UNIVERSITETI I EVROPËS JUGLINDORE
УНИВЕРЗИТЕТ НА ЈУГОИСТОЧНА ЕВРОПА
SOUTH EAST EUROPEAN UNIVERSITY

Validity Impact in Assessment

Professor: Candidate:

Prof. Dr. Veronika Kareva Arjeta Rexhepi

Tetovë, 2020

Left blank intentionally

Declaration of authorship

I declare that I have worked on my master thesis on my own - pursuing the principles of academic work in word and spirit – and that I have used the sources mentioned in the bibliography.

Signature: Arjeta Rexhepi

Date:

Acknowledgement

Reaching my life goals would have been impossible without the people who supported and mentored me. Therefore, I would like to thank Professor Veronika Kareva, from the bottom of my heart, for her irreplaceable support during the studies and especially in the completion of this thesis. Her commitment in clarifying every single detail was a milestone.

Additionally, I also would like to pay my special regards to my family, for their unconditional support and love in each step of mine. Their presence eases each life stage. In regard to them, my husband and my son have been my greatest inspiration ever, thus my thesis is dedicated to all of them.

Lastly, very warm thanks to my colleagues and students, who constructed this thesis with their extremely useful opinions. Their contribution was crucial in the process of inquiry. Therefore, my gratitude will always be extraordinary towards their contribution.

Abstract

This study aims to exemplify the importance of assessment in teaching and its main focus is the impact of validity in assessment, which shows the accuracy and appropriateness of measuring the intended content. Assessment plays an integral role in the process of teaching a second language, thus the process of evaluating students' performance refers to the variety of ways that teachers use to collect data, which include tests, more specifically reliability and validity tests. Hereto, reliability and validity tests are the most influential classification that has been done in assessing students' achievement, as they have a consistent and clarified aim. The former mentioned leads to the accuracy of measurement, in terms of the same student subduing the test in different circumstances, meanwhile the latter mentioned, which is considered the key to assessing, refers to how well a test measures what it is supposed to be measured. Reliability is insufficient without validity; moreover it is the best way to see students' results and teachers' strategies in teaching.

A convenient way of teaching ESL is assessing the four basic skills of learning it, which easily brings effectiveness to the realization of syllabi outcomes. Validation procedure highlights the adequacy and appropriateness of testing the intended material, which contributes in students' motivation, but raises the responsibility of teachers to teach as effectively as possible and with a greater expertise in realizing the outcomes. Hence, establishing and raising awareness about the importance of validity impact in assessment is crucial, thus this study gives very useful information to teachers in compiling and applying validation procedure, proving the necessity to compile reliable and valid tests at the same time.

The inquiry of this study was conducted in two primary schools and a private language school of Viti Municipality; with a higher proportion of time in "Skender Emerllahu" and "Oxford Education Corner" school, continuing in "Bafti Haxhiu" as well. The determined objectives for this research were achieved through different quantitative and qualitative methods, such as: interviews with 10 teachers, surveys for 150 students, direct observation in 6 English classes and a focus group; 10 teachers were supposed to talk about challenges of assessing students' progress and the activity was realized in "Oxford Education Corner" school, Viti. In order to

completely realize the aim of this research 20 English teachers and 150 students were the main stakeholders, whose opinion offers beneficial data on the importance of testing appropriately and accurately.

Keywords: Validity, assessment, evaluation, reliability, impact, grading, tests.

Abstrakt

Ky studim synon të jap shembuj rreth rëndësisë së vlerësimit në mësimdhënie dhe qëllimi kryesor i tij është ndikimi i vlefshmërisë në vlerësim, që shfaq saktësinë dhe përshtatshmërinë e matjes së një përmbajtje të planifikuar. Vlerësimi luan një rol përbërës në procesin e mësimdhënies së një gjuhe të huaj, ashtu që procesi i vlerësimit të performancës së nxënësve i referohet mënyrave të llojllojshme që mësimdhënësit i përdorin për të mbledhur të dhëna për arritjen e nxënësve, të cilat përfshijnë testet, më saktë testeve valide dhe të besueshme. Për më tepër, testet e besueshme dhe valide janë klasifikimi me ndikimin më të madh, që ndonjëherë është bërë në matjen e arritjes së nxënësve, pasi që ato kanë një qëllim qartë dhe të qëndrueshëm. E para e lartpërmendur nënkupton saktësinë e matjes, në kuptimin që i njëjti nxënës t'i nënshtrohet testit të njëjtë në rrethana të ndryshme, ndërsa e dyta e lartpërmendur, që është çelësi i vlerësimit, nënkupton se sa mirë është matur ajo që është planifikuar të matet. Serioziteti i testeve është i pamjaftueshëm pa vlefshmërinë e testeve, në të vërtetë është mënyra më e mirë të shohim rezulatet e nxënësve dhe strategjitë e mësimdhënësve në procesin e mësimdhënies.

Një mënyrë e përshtatshme e mësimdhënies së një gjuhe të dytë është vlerësimi i aftësive themelore të mësimit të saj, e cila sjellë lehtësisht efikasitet në realizimin e objektivave të planprogrameve të caktuara. Procedura e vlefshmërisë së testimit thekson përshtatshmërinë dhe saktësinë e testimit të materialit të caktuar, që kontribon në motivimin e nxënësve, por rrit përgjegjësinë e mësimdhënësve që të japin mësim sa më efikas me një ekspertizë më të madhe në realizimin e objektivave. Prandaj, të themeluarit dhe të ngriturit e vetëdijësimit rreth rëndësisë së ndikimit të validitetit në vlerësim është thelbësore, kështu që, në kjartësimin dhe

arritjen e tij, ky punim jep të dhëna të dobishme se si te strukturohet dhe të aplikohet procedura e vlefshmërisë, duke dëshmuar në të njëjtën kohë nevojën e hartimit të testeve të besueshme dhe valide.

Ky hulumtim u realizua në dy shkolla fillore dhe një shkollë private të gjuhëve të huaja të komunës së Vitisë: me një proporcion më të lartë të kohës në “Skender Emërllahu” dhe “Oxford Education Corner”, duke vazhduar me shkollën fillore “Bafti Haxhiu” . Objektivat e përcaktuara për këtë hulumtim janë arritur përmes metodave sasiore dhe cilësore, sikur se: Intervistat me 10 mësimdhënës, anketimet për 150 nxënës, vëzhgimi i drejtpërdrejt në 6 orë mësimore dhe fokus grupi; në të cilin 10 mësimdhënës diskutuan sfidat e vlerësimit të punës së nxënësve. Dhjetë mësimdhënës të gjuhës angleze dhe 150 nxënës ishin palët kryesore të përfshira në arritjen e qëllimit, mendimi i të cilëve ka sjellur të dhëna fitimprurëse në rënësinë e testimit të saktë dhe të përshtatshëm.

Fjalët kyqe: vlefshmëria, vlerësimi, siguria, ndikimi, notimi, testet.

Working title of the thesis:

“Validity Impact in Assessment”

Translation of the working title in Albanian:

“Ndikimi i Validitetit të Testeve në Vlerësim”

Translation of the working title in Macedonian:

“Валидноста при оценувањето”

Table of Contents

List of figures

List of tables

List of abbreviations

ESL – English as a second language

CV – Construct –oriented validity

TEFL – Teaching English as a foreign language

CA – Classroom assessment

SA – Summative assessment

FA – Formative assessment

EFL – English as a foreign language

TOEFL – Test of English as a foreign language

CEFR – Common European Framework

Chapter I

Introduction

Teaching is considered closely related to countries development. Thus, the educational system needs to be priority ranged in governing, in order to develop and train the future generations for trades-necessities and have qualified and well-prepared teachers. Hereto, one of the most important in education is assessment. Why assessment? This study gives to readers a proven assumption about the evaluation process and reliability and what is most important; validity and validation procedure. Teachers' responsibility is doubtless the hugest, since they mark students' achievement by assessment in general.

Assessment is considered an inseparable part of teaching and learning process, leading to observation of students' achievement, done by formative assessment, proceeding with summative assessment; which means the grading system. The two above mentioned terminologies are useless if teachers are not totally sure what to measure and how. For school and sometimes governmental purposes, students are consistently being assessed, meaning that teachers are obliged to provide suitable techniques and methods to best measure the completion of the syllabus and the realization of the determined outcomes. In consequence, teachers' familiarity with the closest concepts of assessment, such as reliability and validity, are crucial. (Glenn & Fred, 2007) The latter mentioned and its impact in evaluation is the one analyzed in this study, as it is the most important concept in the process of assessing students. Validity is very important and teachers who compile tests based on 'How' and 'What' raises students' awareness about the importance of everyday learning and makes them know what to expect in regard to that certain syllabus and its outcomes.

This study helps teachers find solution on the anxiety created by the idea of assessing students and not being fair enough. Surveys realized with teachers and students made its analysis and solutions possible. The reader has both modern and classic ways in which assessment has been continuously laid on. Furthermore, the results from each method of inquiry are explained in details within this study.

1.1 Research aim

Teaching a second language is a process that looks for a great amount of time in planning every single part of it, including assessment as the most difficult branch of teaching. Accuracy is highly required in the process of assessing students' achievement. Despite the fact that it mostly deals with the observation of students throughout the year, the difficulties are inevitable if the teacher does not know testing purpose and key concepts of what students have to benefit at the end of that course. Therefore, the aim of the thesis is to identify the difficulties in the process of assessing students' progress, highlighting the importance and impact of validity/validation in regard to accuracy and appropriateness of testing methods. It is considered one of the most important and useful branches in the process of teaching and learning stated by Scriven:

> Evaluation itself is a logical activity which is essentially similar whether we are trying to evaluate coffee machines or teaching machines, plans for a house or plans for a curriculum. The activity consists simply in the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings" (Scriven, Tyler, & Gagne, 1967, p. 40)

Assessment is considered like the most influential sequence in education, made at different levels and for many different purposes, such as: Diagnostic tests, placement tests, achievement tests and proficiency tests. Therefore, reaching the goals teachers need to take into consideration the used methodology to complete that specific course. Teacher judgments are the most important activities for realizing all the planned learning outcomes (Lundahl, 2011). Their judgments play an important role in determining grades and identifying pupils with disabilities or those in need of special assistance. Relating the purposes and structuring a test appropriately, teachers need to take into consideration the main important types of assessing that are related to validity, such as: Criterion validity, content validity, face validity, construct validity, as incredibly worshiped source of evidence that gives teachers the opportunity to discuss the key skills that are necessary to be measured.

Moreover, identifying testing purpose facilitates its compilation. Measuring achievement or outcomes realization, teachers have to collect evidence which supports that the test measures the intended material. According to Johansson (2013), evidence in validity is a priority, in order to raise confidence in these inferences (Johansson, 2013). In addition to students' progress measurement, validation procedure offers results on teachers' effectiveness in regard to the completion of the material and their teaching methodology.

The strategies for establishing validity are well aligned with the mixed methods approach. Validity tests can be quantitative, qualitative and quantitative/qualitative according to Ridenour and Newman (2008), who stated that mixed methods approach to validity is based on an inductive–deductive philosophy (Ridenour & Newman, 2008) . The validity of tests is often related to content-oriented strategy, as one of the best ways to prove the outcomes brought by the results. There are some other important validity types and each of them has a great impact on measuring what a test is supposed to measure. Its methods are focused on relevance and representation. A research ensuring the quality of assessment used two different methods of evaluating - validity tests and traditional ways of assessing, focusing more on the former mentioned resulted very productive. This helped on compiling and verifying the validation of tests. Establishing strategies helped language experts compile tests and reach the outcomes they have planned, providing them accuracy and effectiveness in teaching. Moreover, those strategies created possibilities to highly work on their weaknesses. (Norris, 2004).

1.2 General and specific objectives of the study

- The main aim of this study is to show the importance of assessing students' achievement by using reliability and validity, with the focus on the latter mentioned. It also highlights the major types of validity and their impact in the process of accurately assessing the realization of the outcomes in a certain syllabus.

  This study is going to offer different important approaches which can be included in evaluating students and at the same time is going to help teachers reflect on the effectiveness of their teaching process.

Teachers will be able to find out some specific objectives as well:

- Find out the importance of assessment in the process of teaching a second language.
- Define the most useful validity approaches in evaluating students.
- Identify and analyze the impact of validity in assessment.
- Understand how compiling a reliable and valid test results positively.
- Identify the impact that validity tests might have in students' motivation.

1.3 Research questions and hypotheses

The key questions throughout this study are:

- What is the impact of validity in assessment?
- What is the contribution of validity in the process of teaching and learning a second language?

Other research questions:

- Does the use of different validity tests types increase students' motivation and confidence in learning a second language?
- How can validity tests types contribute in the accuracy of measuring what a test is supposed to measure?
- Has validity got any impact in teachers' self-reflection? How is it achieved?

Hypotheses:

1. Validity tests have a great and crucial impact in assessment.
2. The accuracy in measuring the exact skills observed throughout a year influences students self-confidence and awareness of their own achievement.
3. Validity helps teachers reflect on their own teaching effectiveness.

4. Assessing the four basic skills of learning a second language can be realized by being accurate what and how to assess students' achievement.

