for the expected and unexpected results. Moreover, the limitations of the study were pointed out and future research on certain aspects was suggested.

CONCLUSIONS

This chapter of the master thesis will present a brief conclusion as to what was explored in the previous chapter, as well as implications and recommendations for future research.

Returning to the hypothesis posed at the beginning of the study, it is now possible to state that in-service teachers do not receive adequate training in assessment. One implication of the research might be that practising teachers need regular in-service training in assessment and the competent authorities should organize programs wherein teachers are introduced to newer, alternative assessment methods and their most appropriate use tailored to the needs of their classes. Otherwise, instead of applying newer innovative techniques, teachers may persist on the application of traditional methods, which no longer satisfies the effective language learning of today, nor the cognitive needs of learners.

The findings also confirmed the assumption of teachers applying traditional assessment methods more often than alternative ones. The infrequent use of alternative assessment methods is most often due to their time-consuming nature and the much more energy investment involved both on the part of the teacher and the student. However, it can be also assumed that they are simply not used because they are unknown to teachers. A more comprehensive and detailed survey could help us to establish a greater degree of accuracy on this matter, which would examine whether teachers are familiar with the different alternative assessment methods and express their opinion about them, since the present research only assessed the frequency of their use.

The research did not reveal any anticipated significant differences between the assessment practices of lower and upper secondary school teachers. The assumption that upper secondary teachers use alternative assessment more often than lower secondary school teachers has been rejected.

The study has shown that self- and peer-assmessment are less commonly used methods, while assessment of individual participation in whole class lessons

is quite common. This raises a number of questions about whether this type of assesment can be discriminatory against learners with different learning styles, which could also be a subject for further research.

Furthermore, the results suggest that students are tremendously affected by oral feedback implying the need for teachers to pay attention to the consciuos and appropriate use of it as to a secret weapon of instructional effectiveness.

This paper has given an account of how teachers view "teaching to the test" and what methods they use when preparing students for taking an exam. The study partially confirmed the assumption of teachers using "teaching to the test", although considering it bad for the students' overall language performance. This part of the research raises a number of questions for future research, as this is a very complex issue that this research does not provide sufficient insight into.

The major limitation of the present research is that it involves only certain members of the target population who were selected using non-probability methods. In order, to get a more accurate picture without any bias, it is necessary that all members of the target population have a chance to participate in the research.

REFERENCES

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C. (2001). Language testing and assessment (1). In *Language Teaching*, *Vol 34 (4)*, pp. 213-236.

Alderson, J. C. (2002). Language testing and assessment (2). In *Language Teaching*, *Vol 35 (2),* pp. 79-113.

Alderson, J. C., Clapham, C. C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press, pp. 9-207.

Alderson, J.C. (1990). Testing Reading Comprehension Skills (Part Two): Getting students to Talk about Taking a Reading Test (A Pilot Study). In *Reading in a Foreign Language, Vol 7 (1),* pp.465-502.

Alderson, J.C., & Y. Lukamani. (1989). Cognition and Levels of Comprehension as Embodied in Test Questions. In *Reading in a Foreign Language*, *Vol 5 (2),* pp. 253-270.

Angelo, T. A, & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers* (2nd ed.). San Francisco, CA: Jossey-Bass, p. 3

Armstrong, T. (1994). *Multiple intelligences in the classroom*. Alexandria, VA: Association for Curriculum Development.

Assessment Types: Diagnostic, Formative and Summative. Teaching and Learning in Higher Education. Available at: http://www.queensu.ca/teachingandlearning/modules/assessments/09_s2_01_intro_section.html

Azarnoosh, M. (2013). Peer assessment in an EFL context: attitudes and friendship bias. In *Language Testing in Asia, Vol 3 (11)*, pp.1-10.

Bachamn, L. F. & Palmer, A. S. (1996). *Language testing in practice*. New York: Oxford University Press, pp. 9-23.

Bachman, L. F. (1991). What does language testing have to offer? In *TESOL Quarterly, Vol 25 (4), pp. 671-704.*

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press, p. 241.

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. Cambridge, MA: Heinle & Heinle, p. 207.

Berger, A. (2012). Creating Language Assessment Literacy: A model for teacher Education. In *Theory and Practice in EFL Teacher Education: Bridging the Gap* (pp. 57-59). Bristol: Multilingual Matters.

Birnbaum, A. (1958). Some latent trait models and their use in inferring an examinee's ability. In F. Lord and M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Black, P. & Wiliam, D. (1998). *Assessment and Classroom Learning*. London: Routledge.

Boyles, P. (*2005). Assessment literacy*. In M. Rosenbusch (Ed.), *National assessment summit papers* (pp. 11-15). Ames, IA: Iowa State University.

Brindley, G. (2001) Assessment. In Ronald Center& David Nunan (Eds.), *The Cambridge guide to teaching English to speakers of other languages* (pp. 137-143). Cambridge: Cambridge University Press.

Brown, H. D. (1994). *Teaching by principles*. New Jersey: Prentice Hall Regents, p. 387.

Brown, H. D. (2004). *Language assessment principles and classroom practices*. Pearson Longman, pp. 3-48.

Brown, H. D. (2005). *Testing in language program: A Comprehensive Guide to English Language Assessment*. New York: McGraw-Hill, p. 21.

Brummitt, Y. J. (2017) What is Diagnostic Assessment? – Definition and Examples. Available at: https://study.com/academy/lesson/what-is-diagnostic-assessment-definition-examples.html

Carroll, J. B. & Sapon, S. M. (1959). *Modern Language Aptitude Test: MLAT; manual*. New York: Psychological Corporation.

Carroll, J. B. (1954). *Foreign language teaching: the state of the art.*

Carroll, J. B. (1961). Fundamental considerations in testing for English proficiency of foreign students. In *Testing the English proficiency of foreign students* (pp. 30–40). Washington, DC: Center for Applied Linguistics.

Carroll, J. B. (1970). *Problems of Measuerement Related to the Concept of Learning for Mastery.* Educational Horizons.

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K: C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83-118). Rowley, MA: Newbury House.

Chappuis, S., & Stiggins, R. J. (2002). Classroom Assessment for Learning. In *Educational Leadership*, *Vol 60 (1),* pp. 40-43.

Clark, J. L. D. (1983). Language Testing: Past and Current Status – Directions for the Future. In *The Modern Language Journal*, *Vol 67 (4),* pp. 431-443.

Coombe, C. & Hubley, N. (2007). Fundamentals of Language Assessment. Available at: http://www.slideshare.net/marcomed/fundamentals-of-language-assessment-manual-by-coombe-and-hubley

Crosthwaite, P. R., Bailey, D. R., & Meeker, A. (2015). Assessing in-class participation for EFL: considerations of effectiveness and fairness for different learning styles. In *Language Testing in Asia, Vol 5 (9),* pp. 1-19.

Cziko, G. (1982). Improving the psychometric, criterion-referenced, and practical qualities of integrative language tests. In *TESOL Quarterly, Vol 16 (3),* pp. 367-379.

Davidson, F., & Lynch, B.K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications.* London: Yale University Press.

Davidson, F., Hudson, T. and Lynch, B.K. (1985). Language Testing: Operationalization in Classroom Measurement and L2 Research. In Celce Murcia, M. (Ed.) *Beyond Basics in TESOL: Issues and Research in Language Teaching* (pp.137-152). Rowley, Mass. Newbury House.

Davies, A. (1990). *Principles of Language Testing*. Oxford: Basil Blackwell, Ltd, p. 10.

Dietel, R. J., Herman, J, L., & Knuth, R.A. (1991). *What does research say about assessment?* NCREL, Oak Brook.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press, pp. 110-113.

Earl, L. (2003). *Assessment as learning: Using classroom assessment to maximize student learning.* Thousand Oaks, CA: Corwin Press.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall, p. 108.

Ellis, R. (1999). *Learning a Second Language through Interaction.* John Banjamins Publishing, p. 490.