Chapter II

Literature review

The rapid revolution of bilingualism has influenced people's way of living. Specifically, English langauge has widely spread all over the world. Nowadays, teaching ESL requires an intermitent that needs to be consistently well-prepared and an expert in teaching and assessing your own effectiveness simultaneously with students' progress in learning ESL. Thus, to facilitate and make teaching more effective all the teachers need to know the classification of assessment and its importance.

Detailed analysis of validity impact is provided to the reader in this study, including the accuracy and appropriateness of test compilation by teachers, equipping them with proper information on comprehending their effectiveness preferably. The accuracy of testing adequately the specific outcomes within a year increases students' motivation level, making them conscious about what they are supposed to achieve within an academic year and the way their assessment is going to take place. Validation procedure brings freshness, determination and adequacy in evaluation, raising teachers' authority, as it offers the possibility to create a correlation of test compilation to the outcomes of that specific course, equipped with proves of each student achievement. Its effectiveness is related to teachers' professional development in assessment, as teachers' training on compiling standardized tests is closely related to validity.

Assessment is a process that occurs consciously by observing students through the whole academic year, in order to have evidence about each student's achievement. Most of the books in language testing deal to some extent with validity. The language testing and assessment can be compared to our everyday lives. To clarify it, when people ask whether someone loves us or not, coming to a conclusion based on the person's reactions, feelings and behavior, prompts people allude on the impact of validity in our daily basis, related to the assessment process in education. (Glenn & Fred, 2007) Students' achievement is the crown given to each of them at the end of each semester, which is mostly done by testing their knowledge and the completion

of the outcomes set by teachers for each specific syllabus at the beginning of the academic year. By all the means the accuracy meaning of this can be achieved through validity as the central concept in assessment and testing.

Linguists shared different views about validity types. According to Messick (1989) emphasizing that construct validity is shown from different perspectives and the questions needed to be paraphrased (Messick, Educational Measurement, Third Edition, 1989) meanwhile Chapelle (1999) provided different views of language testing, highlighting the acceptability of more than one answer for the question “Does the test measure what is supposed to measure?”. (Chapelle, 1999, p. 19) Another similar view to Messick’s is the one of Bachman and Plamer (1996) who consider validity as the crucial point on which tests should be judged, incorporating the view of Messick about construct validity and adding extra details on developing tests. According to them, validation is an ongoing process which can never be totally completed and its usefulness consists of six qualities which are going to be mentioned further on in this study, including practicality and authenticity. Their model serves as a plain sketch of validity impact, importance and the advancement of tests in general (Bachman L. F., 1996).

Validity is considered a property of tests, not just an influential part of it and it is directly related to the process of scoring. All the above mentioned linguists shared about the same view related to the concept of validity, which is going to be further explained in details, but the differences are mostly on its impact and broadened meaning of its concept. Their views differ mostly in the classification of validity into three types and their importance.

2.1 Introduction of validity impact in assessment

Validity is not a common notion and not very familiar to test-givers. Its meaning is not different from the one used in daily life, but deeper to understand in the world of education, especially in language testing and assessment. According to Messick (1989), validity is the process of appropriateness and adequacy of actions on the basis of test scores supported by empirical evidence and theoretical rationales. (Messick, Educational Measurement, Third Edition, 1989)Its traditional concept interlocks in three main classifications, such as: Construct-related,

content-related and criterion-related Validity, which are not considered enough completed, but the new unified perspective integrates considerations of content, criterion and consequence into construct framework. The former mentioned are now considered as a part of CV framework. (Messick, Educational Measurement, Third Edition, 1989) Following the procedures of testing based on validity facilitates teachers' work, giving them the possibility to reflect on their work and the details that should be changed in the future. Validity is seen as a unitary concept, meaning that all those examining and observing a test come up with the same opinion about the scoring results, which can be ranged from low to high. This component is helpful and necessary when compiling tests, not letting a chance for questioned validation. Even students are going to better and easier understand the accuracy of their answers and based on the criteria the results are going to be easily guessed. It is a complex procedure, as a close relationship exists between the inferences and assessment, but it is an achievable process and ongoing process to realize the intended goal to be measured.

Traditionally, testing has not been taken seriously and teachers most of the time worked on their own, compiling tests without any common framework. They compiled tests based on their opinions, so Validity performance was not mentioned. Assessment is a hard process that needs proves to demonstrate the difficulties and has an ethical dimension that influences people's lives, thus fairness is crucial in assessing students' knowledge and deals with the consequential validity. Fairness in testing is a duty of teachers while testing, not an ideal of assessment. (Cyril, 2005) Moreover, fairness is a key concept in the procedure of validation, which makes accuracy and appropriateness achievable. Validity is multifaceted, thus evidence and different types are needed to support claims for scoring interpretations. They have the same importance according to the traditional way of assessment and validation, but the new perspective of validation ranges construct as the most important one. Assessing students and being accurate means a detailed plan of evaluation and testing, dividing the assignments and its percentage since the beginning of each year, compiling a syllabus with students' consent, in order to fulfill their needs and requirements.

Validity raises teachers' accuracy and helps them gather evidence on scoring results throughout the academic year. The relationship of testing and teaching has got difficulties and problems in the field of TEFL. This comes as a result of not enough training programs for ESL teachers, who have an immense need to professionally advance themselves in assessing students. A common national framework is necessary, which fairly and similarly teach and assess students' learning progress. (Michael, 1988) Validity is not just used for interpretation inferences about test scores, but also about inferences made on observations of different attributes. The concept used in validity about scoring includes the procedure of gathering data, observing classes, portfolio and questionnaires. The correlation of scoring and assessment is completely clear and close, as every technique of assessment shows nothing without the scores and criterion set for that specific test. (Messick, Educational Measurement, Third Edition, 1989)

This study offers to readers a clear idea about how should teachers manage the most untouched and difficult area of teaching. Validity influences positively the way a teacher thinks and gives greater opportunities to choose techniques of compiling tests and organizing the whole procedure of testing time, equipped with evidence and enough demonstrative documents for each students' achievement and knowledge. The Ministry of Education has to invest on teachers' trainings, enriching them with methods of assessing and reflecting on the details that should be changed in the future, if they want to have a larger amount of accuracy in the process of evaluation. Teachers in the past never had the opportunity to professionally grow, because of low interest from the government which was led by Kosovo's invaders. Their interest was not on advancing teachers, but no such issues anymore, therefore the government should take this seriously into consideration in order to have a developed country, as each section is depended on the education system. Teachers never had a common framework for students' grading scale. A common national framework for grading students' achievement should be designed by experts of this field, respecting the rules of Validation. As a result, respecting validity rules means becoming an inseparable part of this new era. Even feedback becomes easier if a plan is structured in details.

2.2 Assessment

Testing and assessment are extraordinarily important in ESL. The terminology of assessing in general is not related to just education, but to a wide range of situation such as; the adults when applying for a new job or want to advance themselves further on, meanwhile, children's assessment happens mostly at school, in consequence to their future preparation and professional advancement. Therefore, making learning happen is not just about happening, but about witnessing its results. (Phil, 2005)

What do people or even novice teacher know about testing and assessment? There is a chance to know basic things or all its classifications, but not any experience yet. The word assessment itself should be understood by each teacher, analyzing its advantages, its correlation with the process of teaching/learning and its impact in the before mentioned processes. Its meaning in education is not just the one given in the dictionary, so not superficially use it, but it is a much broader term. Assessment means decision making, observation, coronation, engagement, scores, results and its closest synonym in education is grading. The broadness of its meaning is limitless because it has the greatest role in the process of teaching and learning a language.

Markedly, experienced teachers have a clearer idea about testing in general, as throughout years every teacher learns and practices new assessing techniques. Different factors influence students' readiness to being tested when it comes to grading. Thus, teachers have to know their students if they want to fairly assess them. Assessment is the section of education that no one does it perfectly, but new experiences taken year by year help teachers gather and use methods that function better. Testing helps teachers' measure students' ability to write, speak, talk and listen in a foreign language. Tests are performance-based because they are held in a social environment that encourages communication, interaction, sharing goals and giving feedback students to students and teacher to students (Glenn & Fred, 2007). The feedback, teacher to students, happens most commonly, offering them evidence about scoring results and providing opportunities to students to improve their mistakes and focus on their weaknesses.

2.2.1 The general impact of assessment in teaching and learning

Assessment serves as a tool to keep students engaged all the time in the process of learning a foreign language and teachers observe their advancement during a whole academic year, in order to coronate their achievement. Having in mind this, they struggle a lot when they are supposed to be graded, but at the same time without the grading idea, their engagement in the process of learning would be more difficult.

Making students aware about the purpose they are learning for is primarily ranges, as they are not capable of thinking. Nowadays, classroom assessment can be multi-component in comparison to the traditional assessment; characterized for single focus tests. Modern tests in CA have multiple focuses, demonstrating students' engagement throughout a specific period and teachers' effectiveness. CA impact is enormous, specifically teaching and learning is difficult to be imagined without assessment. (MacMillan, 2013)

Students experience written assessment in different ways. There are some that express themselves easier in writing, some others in speaking, but teachers set assessment techniques for all the students, as differentiation can be realized, but related to the content, not to change techniques of assessment. What should teachers do in this case? Students might have fear of tests, consequently the teachers' responsibility is to motivate and prepare them in advance. Another option is to decide about the assessment together with the student at the beginning of each year and give them the opportunity to be part of decision making, raising their confidence towards their own progress. When students are involved in the discussion of creating syllabi, discussions about the appropriateness of assessment techniques will not be questioned (MacMillan, 2013). Assessment role is not questionable in the process of education, as it engages students all the time in the process of gaining new information, retrieving and putting it in use. Its impact is considered a driving force that prompts students to learn and wait results about their own achievement. (Phil, 2005)

2.3 Types of assessment

Assessment has got purposes, which serve to define the arguments about their interpretation and use. The main purposes are described and classified to purport better the relationship of teaching and assessment and to better describe the way assessment happens throughout the academic year. The biggest classifications are formative and summative assessment (MacMillan, 2013). Teachers are highly familiarized with these two largest branches of assessing students and they use appropriate techniques to accomplish the appropriateness of assessment in general. Formative and summative assessments are both used in each academic year. FA is related to observation, unlike SA is needed to crown up students' achievement with grades at the end of each semester. FA and SA are closely connected to reliability and validity, if teachers want to have fairness in the process of evaluation.

Formative assessment or referred to as assessment for learning, is the one focused more on students' gaining process of necessary data about a specific subject. FA helps teachers observe their way of teaching and the progress of their students in EFL specifically. FA can be interchangeably used with observation because its function is about observing and revealing the realization of the specific outcomes through students' achievement. Students are capable of understanding about their work and progress daily, weekly or monthly, depending on the teacher' preferable timing about giving feedbacks on different portfolio work and project based approaches. Henceforth, this way of assessing students' achievement makes them an inseparable part of their own evaluation, through qualitative feedbacks. Recently, FA is more common and more preferable, as evidencing someone's progress raises accuracy and convinces more the other pair involved in by observing their work periodically. This also contributes to two most important factors of assessment, such as: Validity and reliability (considered now as an attribution of Validity) (MacMillan, 2013).

On the other hand, summative assessment is used to systematically give evidence of students' achievement to a specific audience. Its purpose is to come up with certain inferences for the sake of others' necessity. It is used to grade students work with numbers or letters, which have

a specific meaning according to a specific framework, such as the common European framework. The inferences are usually elicited to be given to the school principal, administrators and parents, through graded tests mostly or every suitable assessing techniques that can be graded. This happens systematically at certain times during an academic year, periods specified by the ministry of education or school principal. SA is considered authoritative at some points because it causes discouragement to students who are not convinced about the grade, thus teachers have to be extremely careful when it comes to grading, which is inevitably related to validity and reliability. (MacMillan, 2013)

The two main types of assessment, which are closely related to the most important attributes of assessing students, such as validity and reliability, have a very important role in assessing students. Teachers should also define techniques which are usable for each of them and use them appropriately for just one purpose, - for students' progress. Using validity and reliability test in assessment means respecting the basic rules of how evaluation process is supposed to happen, but it does not guarantee fairness in the whole process. Additionally, the above mentioned attributes should be seriously taken into consideration to properly assess.

2.4 Reliability tests

Tests are commonly considered poorly constructed, thus the attributes mentioned throughout this study are the ones needed to be taken seriously, in order to start fading away the incapability of teachers to form test. A common framework and standardized tests would help teachers in the process of fairness likewise prompting students to subdue the test more than once and have the same results is closely related to the concept of reliability. Before Messick's perspective, reliability was considered as important as validity, but now it is seen like an attribute of validity itself (Messick, Educational Measurement, Third Edition, 1989).

Tests are reliable when students' achievements bring the same results even after subduing the test more than twice and being observed by different experts. Reliability has never opposed validity, in fact it has always served to the process of validation. Reliability is easily measured and verified, thus nowadays it is being put into validity, as an attribution of it, which highly

contributes in the process of validating tests, creating an overlap of validity above reliability. Tests can be reliable and give the same results when measuring the same phenomenon, but cannot be valid without being reliable, giving us the reason why reliability is set in validity process. Reliability refers to high internal consistency if the phenomenon being measured brings the same result after repetition of it more than once. (Messick, Educational Measurement, Third Edition, 1989)

Another key point is that there are four types of reliability, which are used to verify whether a test is reliable or not, such as:

a) Test – retest reliability (related to stability of scores after repeating tests)
b) Parallel forms of reliability (meaning equivalence of two different forms of the test)
c) Internal consistency reliability (providing measurement of the construct by each item of the test)
d) Inter rater reliability (talking about the consistency of the rater measuring the rates).

The four classifications are used to measure, prove and define reliable tests and reliability in general as a process of measuring and attribution of assessment (Messick, Validity and Washback in Language Testing, 1996).

2.5 Validity in assessment

Traditionally, validity has been considered a separate concept from reliability, with a close relationship to each other, but lately this has been changed, putting validity as the main attribute in testing fairness. Furthermore, validity was considered a property of the tests, a shifted issue by the famous work of the American psychologist Samuel Messick, who made validity more comprehensive to the world, in the 70s and 80s, coming to a unified model in the late 80s, more accurately culminating in his seminal in 1989, at the educational testing service, the organization of American educational and measurement. Messick worked on simplifying the concept of validation, coming to culmination of inexistence of reliability without validity, leading to the close relation of validity and reliability. Secondly, important and closed discussion

is the one whether validity is a property or a characteristic of tests in general, which crowns it as a characteristic of tests, but a property of test interpretation. The last issue he dealt with was about one type of validity, precisely construct validity, which is going to be elaborated in details in the next pages of this study (Messick, Educational Measurement, Third Edition, 1989). Messick presented a new unified model of validity, in contrast to the traditional one, considering CV an inclusion of both content and criterion-oriented validity, ranging it as the most important component of validation in general. Testing requires accuracy and appropriateness, but how can it be accomplished? Validity is the most suitable answer for formulating an accurate and appropriate test. Its meaning is closely related to fairness, as it intends to measure what it is supposed to measure, stated by Messick (1989) as follows:

"…always refers to the degree to which empirical evidences and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores." (Messick, Educational Measurement, Third Edition, 1989, p. 6)

Another similar view is Hughes's, who relates the accuracy of tests to validity, traditionally known as a concept that is used to discover the focus of the test or uncover of the appropriateness. Its meaning is the presupposition that each test has a specific focus to measure and this can easily be realized the procedure of validation. (Glenn & Fred, 2007) Validity concept involves the question whether a test measures its focus and validation procedure precedes the results, that is, why it is considered a property of results' interpretation, but a characteristic of the test in general. Henceforth, Messick (1989) presented a progressive matrix, which has four classifications within it, such as: Justification, which includes evidence or consequence and the function or outcome; which has use or interpretation of the test. (Messick, Educational Measurement, Third Edition, 1989)His framework has been widely accepted, but considered difficult to put in operation. (MacNamara, 2006)

	Function of Testing	
Source of Justification	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity +

		Relevance /Utility
Consequential Basis	Construct Validity + Value Implications	Construct Validity + Relevance/Utility + Social Consequences

Table 1) Messick's facets of validity classification (Messick, Educational Measurement, Third Edition, 1989)

Messick (1989) emphasizes the importance of social consequences and its impact in testing, including the administrative procedures and even the environment itself. According to him, both interpretation and test in general, involves the terminology of value. Even the whole concept of validity is related to value in general. Validity is the most suitable answer for formulating an accurate and appropriate test. Its meaning is closely related to fairness, as it intends to measure what it is supposed to measure. (Messick, Educational Measurement, Third Edition, 1989)

Messick's assumption has been widely accepted and taken into consideration in further exploration of validity and validation as a procedure, such as the one of Bachman (1990) in his influential fundamental considerations in language testing. (Bachman L. F., 1990)Weir and Shaw (2007) used the influential shift that has been made about the reliability as an aspect of validity, according to Messick's unified model of validity. (Shaw, 2007) Messick's unified model created a shift in validity types, which are extremely necessary, in order to rely testing and assessment on Validation in general. Validity types are going to further be explained and analyzed, in order to come to a better understanding how the whole process of validity takes place in testing.

2.5.1 Types of validity

Validity classifications are made by the traditional theory, which emphasizes three main types and analyzes each separately. The three of them are used to offer evidence in the procedure of validation in evaluation.

1.Criterion-oriented validity – known as concrete-oriented validity is the relationship between a specific test and the predictions that the test-givers might create. For example, if a student is having extra English classes to prepare for TOEFL test, the teacher might wonder whether that student will pass TOEFL test after finishing the ESL classes. The scenario mentioned above cannot be witnessed before the adequate time to apply and take the TOEFL test. A very close relationship coexists between the prediction and the performance of the action being performed. Teachers have to focus on the criterion or criteria when working on a specific academic year or even working with an ESL student, as this is one of the possibilities of accurately and appropriately measuring and evaluating. When there is a goal in mind, means even the criterion is created at the same time. (Glenn & Fred, 2007) This is a central part of validating or investigating validity in testing. Hereto, parts of criterion-oriented are two other subtypes, which are tightly related to each other, but analyzed separately:

Predictive validity – in the above case, the predictive point is the criterion related to the scores whether that TOEFL student will pass the final exam or not. Consequently, the tests are linked to extra English classes, in order to pass a preparative test, precisely the academic achievement that the student succeeds after taking the real TOEFL exam.

Concurrent validity – is when teachers predict on a criterion at the same time when the test is being taken. For example, if the student is taking an exam the teacher creates a criterion about that test bases on test scores. (Glenn & Fred, 2007)

Criterion-oriented validity is a combination of scores and predictive criteria. Henceforth, the combination is created about the future or at the same time when a test is being taken based on the classification that is created and explained above.

2.Content validity - is often known as logical validity and definition validity (Newman, Brow, & McNeely, Conceptual statistics for beginners., 2006), emphasizing how different instrumental items seek to measure students' achievement in a specific course. It helps teachers focus on the key concepts dealt with throughout the course, resulting with very accurate measurement of the most useful terminologies and topics when students subdue the assessing process. According to Scriven (1967):

> Evaluation itself is a logical activity which is essentially similar whether we are trying to evaluate coffee machines or teaching machines, plans for a house or plans for a curriculum. The activity consists simply in the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical rating. (Scriven, Tyler, & Gagne, 1967, p. 40)

Despite of the fact that it just deals with gathering and combining information, the difficulties are inevitable if the teacher has not found the purpose and key concepts of the test. It is considered like the most influential sequence in education, which can be made at different levels and for many different purposes, thus, to reach the goal teachers should take into consideration the content they have been using through the completion of the specific course. Teachers' judgments are the most important activities for pupil learning outcomes (Lundahl, 2011) . Their judgments play an important role in determining grades and identifying pupils with disabilities or those in need of special assistance. Relating the purposes and structuring a test appropriately, teachers should take into consideration the main important type of assessing that is content validity, as one source of evidence that gives us the opportunity to discuss on the measurement a test is required to do (Johansson, 2013). Thus, identifying the purpose of the test facilitates teachers' job when it comes to compiling a test. In order to use a test to describe achievement, we must have evidence to support that the test measures what it is intended to measure. Without evidence of content validity, we cannot have confidence in these inferences (Johansson, 2013).

Content validity tests serve like observing points of students' achievement, but at the same time teachers' completion of syllabus outcomes for the whole course. Content validation is a multi-method process (Haynes, Richard, & Kubany, 1995). Based on a qualitative stance, content validity provides oral indication of consensus by experts in the content area (Newman & McNeil, Conducting survey research in the social science, 1998).

Content validity test can be quantitative, qualitative and quantitative/qualitative according to Ridenour and Newman, who state that mixed methods approach to content validity is based on an inductive–deductive philosophy (Ridenour & Newman, 2008)which can be represented

visually by figure 1, referring to the relation of methods leading to compilation of a content validity test:

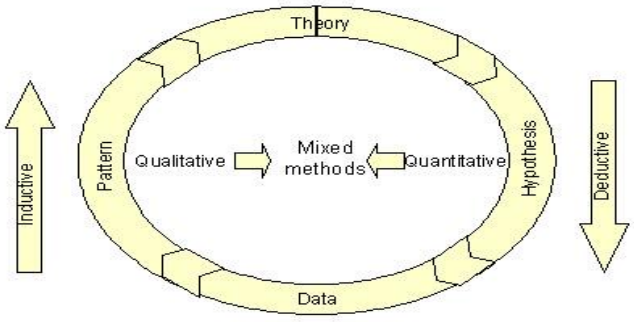


Fig.1 Mixed methods approach in content validity tests (Ridenour & Newman, 2008)

The mixed methods approach are aligned very closely to those strategies that are going to be treated and explained further on (Newman, Lim, & Pineda, Journal of mixed methods research, 2013) The validity of tests is often related to the content oriented strategy, as the best way to prove the outcomes brought by results. Its methods focus on content relevance and content representation ((Stelly, 2007).Content validity is established by showing that the behaviors sampled by the test are a representative sample of the attribute being measured and it is depended on both the test and the method of responding to the test.

3. Construct validity – is a concept used extremely frequently in our daily basis, but not the same meaning when used in testing. Teachers might think they all know the meaning, but in fact is much deeper. Traditionally, CV has been ordered the third and analyzed separately, same as the two other types of validity. The concept fluency is known to everyone, but in construct validity it becomes construct, thus can be related to something measurable or when the concept becomes 'operational'. For example, a teacher might measure students' fluency of English by creating a game and set scoring points and criterion for the accomplishment of that game, creating the chance for CV to take place in the process of evaluation. (Glenn & Fred, 2007) The new unified perspective, which derived from Samuel Messick, ranks construct validity the first, based on the importance of validity types, in contrast to the traditional

perspective. Messick's view was criticized, but has also been accepted all over the world, in the process of education, particularly, testing and assessment. According to Cronbach and Meehl:

> Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are means of confirming or disconfirming the claim.
>
> (Glenn & Fred, 2007, p. 8)

Their assumption makes everyone understand Messick's framework, which has construct validity in each column and row. Messick made clear that CV is multifaceted and is a measuring instrument of the variables a teacher would like to test. Testing fluency of English language requires construct validity, to verify whether the way of testing a teacher is using measures in reality the concept or content mentioned above. It is related to generalization and operationalization of concepts. Furthermore, to address crucial issues in the terminology of Construct Validity are highlighted six very important aspects: content, substantive, structural, generalizability, external and consequential aspects. (Messick, Validity and Washback in Language Testing, 1996) These all function as standards of construct validity, according to Messick:

- The content aspect of CV – it includes evidence about the representatives of quality. Crucial point of the content aspect is the determination of the boundaries to be assessed, that are different skills, which will be revealed by assessment tasks.
- The substantive aspect – is about theoretical rationales, related to the performance, offering evidence that the theoretical processes are linked to the respondents in the assessment. Two important key points of substantive aspects are very important in construct validity, such as: a) providing appropriate sample tasks domain process, in addition to the content domain.
- The structural aspects – it refers to the fidelity, praising it about the score scales to the structure of the construct. According to this aspect, the construct domain has to be relevant with scoring models and not guide only the selection and construction of it.
- The generalizability aspect – refers to the examination of the extent to which scores interpretations generalize to and across groups of population, settings and tasks. The duty of this aspect is to provide a representative coverage of the content and process by performance.
- The external aspect – is about discriminant and convergent evidence from multi method comparisons. It refers to the correlation of assessment scores with other measuring instruments.
- The consequential aspect – praises the implications of score interpretation. It provides evidence for evaluating consequences of score interpretations. (Messick, Educational Measurement, Third Edition, 1989)

Those six listed aspects of construct validity are usable in the process of evaluation and assessment particularly, including performance assessment and other possible ways of assessing a specific domain. Those provide better chances to address all the multiple questions that are necessary to be answered in justifying the interpretation of scores or results in general, as it is also stated according to Messick that “the relation between the evidence and the inferences drawn that should determine the validation focus”. (Messick, Educational Measurement, Third Edition, 1989)

2.5.2 The Impact of validity in assessment

Validity - the process of appropriateness and adequacy of actions on the basis of test scores, supported by empirical evidence and theoretical rationales according to Messick, interlocks in one main classifications, which involves in two others, such as: Construct-related, content-related and criterion-related validity. (Messick, Educational Measurement, Third Edition, 1989) Following the procedures of testing based on validity facilitates teachers' work, giving them the possibility to reflect on their work and the details that should be changed in the future. Validity is seen as a unitary concept, meaning that all those examining and observing a test come up with the same opinion about the scoring results, which can be ranged from low to high.

This component is helpful and necessary when compiling tests, unquestioning validation. Even students are going to better and easier understand the accuracy of their answers and based on the criteria the results are going to be easily guessed. It is a complex procedure, as a close relationship exists between the inferences and assessment, but it is an achievable process and ongoing process to realize the intended goal to be measured.

Testing has not been taken seriously and teachers most of the time worked on their own, compiling tests without any common framework. They compiled tests based on their opinions, so validity performance was not mentioned and the assessing system resulted poor until validity takes place in testing. Assessment, as a difficult process, needs proves to demonstrate mistakes in the process of evaluating and testing has an ethical dimension that influences people's lives, thus fairness is crucial in assessing students' knowledge and deals with the consequential validity. Fairness in testing is a duty of teachers while testing, not an ideal of assessment. (Cyril, 2005) Moreover, fairness is a key concept in the procedure of validation and through it accuracy and appropriateness is achieved.