Farhady, H. (1982). Measures of language proficiency from the learner's perspective. In *TESOL Quarterly* (p. 43-59).

Fathi, J., Mohammad, Y. L. & Sedighravesh, M. (2017). The Impact of Self-assessment and Peer-assessment in Writing on the Self-regulated Learning of Iranian EFL Students. In *Journal of Sociological Research*, *Vol 8 (2),* p. 4.

Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences.* NYC: Basic Books.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st centruy*. NYC: Basic Books.

Gipps, C. V. (1994). *Beyond Testing. Toward a Theory of Educational Assessment*. Falmer Press.

Glaser, R. (1963). Instructional technology and the measurement of learing outcomes: Some questions. In *American Psychologist, Vol 18 (8), pp. 519-521.*

Goleman, D. (1995). *Emotional Intelligence*. New York: Bantam Books.

Goleman, D. (1998). *Working with Emotional Intelligence*. New York: Bantam Books.

Gronlund, N. E. (1998*). Assessment of student achievement* (6th ed.). Boston: Allyn and Bacon, pp. 209 – 226.

Gronlund, N. E. (1985). *Measurement and Evaluation in Teaching*. New York, NY: Macmillan, p.58.

Harmer, J. (2007). *The Practice of English Language Teaching* (4th ed.). Harlow: Pearson Longman.

Heaton, J. B. (1988). Writing English Language Tests. In *Longman Handbooks for Language Teachers, Vol 3 (1),* (p. 17). New York: Longman.

Henner, S. C. & Holec, H. (1985). Evaluation in an autonomous learning schema. In P. Riley (ed.), *Discourse and Learning*. London: Longman.

Henning, G. (1987). *A Guide to Language Testing*. Cambridge, Mass: Newbury House, p. 89.

Hilliard, P. (2015). Performance-Based Assessment: Reviewing the Basics. Available at: https://www.edutopia.org/blog/performance-based-assessment-reviewing-basics-patricia-hilliard

Hock. I. (2003). *Test Construction and Validation*. Budapest: Akadémia Kiadó, pp.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press, pp. 9-25.

Hughes, A. (2003). Testing for Language Teachers (2nd ed.).

Huseyin, O. (2014) Turkish Teachers' Practices of Assessment for Learning in the English as a Foreign Language Classroom. In *Journal of Language Teaching and Research, Vol 5(4)*, pp. 775-785.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. In *Language Testing*, *Vol 25 (3),* pp. 385-402.

Ingram, E. (1968). Recent Trend in Psycholinguistics: A Critical Note. In *British Journal of Psychology* (p. 74).

Ingram, E. (1977). Basic concepts in testing. In J.P.B. Allen and A. Davies (Eds.), *Testing and Experimental Methods* (p. 74). Oxford: Oxford University Press.

Kalra R., Sundrarajun C. & Komintaracat, H. (2017). Using Portfolio as an Alternative Assessment Tool to Enhance Thai EFL Students' Writing skill. In *Arab World English Journal, Vol 8 (4*), pp. 292-302.

Kerlinger, F. N. (1973). *Foundations of Behavioral Research*. New York: Holt, Rinehart and Winston, p. 458.

Lado, R. (1951). *Measurement in English as a foreign language with Special reference to Spanish-Speaking Adults*. University of Michigan doctoral dissertation. Ann Arbor: University Microfilms.

Lado, R. (1961). *Language testing. The construction and use if foreign language tests: A teacher's book*. New York: McGraw Hill Book Company.

Law, B. & Eckes, M. (1995). *Assessment and ESL.* Manitoba, Canada: Peguis publishers.

Letina, A. (2015). Application of Traditional and Alternative Assessment in Science and Social Studies Teaching. In *Croatian Journal of Education, Vol 17 (1),* pp. 137-152.

Lord, F., & Novick, M. (1968). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Madsen, H. S. (1983). *Techniques in Testing*. Oxford: Oxford University Press, p.166.

Malone, M. (2008). Training in language assessment. In Shohamy, E. & Hornberger, N. (Eds), *Encyclopedioa of language and education* (2nd ed.), *Vol 7 (2),* pp. 225-239. New York: Springer Science + Business Media.