Validity raises teachers' accuracy and helps them gather evidence on scoring results throughout the academic year. The relationship of testing and teaching has got difficulties and problems in the field of TEFL, as a result of lack of training programs for ESL teachers, in regard to assessment. A common national framework is extremely needful, in order to fairly and similarly teach and assess students' learning. (Michael, 1988) Validity is not just used for interpretation

inferences about test scores, but also about inferences made on observations of different attributes. Its impact is huge, as it offers enough evidence on test scores interpretation. This is all achieved through the process of gathering enough data about students' skills, respectively including the types of validity. The classification is done in purpose of help to teachers and the whole system of education. Validation is a procedure that interlocks in all the Validity types and all the stages in which the examination goes through. It must be remembered that the classical way of testing students' achievement, based on teachers' own perspective of fairness and formulation of tests, can be considered extremely poor. Hereto, assessment is complex and requires a process that verifies its accuracy, appropriateness and fairness, which now is being better achieved and realized by using standardized tests.

Assessment looks for a common framework of testing. Consequently, CEFR helps teachers assess identically all over a specific region and facilitate their effort on fulfilling the outcomes planned at the beginning of each academic year. As a result, the up mentioned procedure proves the accuracy of assessment in general. (Glenn & Fred, 2007)

2.5.3 Benefits of validity in assessment

The traditional way of assessing students was poor, but is getting advanced each year. Even the ministry of education is highlighting the credibility of its importance in education. Students need to know what to expect when being examined on a specific subject, in this case, specifically, foreign languages. Teachers have to focus on each important stage of learning a language and having a common framework of assessing their students' progress and reflect on their work is extremely beneficiary at the same time. Validation helps the teachers gather enough evidence throughout the whole year and gives them the opportunity to improve their way of assessing students' knowledge yearly. It also gives them enough arguments to prove the work they have done and also what they have achieved. Validity is complex, but when put to work is easily manageable, it just looks for everyone' effort for better results in the assessing process. Even the students know what to focus on and what to expect when being tested and assessed. (Glenn & Fred, 2007)

The accuracy and appropriateness of assessment and testing is laid on validation process in general. Even though, linguists share different views about a certain importance level of Validity types, but their opinion about its impact in the process of assessing students' progress is identically the same. Messick (1989) highlights that construct validity can be researched from different perspectives and the questions should be paraphrased, (Messick, Educational Measurement, Third Edition, 1989), whereas Chapelle (1999) provides different views of language testing, emphasizing the acceptability of more than one answer for the question "Does the test measure what is supposed to measure?". (Chapelle, 1999) Another similar view to Messick's is the one of Bachman and Plamer (1996) who consider validity as the crucial point on which tests should be judged, incorporating the view of Messick about construct validity and adding extra details on developing tests. According to them, validation is an ongoing process which can never be totally completed and its usefulness consists of six qualities, including practicality and authenticity. Their model serves as a plain sketch of validity impact and importance and the advancement of tests in general ((Bachman L. F., 1996), meanwhile Messick's progressive matrix is considered difficult to functionalize, but regardless it has been widely accepted with all its classifications, emphasizing the fact that validation is directly related to the interpretation of scoring. The views of the linguists mentioned above made the process of assessment more accurate and appropriate, raising students' awareness toward the process of being assessed itself and motivating at the same time to have a greater engagement in learning a foreign language, in this case EFL.

Chapter III

Research design and methodology

Quantitative and qualitative methods are used in the process of gathering information. The realization of this study was accomplished by the compilation of surveys, class observations and a focus group.

3.1 Data gathering procedure

This study provides to readers enough data about the upraised topic. The elaboration of this study is conducted in purpose to offer the teachers clues on how fairness, appropriateness and accuracy can be accomplished in the process of testing and assessment. Henceforward, the collected data and information were used correctly and implemented properly.

The data were collected with a questionnaire as the instrument of quantitative method, as well as the focus group and class observation as the instruments of qualitative method. A hundred and fifty pupils participated in pupils' survey. It served as an instrument to identify what creates anxiety to them in an academic year, their favorite method when being assessed and how is assessment happening in their school based on their experiences and assumptions. Furthermore, it had enough instructions in the questionnaire to the participants, in order to make them feel free and give sincere answers in the questions, showing them clearly the main purpose of it. Ten teachers had to complete a survey as well, with clear and simple questions, which are going to be used for analysis of this study from a more professionally spectrum.

Focus group, which was made of ten teachers, was created to gather much more information on assessment, in the new system of curricula, and discussing on the questions, which helped this study to have more accurate information and different opinions on the listed questions about the aim of this thesis. Meanwhile the class observations helped me identify the way teachers assess, students' knowledge on this specific domain, particularly, the impact of an inaccurate way of testing might influence students' motivation. The methods used here to

crown this study with conclusion made the work in it easier, as they gave enough data on the most difficult stage of education, the influence of validity in assessment and moreover the way how validation can be accomplished.

3.2 Participants/stakeholders

Students' survey was conducted in three different schools. The first one was conducted in "Skender Emerllahu" school, Ramjan, the second was done in "Bafti Haxhiu" school, Viti, and the last one was a private English language school "Oxford Education Corner" – Viti.

The participants of this study were a hundred and fifty pupils, eighth graders, and twenty English teachers for the survey and the focus group, ten per each method. Fifteen experienced teachers were women, who have master degree in English Language and five were men with less than ten years. The students' age is from 13 to 14, including both boys and girls. All the instruments used for this study have very clarified instructions in relation to the purpose of this study conduction.

Six classes were observed at almost the end of the semester, as the aim of the observations was to see the methods used in assessing students' achievement. Three observations were in "Oxford Education Corner", two in "Skender Emerllahu" and one in "Bafti Haxhiu" .

The permission for the observations was taken from the director of Education Department, in municipality of Viti, in accordance with those two schools' directors, Mrs. Vjollca Zejnullahu and Mr. Mustaf Aliu, whereas in "Oxford Education Corner" school the permission was taken directly from the management, specifically Ms. Mirjeta Rexhepi and Mr. Pajtim Alidema, who were very collaborative.

3.3 The study instruments

Pupils' survey

The survey conducted with the students was made of 8 multiple types questions. There were variations of questions; five of them were opened questions, letting the students answer based

on their own experience, two others were multiple choice questions and one question was dichotomous question, in which pupils had to answer with “yes/no”. This variety is applied based on the level of their ability to critically think on a specific issue.

The beginning has got instructions for the inquiry that is being elaborated and that the surveys are anonymously. The teachers from the three schools assisted students in case they could not understand any of the questions. The collected information was quite helpful for completing my MA thesis. It helped this study identify the most difficult area in education, teachers’ techniques in assessing students’ achievement and the focus of the study; the way how testing is accomplished and whether teachers knew the terminology that is going to be analyzed in this study.

At the end, each survey has a thanking statement for the time dedicated to this study and appreciation about the sincerity they have shown answering the questions. Without the help of the pupils and teachers, who become part of this work, my thesis would have been uncompleted.

Teachers’ survey

Teacher’s survey includes ten questions. It has some more advanced questions, answered by thirty teachers. Their answers were extremely valuable, as this study needed to be realized by comparing and contrasting teachers’ experiences. The variety of questions is applicable and present there as well; from ten questions, five are opened questions, looking for their own perspectives, one is multiple choice answer and four are dichotomous questions, yes/no or always/sometimes/never.

All the ten English teachers have their marks on my master thesis helped me by sincerely answering the questions. Their time and work on answering was extremely valuable and needful for the analysis. Its aim was to understand what the teachers struggle mostly with, the methods they use in testing and assessing their students’ knowledge, the compilation of the tests and moreover the whole procedure of it.

Focus group

The focus group is used as an instrument of qualitative method, to gather more information on the analysis that the thesis is aiming to achieve. It consists of seven questions. The main purpose is to collect information about this study, through interviewing the teachers at the same time and opening up a discussion about the topic.

The data that was collected from this interview was very valuable and quite useful. Hence, the teachers talked about the most difficult part in education, assessment methods and the procedure of testing and assessment.

Class observation

Observation as an instrument of qualitative method was done on the eighth grade of EFL pupils. Its aim is to observe the assessment methods and it happened at the end of the second semester. During the observation, some very important notes are taken related to the process of assessing and mainly the analysis have been made upon the personal diaries the teachers have used and continue using each academic year. As an advantage prior to the analysis for this study, is the fact that the analysis will be done in two public schools and a private one. The private school of languages uses standardized tests in assessing their students' performance and the grading scale is based on CEFR (The Common European Framework), in comparison to the public schools, where teachers compile tests on their own without any specific framework.

This observation was very useful because it helped me find the gaps and needs on assessing pupils' performance. I was able to identify the techniques that teachers use to evaluate the specific outcomes related to a certain topic.

Chapter IV

Findings and results

The findings are all based on quantitative and qualitative methods of gathering data on certain analysis, such as: Students and teachers' survey, focus group and class observations. Ten English teachers have been part of this study by completing the questionnaire, with the permission of observing six English classes at the end of summer semester, and ten others were part of the focus group. Students helped a more accurate completion of the study with their enormous contribution, which is irreplaceable.

The collected data, from all the instruments listed above, were the core of this study and its aim. The importance of evaluation process and its relation to validation is proved here with accurate and useful results. The responsibility of transmitting and assessing students' progress is the most difficult section in TEFL, as teachers try fairness and appropriateness of assessment based on the new curricula system and its determined outcomes.

4.1Students' survey

Students' questionnaire consisted of six questions. The questions were four multiple choice questions and one dichotomous. This questionnaire was distributed to one hundred and fifty students, showing their concerns about assessment in general. Its purpose was to prove the hypotheses – "Validity tests have a great and crucial impact in assessment." and "The accuracy in measuring the exact skills observed throughout a year influences students' self-confidence and awareness of their own achievement."

The answers are analyzed as follows:

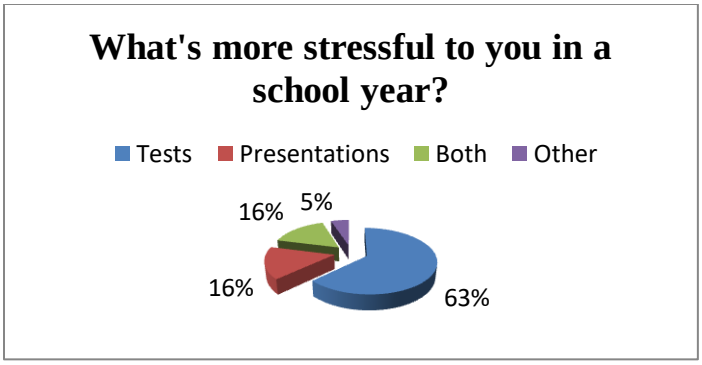


Fig. 2 – Question 1 of students' survey

According to this chart, the readers understand clearly that students are mostly stressful about being tested, which can be proved by the percentage that varies from 63% about tests, 32% shared equally on presentations and both, whereas 5% of the students showed other concerns related to the process of being taught by their teachers. Testing has its advantages and disadvantages, thus a fairly process of it makes it easier and better.

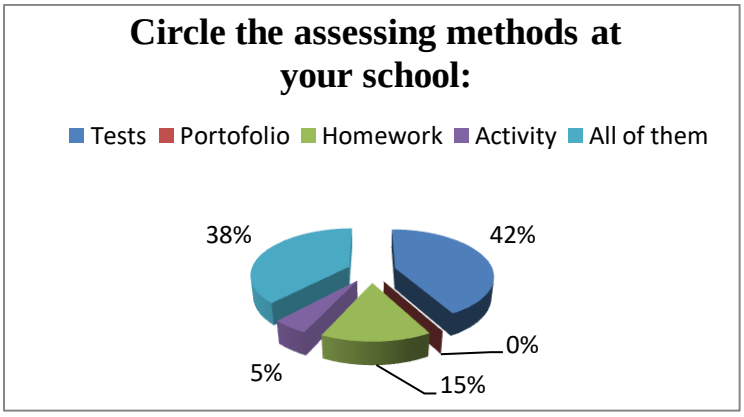


Fig. 3 – Question 2 of students' survey

Traditionally, testing has been one of the most used methods on assessing students' knowledge. Despite of the fact that the new system of education aims to change this, the second question on the survey makes the reader allude that tests continue being highly used in the process of assessment, with 42% based on the conducted survey, followed by homework with 15%, as another tool of classic teaching and 5% by activity. Meanwhile, 38% shows us that

the system of education is in transition of methodology, making the reader create positive thoughts about the new era of education. An immediate transition is almost impossible, but small steps towards that means a big shift in the process itself soon.