Mathews, J. C. (1985). *Examinations: A Commentary.* London: George Allen and Unwin, pp. 90-101.

Mayer, J. D., Roberts, R. D., & Barasade, S. G. (2008). Human abilities: Emotional intelligence. In *Annual Review of Psychology, Vol 59,* pp. 507-536.

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press, p. 31.

Measurement & Evaluation. The Essentials to Measuring the Success of Workplace Learning Programs. (2006). In *Train the Trainer Guide*, *Vol 4,* pp.177-178. American Society for Training and Development.

Mehrens, W. A., & Lehman, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Belmont, CA: Wadsworth/Thomson Learning, p. 888.

Messick, S. (1989). Validity. (In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement*. Macmillan Publishing Co, Inc; American Council on Education, pp. 13-103.

Millman, J. & Greene, J. (1993). The specification of test and development of test of achievement and ability. In Linn, R.L. (Ed.). *Educational Measurement* (pp. 105-146). Phoenix, AZ: Oryx Press.

Mislevy, R., & Bock, R. (1982). *BILOG: Item analysis and test scoring with binary logistic models.* Chicago, IL: Scientific Software.

Monib, W.K. (2020). Effects of Alternative Assessment in EFL Classroom: A Systematic Review. In *American International Journal of Education and Linguistics Research, Vol 3 (2),* pp. 7-18.

Mousavi, S. A. (2002). *An acyclopedic dictionary of language testing* (3rd ed.). Taiwan: Tung Hua Book Company, p. 244.

National Council on Measurement in Education. Glossary of Important Assessment and Measurement Terms. (2017)Available at:

Neil, M. (2003b). The dangers of testing. In *Educational Leadership, Vol 60 (5)*, pp. 43-46.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago: Consortium on Chicago School Research, p. 10. Available at: http://www.consortium-chicago.org/publications/pdfs/p0a02.pdf

Oller, J. W. (1979). *Language Tests at School: A Pragmatic Approach*. London: Longman.

Oller, J. W. (1983). *Issues in language testing reasearch*. Rowley, MA: Newbury House, p. 352.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. In *Phi Delta Kappan, Vol 76* (7), p. 561.

Palm, T. (2008). Performance Assessment and Authentic Assessment: A Conceptual Analysis of the Literature. In *Practical Assessment Research and Evaluation*, *Vol 13 (4).*

Parkes, J., Abercrombie, S., & McCarty, T. (2013). Feedback sandwiches affect perception but not performance. In *Advances in Health Sciences Education*, *Vol 18 (3),* pp. 397-407.

Parviz, B. (2002). Dynamic Assessment (DA): An Evolution of the Current Trends in Language Testing and Assessment. In *Theory and Practice in Language Studies*, *Vol. 2 (4),* pp. 747-753. Finland: Academy Publisher.

Perren, G. E. (1968). Testing spoken language: Some unresolved problems. In Davies, A (ed.) *Language Testing Symposium*, *Vol 15 (2),* pp.115-138. Oxford: Oxford University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Reeves, T. C. (2000). Alternative assessment approaches for online learning environments in higher education. In *Educational Computing research*, p. 103.

Rezaee, A. A. (2013). Alternative assessment or traditional testing: How do Iranian EFL teachers respond? In *TELL, Vol 7 (2),* pp. 151-190.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. In *Psychological Bulletin, Vol 88 (2),* p. 413.

Saehu, A. (2012). Testing and Its Potential Washback. In Bambang Y. Cahyono. and Rohmani N. Indah (Eds.), *Second Language Research and Pedagogy* (pp. 119-132). Malang: State University of Malang Press.

Salim, B. (2001). *A Companion to Teaching of English.* New Delhi: Atlantic Publishers and Distributions, p. 178.

Savignon, S. J. (1982). Dictation as a measure of communicative competence in French as second language. In *Language Learning*, pp. 32-51.

Sharon, A. S. & William, C. C. (2008). *Criterion-referenced Test Development. Technical and legal guidelines for Corporate Training* (3rd ed.).