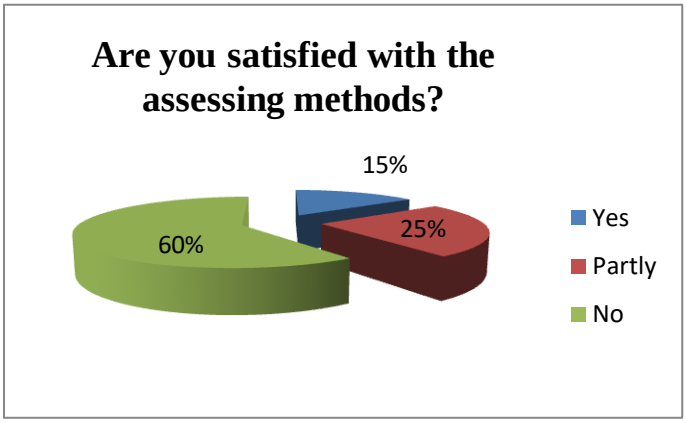


Fig.4 – Question 3 of students' survey

Students show a great non-satisfaction towards the methods used to measure their knowledge, based on the 60% of them, who answered negatively about the approaches used. Even the percentage of those who answered the questionnaire Partly (25%), and Yes (15%) cannot equalize the No answers. Relating this question to the previous one, it clarifies and proves their dissatisfaction about the assessing issue. One criterion evaluation is not relevant, thus teachers are supposed to change their teaching style and adapt it to the one required by the new curricula in Kosovo.

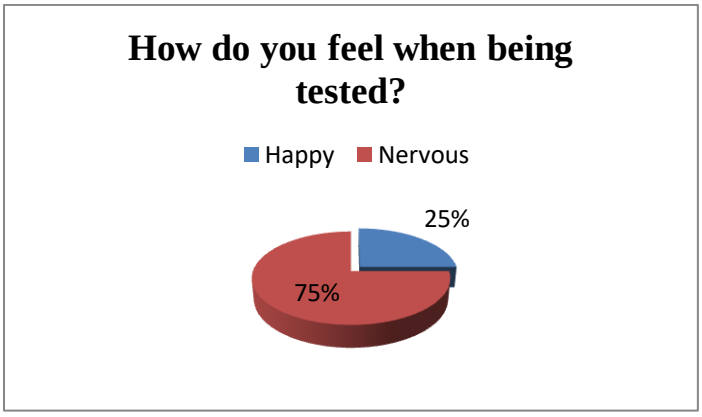


Fig.5 – Question 4 of students' survey

Accordingly, the chart demonstrates that almost all the students are nervous about being tested. Consequently, the percentage of 75% negative and 25% positively, gives allusion to readers about students' confidence, expectancies and preparations. Teachers are supposed to take into consideration even the way students' cultural and educational background when it comes to testing. The gathered information contributes in accuracy and appropriateness according to the procedure of validation.

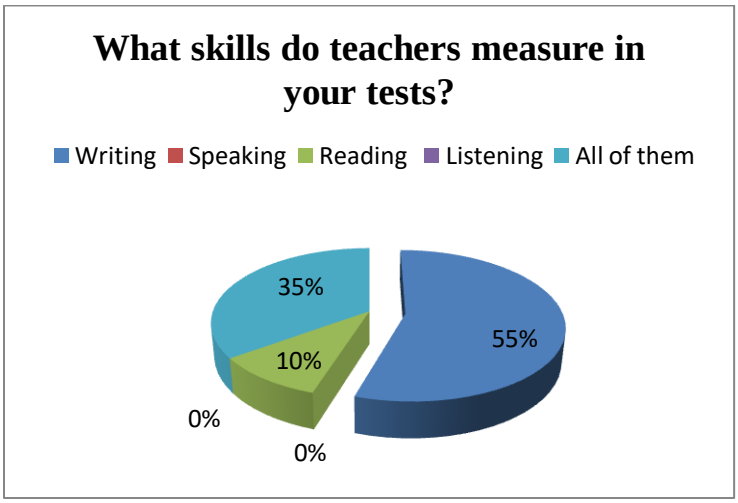


Fig. 6 – Question 5 of students' survey

The system of education in Kosovo is in transition, making teachers' adaptation obligatory. Henceforth, the process of assessment cannot have validation and functionalize it if teachers compile tests on their own. Teachers' training on standardized tests brings credible results if the ministry of education emphasizes the impact of this issue in TESL. Moreover, seeing the focus of teachers on figure 6 proves the hardships of teaching process, as a result of unprofessional compilation of tests by non-experts of ESL.

The chart shows the emphasis on writing with 55%, as a very old method of being tested, followed by reading 10%, meanwhile 35% on all of them. The percentage on the last option is from the private school "Oxford Education Corner", which uses the Common European Framework (CEFR) in grading students' achievement. The tests used in that private school are all standardized, in comparison to the ones used in public schools compiled by teachers themselves.

What do you think that should be changed in assessment?	
1.	Grading properly
2.	Having more criteria to assess us
3.	Focus more on what we learn in our classes and assessing us based on it
4.	Tests
5.	Online tests
6.	Focusing on just the subject that a teacher teaches, not on other grades as well.
7.	Not to have two or more test on the same day.
8.	Easier tests
9.	Not having open questions on tests.
10.	Having a multiple choice test.
11.	Giving us the chance to improve our grade by speaking.
12.	Not more than two tests on a period in the same subject.
13.	Not having tests, as they make us feel stressed.
14.	Nothing needs to be changed.
15.	Everything is okay.
16.	Group work.
17.	Having tests, but in ways that make us feel entertained and not stressed.
18.	Portfolio a good way of being assessed.
19	Teachers should assess us by project based learning.
20.	Talking to us about testing time.

Table 2 – Question 6 of students' survey

The intent of this questionnaire was to reveal the impact of testing in stduents' brain. The results above show that testing makes students feel anxious and not very confident about their knowledge. Moreover, more criteria of grading students achievement elevate their aptitude and contributes to the effectiveness of their learning process and better performance. Most importantly, organizing the course syllabus with students' assumptions on what is more beneficiary to their engagement is primariy ranged in the process of ESL. Hereto, the discussion about course syllabus, its possible grading criteria and their momental emotional state raises students motivation towards learning and better achievement, which leads us to the process of validation. Lastly, teachers preparation in compiling standardized tests to provide validation

includes: Students' emotional state at the moment of testing time, their cultural/educational background and their necessities. Thus, the hypotheses "Validity tests have a great and crucial impact in assessment" and "The accuracy in measuring the exact skills observed throughout a year influences students self-confidence and awareness of their own achievement" are truly realized. English is part of each student life, as the edcational system of Kosovo requires students to start acquiring English since the early stages of life. Therefore, the teaching and assessing process need to be fairly and properly done. Students' variety of answers, give to the readers different spectrums of assessment that need shifts. One skill based tests are useless, thus testing the achievement in a language looks for testing the four basic skills, which are related to standardized tests.

4.2 Teachers' survey

The questionnaire was conducted in with three in three schools, two public, primary ones - "Bafti Haxhiu" and "Skender Emërllahu", and a private language school in Viti, "Oxford Education Corner". Ten English professors were part of this survey, which was consisted of eight questions, four dichotomous and four multiple choice questions. This questionnaire was done in order to get proper information if English teachers were familiar to validity notion and standardized tests. In addition, the aim of the questionnaire was to prove the hypothesis – "Validity helps teachers reflect on their own teaching effectiveness."

The answers of this survey are analyzed as follows:

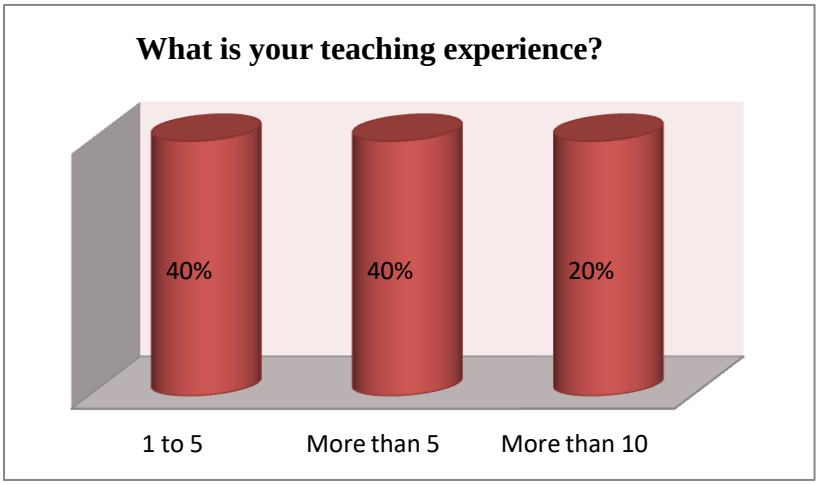


Fig. 7 – Question 1 of teachers' survey

In the first question, English teachers had to write their teaching experience. All ten teachers involved in this study had experience in teaching students and their collaboration helped the realization of the analysis, with the greatest impact in all its elaboration. Two teachers had more than 10 years of experience of teaching, meanwhile four from the others had one to five and four others more than five.

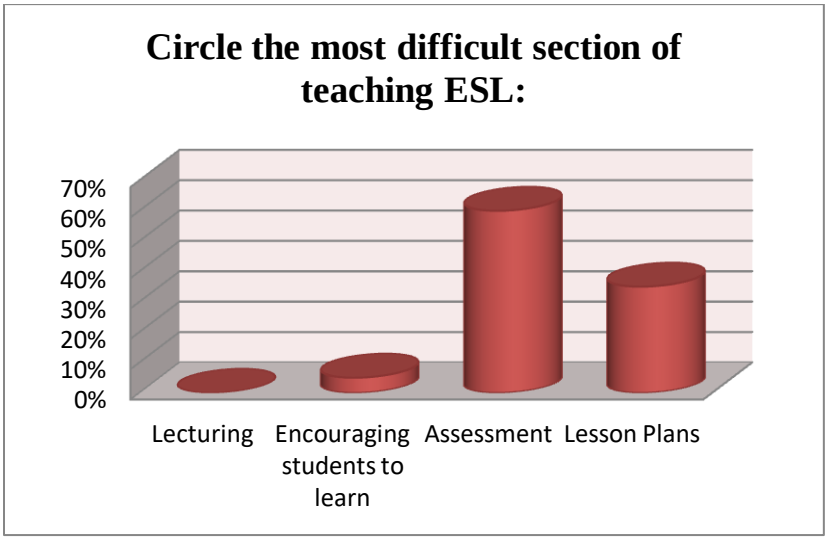


Fig. 8 – Question 2 of teacher's survey

The transition of education has made the process of teaching even more difficult than it was before. In Kosovo the new curricula requires a totally new way of teaching students and assessing them. Assessment has always been considered as the most difficult section in teaching, as it is directly related to the accuracy and appropriateness of spreading and transmitting information. The chart about the second question of teacher's survey, proves that when it comes to difficultness, assessment takes place with the highest percentage 60%, being followed by lesson plans and with just 10% comes the encouragement of students to learn.

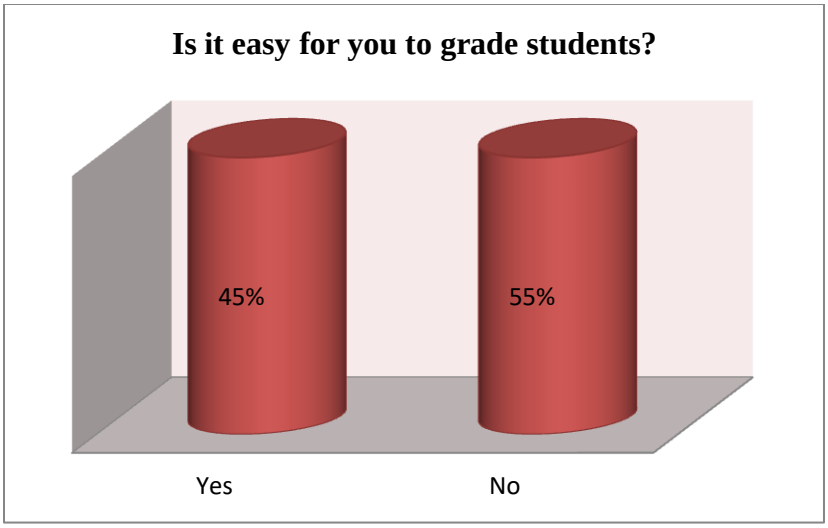


Fig.9 - Question 3 of teachers' survey

The responsibility of whether you graded students fairly or no, burdens all teachers, which is shown from the answers taken on the third question. Grading and assessment is not an easy process based on the answers, thus having a common framework for grading and standardized tests for all, will bring better results, more accuracy and validity in assessing.

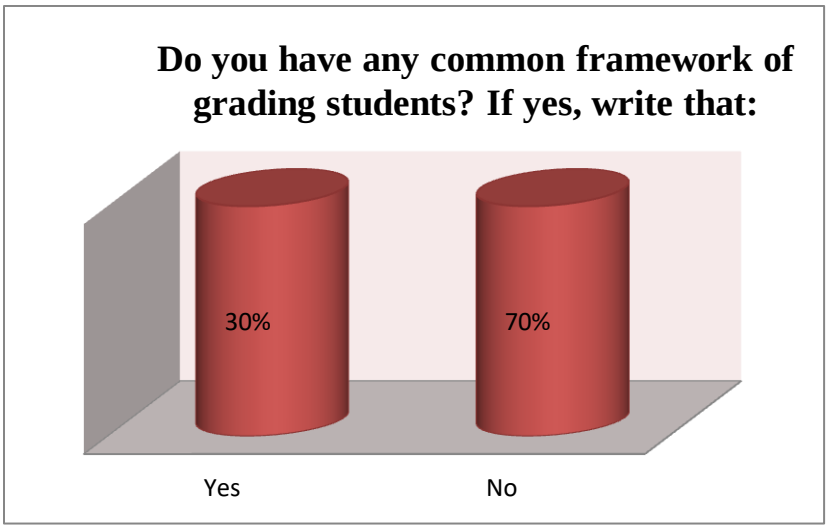


Fig. 10 – Question 4 of teachers' survey

The framework of grading students:
1. Common European Framework
2. CEFR
3. CEFR

Table 3 – Question 4 of teachers' survey

Almost all the teachers were from the public schools and even the results show that, with 70% No and 30% Yes. The teachers, who answered positively about the policy of grading students, are part of the private school. Additionally, they use CEFR as a framework for diagnosing students' knowledge in second language learning, which facilitates teacher' assessing process easier. The teachers were asked to mention the framework they base their grades on and three of them mentioned CEFR. Some more details about this issue will be given on observation section, as two classes were observed on that school and the tests they use, as well. Moreover, this paper is going to offer a comparison of standardized tests and non-standardized ones. Standardized test are more useful when it comes to verifying the accuracy and appropriateness of tests through validation.

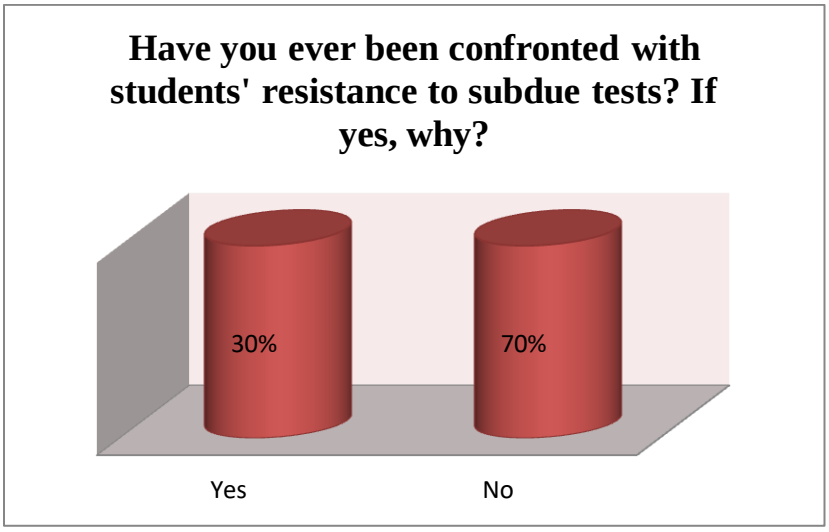


Fig. 11 – Question 5 of teacher's survey

It is important to remember that the students' anxiety in tests happens to make them resist subduing a test, but based on the results elicited from the fifth question, we can conclude that teachers did experience students' resistance to take a test. The percentage shows clearly that 70% of them never experienced it, whereas 30% experienced the resistance. Students resistance can happen as a result of unfairness, an issue that can be regulated by standardized tests and moreover measurement and accuracy of the test.

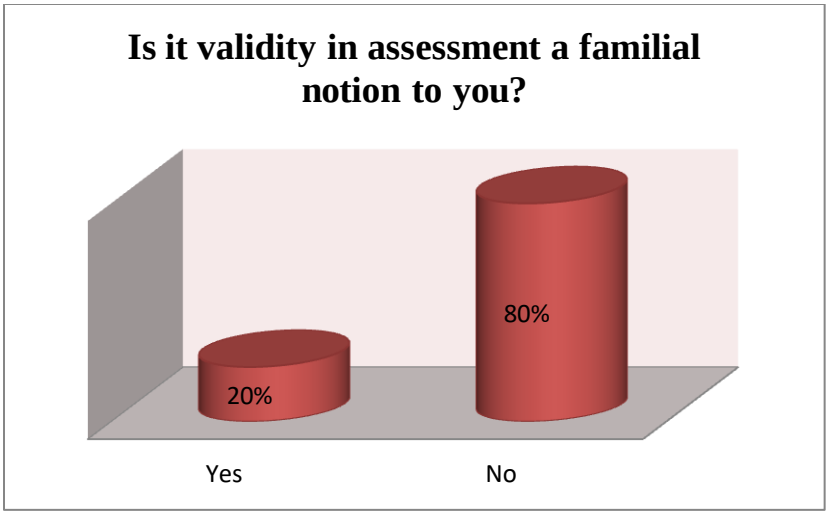


Fig. 12 – Question 6 of teacher's survey

Validation helps the teachers collect necessary information and proves about the work of each student, giving them the opportunity to reflect on their own way of teaching and assessing the information transmitted to the students. It is related to formative assessment, as it lies on observation. It is quite complex, but when put to work is easily manageable, it just looks for everyone' effort for better results in the assessing process. Through validity tests, even students know where to focus on and what to expect after a test, (Glenn & Fred, 2007) but based on the chart, despite of all the advantages that validation has in assessment, teachers are not familial to it as a notion 80% of them, do not have any idea about the meaning, whereas just 20% knew it as a notion.

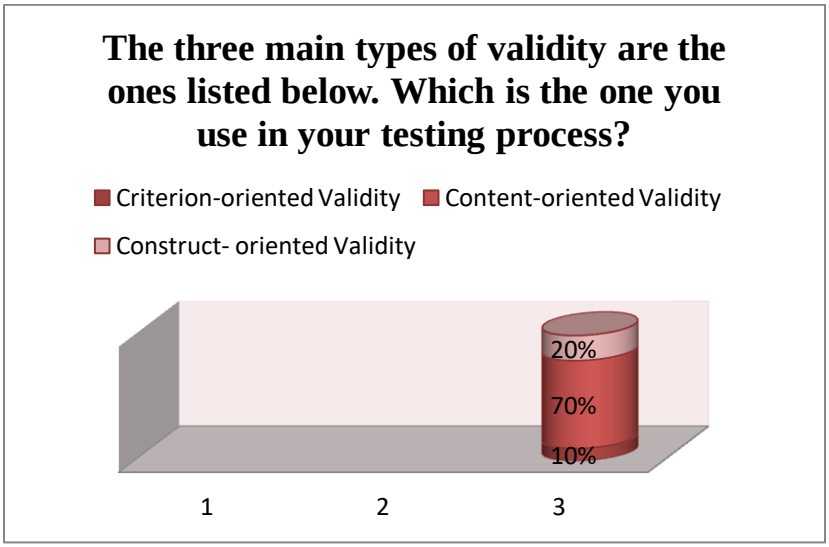


Fig. 13 – Question 7 of teachers' survey

Based on a brief description about the main types of validity, on the questionnaire, teachers had to write their assumptions on the test they use. They related their testing process mostly to content-oriented validity, with construct-oriented then and finalizing it with criterion-oriented validity.

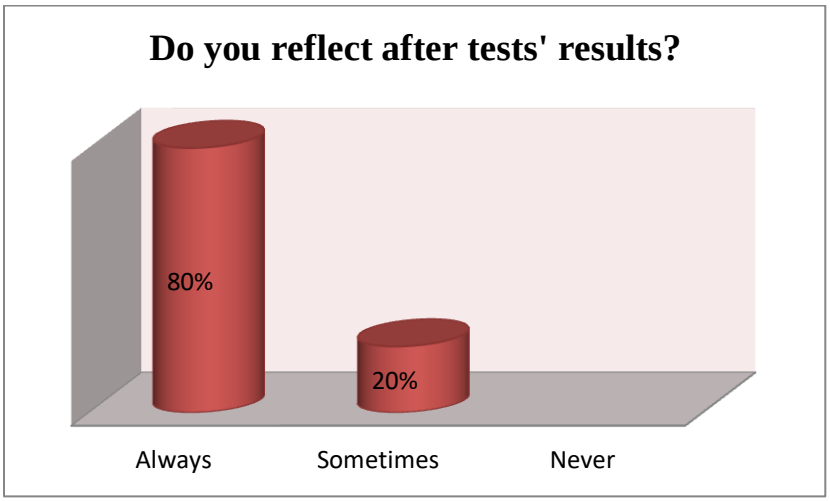


Fig. 14 – Question 8 of teachers' survey

The new educational system has brought freshness in teaching and learning ESL, as almost all teachers reflect on test results after taking them, with 80% answered positively and 20% with suspicion whether their reflection is accurate or not. Validity is related to reflection, as accuracy and appropriateness cannot be achieved in any other way. Reflection helps teachers improve the results of the intended material or methodology shift if necessary. Reflecting means the collection of evidence and improvement of the effectiveness teachers' own teaching.

The goal of this questionnaire was to figure out the impact of reflection upon their own improvement. The hypotheses "Validity helps teachers reflect on their own teaching effectiveness" is proven and frequently realized after testing students knowledge. It was easily achievable in "Oxford Education Corner" school because of the fact that they use standardized tests and CEFR in grading students' achievement. Their tests help them effectively reflect on their teaching methodology and its weaknesses. The teacher in the public schools reflect and see reflection very useful in advancing themselves professionally, but the accuracy is low rated

in comparison to the school that uses standardized tests. Additionally, in accordance to English teachers' perspectives, assessment is the most difficult section of teaching ESL. Compiling tests on their own is quite challenging, thus professional advancement is highly required.

Furthermore, English teachers need to make teaching creative, in order to attract students' attention towards learning the intended material. Consequently, the discussion on testing time in an academic year is cruically in the process of motivating and raising their confidence towards their capabilities. Likewise, teachers to engage and talk to them, making them know their importance and the relief they bring in the process of teaching and assessing when are par of decision making. Students' satisfaction with grading results establishes a total new perspective to the way the see the process of learning ESL.

4.2Class observation

This study contains six observed classes, which had the focus on the assessing instrument that teachers used at the end of each class:

Three classes observed at "Oxford Education Corner" school:

Number of Students	15	10	20
Level	Elementary	Intermediate	Beginner
Date	12th of June 2019	13th of June 2019	13th of June 2019
Time	3:30 p.m	2:00 p.m	3:30 p.m

Table 4 – Observations at "Oxford Education Corner" school

The focus of observation was assessment, so the observations happened on testing day in two classes and a regular class the last observation. While the students were being tested, the tests were checked and compared to the material that has been covered.

Another observation happened on a regular class. The teacher greeted the students warmly and started the class with a warm up activity, which served as an assessing instrument on

recalling the words learned at the other units (guess the word). After that, together with the students, they elicited the topic and started working on it altogether. At the end, the teacher used some practice sheets taken from an extra book related to the books they use daily, in order to measure the comprehensiveness of that topic.

Two more classes observed at “Skender Emërllahu” Ramjan.

Number of Students	24	17
Grade	9th	8th
Date	4th of June 2019	6th of June 2019
Time	10:25 p.m	11:05 p.m

Table 5 – Observations at “Skender Emerllahu” school

Likewise, the emphasis of the observations in this school was assessment. The ninth graders had test and the validation of testing was observed, comparing it to the intended material and teachers’ attitude towards students’ assessment. Its form was classic, not standardized, but it met the requirement of measuring what it is supposed to measure, not all skills, just the writing one, but it corresponded to the units that they have learned. The teachers’ personal diary has been checked, in order to witness the criteria used to assess students in a specific period and the percentage per each of them. Henceforth, the criteria were just about writing and speaking.

The other observed class was a very warm one. The teacher had great relationship with the students and the class was students’ centered as well. It was about past simple, the students tried giving examples, assisted by the teacher, but at the end there was not any assessing instrument in order to understand whether the students understood what they learned.

Two classes observed at “Bafti Haxhiu”

Number of Students	20	26
Grade	9th	9th
Date	7th of June 2019	7th of June 2019
Time	9:35 a.m	10:20 a.m

Table 6 – Observations at “Bafti Haxhiu” school

Both the observed classes in this school had tests on that day. The teachers were different. Same as at the other schools, the focus was on the methodology of assessment, checked by observation and based on teachers’ personal diaries. The tests were not standardized and the diaries were with just four criteria, which measure the writing and speaking skills of students. The atmosphere was positive in the class, but the validation process was hard to be realized in that school because of the classic way of testing and assessing students.

The absence of a common framework on assessment and testing is in the responsibility of the Ministry of Education in Kosovo. Teachers need to collaborate more to each other, in order to compile more accurate and appropriate tests according to each grade. Students need to know what is more important in the supposed material. Another worth mentioning factor was the measurement of four basic skills of learning and speaking a second language, which is left aside in public school. In contrast to non-standardized tests, the samples taken from the private language school in Viti uses standardized tests, which gives more accurate evidence to teachers on reflecting and professionally advancing themselves in assessing students’ achievement. Students were comfortable with the procedure and emotionally relaxed, creating the impression of fair, accurate and appropriate testing methods. This instrument served in strongly proving the fourth hypothesis “Assessing the four basic skills of learning a second language can be realized by being accurate what and how to assess students’ achievement”, which brings even motivation and self-confidence to students.

4.4 Focus group

The focus group was made of ten teachers. They shared their opinions about the questions related to this study and their contribution is worth it.