Shepard, L. A. (1998).The role of assessment in a learning culture. In *Educational Researcher, Vol 29 (7),* pp. 4-14.

Spolsky, B. (1978). Approaches to Language Testing. Advances in Language Testing Series: 2. In *Papers in Applied Linguistics,* pp. 5-10. Arlington, Virginia: Center for Applied Linguistics.

Sternberg, R. J. (1988). *The triarchic mind: A new theory of human intelligence*. New York: Viking Press.

Sternberg, R. J. (1997). *Thinking styles.* Cambridge: Cambridge University Press.

Sternberg, R. J. (2003). *Wisdom, Intelligence and Creativity Synthesized*. Cambridge University Press.

Sternberg, R. J. (2005). WICS: A model of positive educational leadership comprising wisdom, intelligence, and creativity synthesized. In *Educational Psychology Review*, *Vol 17 (3),* pp. 191-262.

Stevenson, D. L. (1985). Authenticity, Validity and a Tea Party. In *Language Testing*, *Vol 2 (1),* pp. 7-41.

Stiggins, R. J. (1991). Assessment literacy. In *Phi Delta Kappan*, *Vol 72 (7),* pp. 534-539.

Stoynoff, S. & Chapelle, C.A.(2005*). ESOL Tests and Testing: A resource for teachers and program administrators.* Alexandria, VA: TESOL Publications.

Szénásiné, S. R. (2017) Az ellenőrzés és önértékelés képességének kialakítása a „Rajtam múlik" programban. In E. Gaborják Ádám, *Mérési és értékelési módszerek az oktatásban és a pedagógusképzésben* (282-295 old.). Komáromi Nyomda és Kiadó Kft.

Topping, K. (1998). Peer-assessment between students in colleges and universities. In *Review of Educational Research*, pp. 249-276.

Underhill, N. (1987). *Testing Spoken Language: A Handbook of testing techniques*. Cambridge: Cambridge University Press, p. 8.

Weir, C. (1998). *Communicative language testing.* Hemel Hempstead: Prentice Hall, p. 46.

Weir, C. J. (1990). *Communicative Language Testing.* London: Prentice Hall, p. 27.

Winking, D. (1997). *Critical issue: Ensuring equity with alternative assessments*. NCREL, Oak Brook: IL.

Wolf, D. P. (1991). To Use Their Minds Well: Investigating New Forms of Student Assessment. In G. Grant, ed., *Review of Research in Education, Vol 17,* pp. 31-74. Washington, DC: American Educational Research Association.

РЕЗЮМЕ

Мало що відомо про думки та переконання вчителів Закарпаття щодо тестування та, що більш важливо, про те, як вони використовуються в контексті освіти.

Виходячи з цього, темою моєї роботи є методи оцінювання в класах англійської мови в угорськомовних школах Закарпаття.

Темою роботи є використання традиційних та альтернативних методів оцінювання та зворотного зв'язку викладачів англійської мови, а також застосування "тестового навчання" в класах англійської мови в школах з угорською мовою Закарпаття.

Метою дослідження є вивчення традиційних та альтернативних методів оцінювання, що застосовуються серед учителів середніх угорських шкіл Закарпаття, та виявлення подібності та відмінності в практиці оцінювання викладачів молодших та старших класів середньої школи. Подальшою метою є виділення областей мовного тестування та оцінювання, які, можливо, доведеться вдосконалити, та підкреслення важливості змістовного зворотного зв'язку. Також маємо на меті дослідити ставлення викладачів до «тестового навчання » та вивчити, як відбувається підготовка до іспиту до зовнішнього незалежного оцінювання в 11 класі.

У дослідженні використовуються як теоретичні, так і емпіричні методи. Сюди входить емпіричне дослідження з використанням якісних та кількісних методів збору даних. Інтернет-анкета використовується як інструмент збору даних.

Практична цінність дослідження полягає в тому, що воно дає корисне розуміння практики оцінювання в класах англійської мови в школі з угорською мовою на Закарпатті, що може забезпечити важливу основу для подальших досліджень та виявити ряд недоліків на користь майбутніх розробок.