Do you prefer more Formative or Summative Assessment?	
1.	Formative Assessment
2.	Both
3.	Formative Assessment
4.	Formative Assessment
5.	Both
6.	Formative Assessment
7.	Summative Assessment
8.	Formative Assessment
9.	Formative Assessment
10.	Formative Assessment

Table 7 – Question 1 of focus group

Nowadays, teachers are dealing more with formative assessment than summative one. They prefer observing and then crowning their results with accurate feedback about their achievement. This kind of assessment gives the teachers opportunities to improve themselves, in case of any mistake while grading students. To both teachers and students, this is beneficiary, as it gives them time to professionally advance themselves. FA makes students understand the importance of learning to know and not to be graded, developing the long lasting idea of learning throughout all our lives.

What criteria do you use for grading students?	
1.	Activity, homework, reading, tests
2.	Participation, homework, reading, speaking, tests, projects, etc.
3.	Participation, homework, reading, speaking, tests, projects, listening, etc.
4.	Tests, activity and homework
5.	Tests and homework
6.	Reading, speaking and tests
7.	Homework and tests
8.	Homework, tests and speaking
9.	Reading, speaking, tests and homework
10.	Participation, homework, presentations/projects, reading, speaking, etc.

Table 8 – Question 2 of focus group

Nowadays, teachers are supposed to adapt their teaching methods based on the requirement of the new curricula here in Kosovo, thus more than three criteria are required to observe students' achievement. The criteria listed up there, based on those ten teachers' answers, needs some improvement, as there are still teachers, who use just two criteria to assess students. Moreover, just two of the teachers include the assessment of the four basic skills of learning a second language, meaning that teachers need to reflect and adapt their diaries based on the inquiries. Testing the four basic skills makes students feel more confident, as they have the opportunity to lay on their strengths and not always all of them can be their strengths.

Do you have ready- made tests or you formulate them on your own?	
1.	We have ready-made ones.
2.	We formulate tests on our own.
3.	We formulate tests on our own.
4.	We formulate tests on our own.
5.	We formulate tests on our own.
6.	We formulate tests on our own.
7.	We formulate tests on our own.
8.	We have ready –made ones.
9.	We have ready-made ones.
10.	We formulate tests on our own.

Table 9 – Question 3 of focus group

The transition of education in Kosovo is also related to standardized tests, which should have been delivered to teachers, but yet not done. Teachers working in the public schools do not have ready- made tests, whereas those working in the private school have standardized tests, as the school provides them for the teachers. Validation process occurs in "Oxford Education

Corner" school, but not in the public schools. The Ministry of Education needs to reflect and work on fulfilling the gaps that education has in Kosovo, moreover needs to provide worthwhile trainings related to assessment for the teachers. Assessing fairly and properly is crucial in teaching and learning ESL.

Have you ever heard about Validity? What are your needs in regard to assessing students' progress?	
1.	No, never. Trainings on standardized tests.
2.	No, I have never heard about it. We need to know more about assessment.
3.	Yes, it is about the accuracy of what a test measures. More trainings on compilation of tests.
4.	No. We need to know more how to become fair to them.
5.	No, never.
6.	Yes, I have heard. Trainings on methodology.
7.	I have never heard before. Methodology trainings.
8.	Yes, I have. Not sure about its meaning.
9.	No. More trainings on assessment specifically.
10.	Never before. Standardized tests trainings.

Table 10 – Question 4 of focus group

The youngest teachers, who finished master degree in English Teaching, knew the terminology related to validity, but not the most experienced ones. This gap is because of misinterpretation of the term of Education in general, which has not got enough scientific research by our experts in the field of it. Teachers, who knew about it, have ideas about the procedure of realizing it and the way it can be achieved, but this is not sufficient, as from ten teachers just three of them were familiar to the notion and its meaning. Teachers need to know more about the procedure of validation and how to compile accurate and appropriate test, bringing better results and fairness in evaluation.

Chapter V

Conclusion

As it has been noted above, teaching and learning process requires the greatest efforts from teachers, to transmit and bring results to the teaching and learning ESL. An inseparable part of it is assessment, which is the focus of this study, as the most difficult section of teaching. The classification of assessment in the new system of education has brought differences to the whole process. Hence, teachers need to understand and identify the importance of each kind in the process of testing and evaluating students' achievement. Summative and formative assessments are the ones mentioned at the beginning, which are supposed to be used in assessment. The former is related directly to the grade, whereas the second one requires more time to give opinions about students' achievement, meaning that it is related to observation.

Nowadays, assessment relies more on FA, in order to have more assessing criteria mentioned throughout this study. Students are consistently being assessed, in school and sometimes for governmental purposes. Furthermore, teachers are responsible to make accurate and appropriate tests to show exact results, which can best be measured and completed based on the determined outcomes. Two more crucial concepts related to assessment are: Reliability and validity. (Glenn & Fred, 2007) The latter mentioned and its impact in evaluation is the one analyzed in this study, as it is the most important concept in the process of assessing students. Validity is very important and teachers who try compiling tests based on the rules 'How' and 'What' will raise students' awareness on the importance of lifelong learning, equipping the students with specific expectations in regard to their achievement results.

The main hypothesis set up at the beginning of this study, – "Validity tests have a great and crucial impact in assessment", has been proven by all the instruments used to gather data, concluding on the necessity of a common national framework to fairly and similarly teach and assess students' progress. (Michael, 1988) The comparison of the public schools and the private school, made clear the distinction of standardized tests from the ones made by teachers, without being trained on standardized ones. Validation is a procedure that interlocks in all the

validity types. Henceforth, all its stages look for a common framework in testing, in order to be identically all over a specific region and facilitate teachers' effort on fulfilling the outcomes planned at the beginning of each academic year. Validation is evidenced in case all the required details are taken in consideration when diagnosing students' achievement.(Glenn & Fred, 2007).

The 2nd and 3rd hypotheses – "The accuracy in measuring the exact skills observed throughout a year influences students' self-confidence and awareness of their own achievement." and "Validity helps teachers reflect on their own teaching effectiveness", were possible in "Oxford Education Corner" school. Consequently, standardized tests help teachers assess students' achievement easily and validation takes place inevitably, resulting with a high level of accuracy in measuring the outcomes for that specific level. On the other hand, teachers working in the public schools, compile tests themselves without any common framework to grade students' achievement of ESL. After taking the results teachers were able to reflect on their own work based on students' weaknesses showed in testing and helped them focus on what is mostly needed. In that private English language school, the material corresponds to each kind of test, such as: achievement, diagnostic, unit, skills tests, resulting on a high level of accuracy in measuring the outcomes determined for that specific level.

At last, the 4th hypothesis – "Assessing the four basic skills of learning a second language can be realized by being accurate what and how to assess students', has also been proven in accordance with the standardized tests and testing results. In contrast, the inconvenient compilation of tests provides results, but not including accuracy and appropriateness in the process.

Summing it all up, the hypotheses raised at the very beginning of this study are proven by the instruments used and especially from the comparison made in public and a private school of languages. The research updated the necessity for teachers' professional advancement in assessment, specifically in compilation of standardized tests. This study offers solution to teachers about fairness, accuracy and appropriateness of testing. Furthermore, some recommendations are listed to help reader easily conclude on the emphasis of the thesis. Teachers did not have the adequate information about validity, so testing resulted poor. It

offers proves and comparison of the usage of a common framework about grading with standardized tests and non-standardized tests.

5.1 Conclusion from the surveys

The concept of standardization and non-standardization were not familiar to all the teachers. In spite of the fact that tests are commonly used in teaching, they are considered difficult for both teachers and students, in different spectra. The duty of teachers is to offer appropriate and accurate tests corresponding to the syllabus of ESL, but the analysis done in three different schools, a private and two public, proves the opposite. The private school uses standardized tests and offers levels according to the Common European Framework of learning a foreign language, known and accepted all over the world. Meanwhile, the public schools work without a specific framework of grading students and even the compilation of tests is on teachers' responsibility, with no help about what should a test include to measure the four skills of learning a second language. Those surveyed students of "Oxford Education Corner", answered that the test measures all their basic skills, whereas at the public schools, teachers have three or less criteria to assess and the tests are not compiled based on the standardized tests. The process of validation can be achieved just when the tests meet the standards of fulfilling the outcomes specified about learning a foreign language. Validation has the strength to make teachers' job easier and students' learning process less stressful, but to reach the requirements this study is going to write some recommendations at the end.

5.2 Conclusion from the observations

Six classes observed brought different views about the needs that teachers and students have to bring effectiveness in evaluation. The collaboration between students and teacher at the beginning of each school year is necessary in order to have a fair and satisfying testing time. The comparison of the school working with standardized tests is huge with those ones, where teachers make the tests themselves, trying to do their best and bring the results they have been

working for. Standardized tests make a big difference in assessment, as working becomes easier and teachers are responsible about the validity of testing system.

5.3 Conclusion from the focus group

Teaching requires responsibility for what you serve to students, thus being accurate and aware of students' needs brings better results. The teacher discussion during the focus group, revealed the gaps left in one of the first developmental fields of each country. Teachers grade students based on their own assumptions of accuracy and fairness, making tests and processing the whole crucial procedure of educating students themselves. Markedly, their information towards measuring students' achievement is not necessary, including here the procedure of validation as well. The validation procedure has to take into consideration students' need to teach ESL, their involvement in deciding about the perfect timing to be tested and collaboration between teachers and students. Almost all the teachers showed their personal diaries that contained three, four and even less than, with 20% of them who take into consideration all the instrument that are needed to test someone skills on learning a second language. Therefore, most of teachers' opinions show that trainings in assessment are necessary. On the other hand , some of them think that methodology training are also welcome. In conclusion, both trainings are useful in the process of teaching ESL.

5.4 Recommendations

This thesis provided important data on testing appropriateness and accuracy, in regard to validation procedure. The outcomes taken from the used instruments in this study pointed out that testing needs to focus on the intended material, but its process takes into consideration even students' emotional state. Therefore, listed below are some of the recommendations taken from the results of this research, such as:

For the Ministry:

- The Ministry of Education needs to decide on a specific framework that can be used in schools about grading scales.
- Trainings on standardized tests.

For the teachers:

- Testing the four basic skills in assessment.
- Students need motivation about the process of learning, which can be realized by making standardized tests and giving them the opportunity to know what to expect after subduing it.
- Syllabi should be discussed at the beginning of each academic year, with students' collaboration about specific details related to their decisions.
- Teachers should use contemporary techniques to compile tests, even if they are not trained for compiling standardized ones, in collaboration with other teachers of the same region.
- Teachers need to reflect daily on each class completion and results.

Bibliography

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University.

Bachman, L. F. (1996). The use of Test Method characteristics in the content analysis and design of EFL proficiency Tests;. In *Language Testing* (pp. 125 - 50).

Brown, H. D. (2006). *Principles of Language Learning and Teaching* (5th ed.). New York: Pearson Education.

Carmines E. G, &. Z. (1991). *Reliabillity and Validity Assessment.* Sage: Newbury Park.

Chapelle, C. (1999). Annual Review of Applied Linguists. In *Validity in Language Assessment* (p. 19).

Claudia, F. (2020, November 14). *Validity in Assessment: Content, Construct and Predictive Validity.* Retrieved from Study.com: https://study.com/academy/lesson/validity-in-assessments-content-construct-predictive-validity.html

Cyril W., C. (2005). *Language Testing and Validation.* New York: Palgrave Macmillan.

Cyril, W. (2005). *Language Testing and Validation.* New York: Palgrave Macmillan.

Glenn, F., & Fred, D. (2007). *Language Testing and Assessment.* New York: Routledge.

Harmer, J. (2001). *The Pactice of English Language Teaching* (3rd ed.). Essex, England: Longman.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). *Content validity phsychological assessment. A functional approach to concepts and methods.*

Hyland, K. (2003). *Second Language Writing.* New York: Cambridge University Press.

Isadore Newman, F. P. (n.d.). *Content validity using mixed methods approach; its application and the use of the table of specifics.* Florida International University.

Isadore Newman, J. L. (2013). *Journal of mixed methods research.*

Johansson, S. (2013). *On the validity of reading assessments.* University of Gothenburg.

Krashen, S. D. (1982). *Principles and Practice in Second Language Acquisition.* Southern California: Pergamon Press Inc.

Larsen-Freeman, D. (2000). *Techniques and Principless in Language Teaching* (2nd ed.). (W. E. Russell N. Campbell, Ed.) New York: Oxford University Press.

Lundahl C. (2011). *Assessment for learning.* Stockholm: Norstedt.

Lundahl, C. (2011). *Assessment for learning.* Stockholm: Norstedt.

MacMillan, J. H. (2013). *Sage Handbook of Research on Classroom Assessment.* Lodon: Sage.

MacNamara, T. ,. (2006). *Language Testing: The Social Dimension.* Oxford: Oxford.

Messick, S. (1989). *Educational Measurement, Third Edition.* New York: Macmillan.

Messick, S. (1996). *Validity and Washback in Language Testing.* New Jersey: Educational Testing Service.

Michael, M. (1988). *The Construction and Validation of a Performance-based Battery of English Language Progress Tests.* London: University of London Press.