Робота складається із вступу, трьох частин, короткого викладу, списку використаної літератури, додатка та короткого викладу. Частина 1 забезпечує

теоретичну та концептуальну основу для дослідження шляхом перегляду літератури з мовного оцінювання, включаючи зміни в часі концепції та практики оцінювання та тестування, а також характеристик та застосувань різних методів та засобів тестування. Частина 2 описує етапи створення випробувань, починаючи від специфікацій випробувань (умов), закінчуючи звітами про перевірку та після тесту. Частина 3 представляє хід та результати згаданого емпіричного дослідження, їх обговорення та зроблені висновки.

Результати дослідження підтвердили низку гіпотез, але є й такі, які спростовуються. Результати дослідження говорять про те, що вчителі потребують регулярного підвищення кваліфікації, а відповідальні за це повинні організовувати програми, в рамках яких вчителі можуть дізнаватися про нові, альтернативні методи оцінювання та навчитися пристосовувати їх до потреб своїх учнів.

Отримані результати також підтверджують гіпотезу про те, що вчителі частіше використовують альтернативні методи традиційного оцінювання. Рідке використання альтернативних методів найчастіше займає багато часу і, а також вимагає набагато більше вкладень енергії як викладача, так і учнів. Однак можна також припустити, що вони не використовуються, оскільки раніше їх просто не знали. Більш всебічне та детальне опитування могло б допомогти нам сформулювати більш точні відповіді на це питання, які б дослідили, чи знайомі вчителі з різними альтернативними методами та в яких вони могли б висловити свою думку щодо них, оскільки в цьому дослідженні розглядалася лише частота їх використання.

Всупереч очікуванням, дослідження не виявило суттєвої різниці між методами оцінювання викладачів нижчої та старшої середньої школи. Припущення, що вчителі частіше використовують альтернативні методи оцінювання у старших класах середніх класів, ніж у молодших, не було обґрунтованим. Дослідження також показало, що самооцінка та співоцінка є рідше використовуваними методами, тоді як оцінка індивідуальної активності та участі на заняттях у цілому класі досить поширена. Це

викликає низку запитань щодо того, чи не є такий тип оцінки дискримінаційним щодо учнів, які використовують різні техніки навчання.

Окрім того, результати додатково свідчать про те, що на студентів впливає сильний усний відгук вчителів, що вказує на те, що вчителі повинні приділяти йому більше уваги та використовувати його свідомо та доречно.

Дослідження також повідомляє про те, як викладачі розглядають тест і як вони використовують його для підготовки учнів до іспиту. Наше дослідження підтвердило припущення, що вчителі застосовують викладання до тесту, хоча вони вважають це поганим для загальної успішності учнів. Ця частина дослідження порушує низку питань для подальших досліджень, оскільки це дуже складне питання, яке це дослідження не дає достатнього розуміння.

Обмеженням цього дослідження є те, що воно включає лише певних представників цільової сукупності, які були відібрані шляхом випадкового вибору. Щоб отримати більш достовірнішу картину без упередженості, необхідно, щоб усі представники цільової групи мали можливість взяти участь у дослідженні.

Ключові слова: традиційне оцінювання, альтернативне оцінювання, зворотний зв'язок, навчання для тестування

NYILATKOZAT

Alulírott, Gál Vivien Gréta angol szakos hallgató, kijelentem, hogy a dolgozatomat a II. Rákóczi Ferenc Kárpátaljai Magyar Főiskolán, a Filológia tanszéken készítettem, angol nyelv és irodalom tanári diploma megszerzése végett.

Kijelentem, hogy a dolgozatot más szakon korábban nem védtem meg, saját munkám eredménye, és csak a hivatkozott forrásokat (szakirodalom, eszközök, stb.) használtam fel.

Tudomásul veszem, hogy dolgozatomat a II. Rákóczi Ferenc Kárpátaljai Magyar Főiskola könyvtárának Kézirattárában helyezik el