Mora, J. K. (2019). *ALMMethods*. Retrieved from Mora Modules: http://moramodules.com/ALMMethods.htm

Moskal, B. (2000). *Scoring rubrics development.* Retrieved September Saturday, 2019, from https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1093&context=pare

Newman, I., & McNeil, K. A. (1998). *Conducting survey research in the social science.* New York: University Press of America.

Newman, I., Brow, R., & McNeely, S. (2006). *Conceptual statistics for beginners.* University Press of America.

Newman, I., Lim, J., & Pineda, F. (2013). Journal of mixed methods research.

Norris, J. M. (2004). Validity of assessment in foreign assessment.

Ortega, L. (2013). *Understanding Second Language Acquisition.* New York: Routledge .

Phil, R. (2005). *Making Learning Happen.* London: Sage.

Ridenour, C., & Newman, I. (2008). *Mixed methods research: Exploring the interactive continum.* Southern Illionis Up.

Scriven, M., Tyler, R. W., & Gagne, R. M. (1967). The methodology of evaluation.Perspectives of curriculum evaluation. Chicago: Rand McNally.

Shaw, S. a. (2007). *Examining writing: Research and practice in assessing second language writing.* Cambridge: Cambridge University Press.

Stelly, D. J. (2007). *Application of content validationmethods to broader construct.* San Francisko.

Appendix I – Students' survey

"Validity Impact in Assessment" Survey for Students

The survey aims to identify problems students face in assessment and the dissatisfaction they have as a result of poor testing methods. Your responses are going to be used for analytical reasons and kept extremely confidential.

1.What is more stressful to you in a school year?
a) Tests
b) Presentations
c) Both
d) Others

2. Circle the assessing methods at your school:
a) Tests
b) Portfolios
c) Homework
d) Activity

3. Are you satisfied with the assessing methods?
a) Yes b) Partly c) No

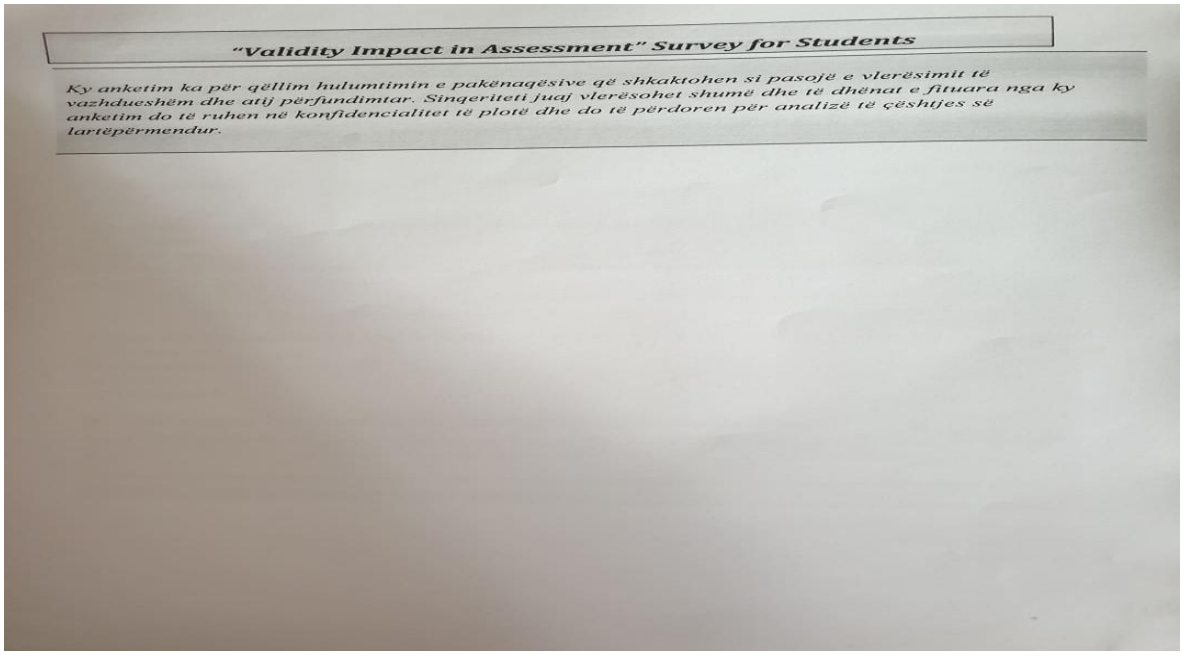
4. How do you feel when being tested?
a) Happy
b) Stressful

5. What skills do you teachers measure in tests?
a) Writing
b) Reading
c) Speaking
d) Listening

7. What do you think should be changed in assessment?
_____
_____.

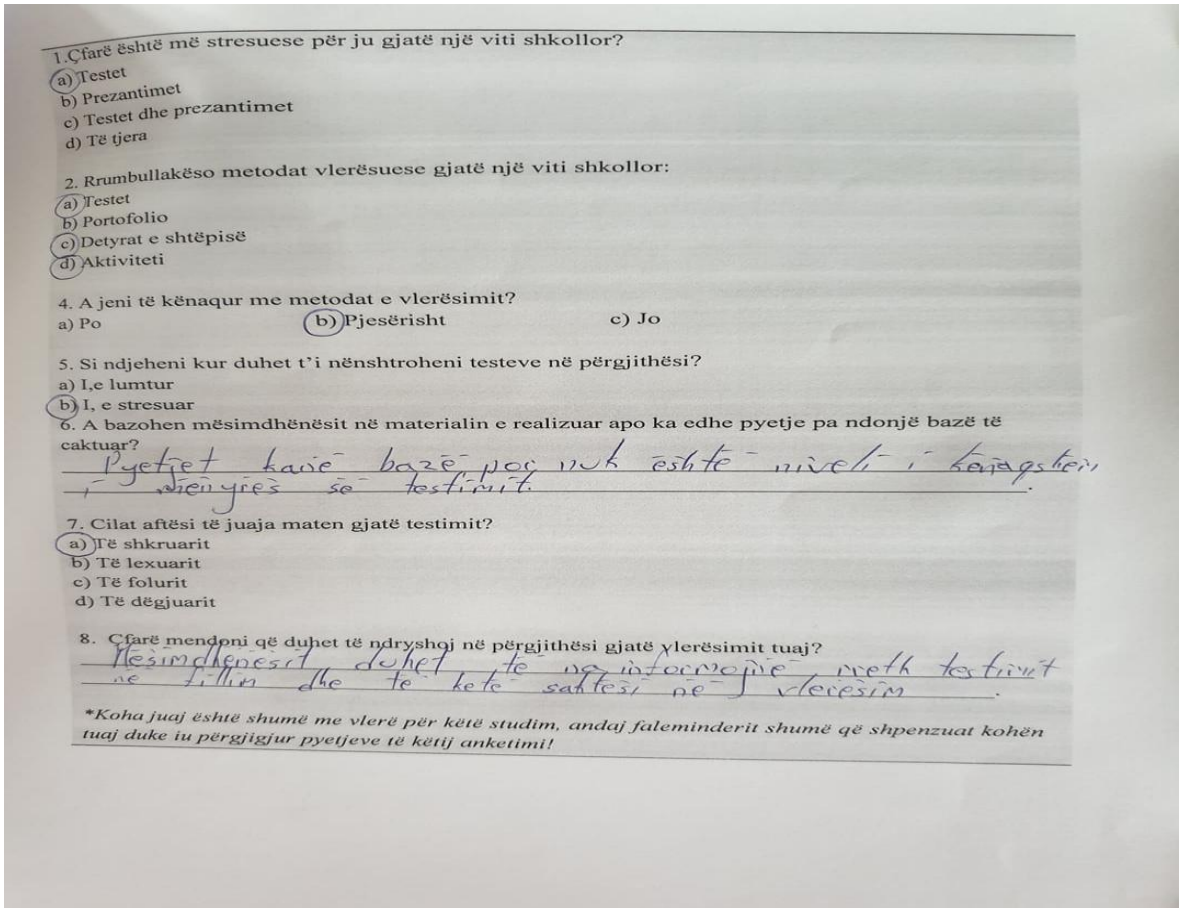
*Your time is irreplaceable in the aim of this study, thus thank you for answering the questions!

Appendix II – Students' survey



"Validity Impact in Assessment" Survey for Students

Ky anketim ka për qëllim hulumtimin e pakënaqësive që shkaktohen si pasojë e vlerësimit të vazhdueshëm dhe atij përfundimtar. Sinqeriteti juaj vlerësohet shumë dhe të dhënat e fituara nga ky anketim do të ruhen në konfidencialitet të plotë dhe do të përdoren për analizë të çështjes së lartëpërmendur.



1.Çfarë është më stresuese për ju gjatë një viti shkollor?
(a) Testet
b) Prezantimet
c) Testet dhe prezantimet
d) Të tjera

2. Rrumbullakëso metodat vlerësuese gjatë një viti shkollor:
(a) Testet
b) Portofolio
(c) Detyrat e shtëpisë
(d) Aktiviteti

4. A jeni të kënaqur me metodat e vlerësimit?
a) Po (b) Pjesërisht c) Jo

5. Si ndjeheni kur duhet t'i nënshtroheni testeve në përgjithësi?
a) I,e lumtur
(b) I, e stresuar

6. A bazohen mësimdhënësit në materialin e realizuar apo ka edhe pyetje pa ndonjë bazë të caktuar?
Pyetjet kanë bazë por nuk është niveli i kërkesës, i shembjes së testimit.

7. Cilat aftësi të juaja maten gjatë testimit?
(a) Të shkruarit
b) Të lexuarit
c) Të folurit
d) Të dëgjuarit

8. Çfarë mendoni që duhet të ndryshoj në përgjithësi gjatë vlerësimit tuaj?
Mësimdhënësit duhet të ngë informojë rreth testimit në fillim dhe të kete saktësi në vlerësim.

Koha juaj është shumë me vlerë për këtë studim, andaj faleminderit shumë që shpenzuat kohën tuaj duke iu përgjigjur pyetjeve të këtij anketimi!

Appendix III – Teachers' survey

"Validity impact in assessment" survey for teachers

1.What is your working experience in teaching?

a) 1 to 5 years of experience
b) More than 5
c) More than 10

2. Circle the most difficult section of teaching:

a) Lecturing
b) Encouraging students to learn
c) Assessment
d) Lesson plans

4. Is it easy to grade students?

a) Yes b) No

5. Do you have any common framework to grade students' achievement? If yes, write that:

a)Yes b)No

_____.

6. Have you ever been confronted with students' resistance to subdue the test? If yes, why?

a) Yes b) No

_____.

7. Is it Validity in Assessment a familial notion to you?

a) Yes b) No

*Validity means how well a test measures what it is supposed to measure.

8. The three main types of Validity are Content-Oriented, Construct-Oriented and Criterion-Oriented. Which one do you think you use in your tests?

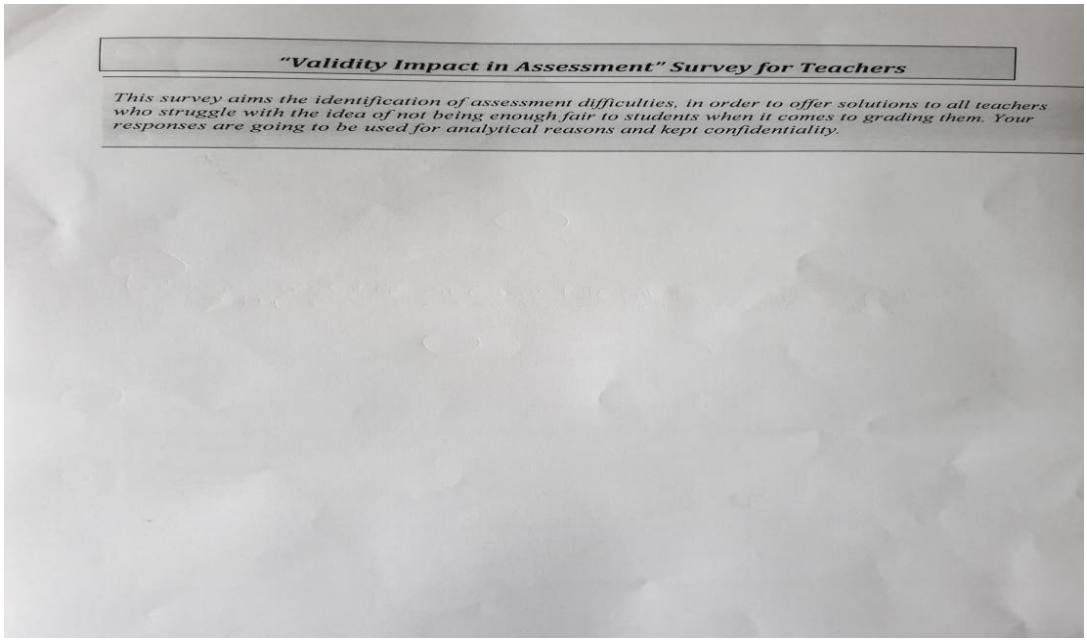
a) Content-Oriented Validity
b) Construct-Oriented Validity
c)Criterion-Oriented Validity

9. Do you reflect after taking results from the tests?
a) Always
b)Sometimes
c)Never

*Your time has been very precious to me, so thank you for spending it answering this survey.

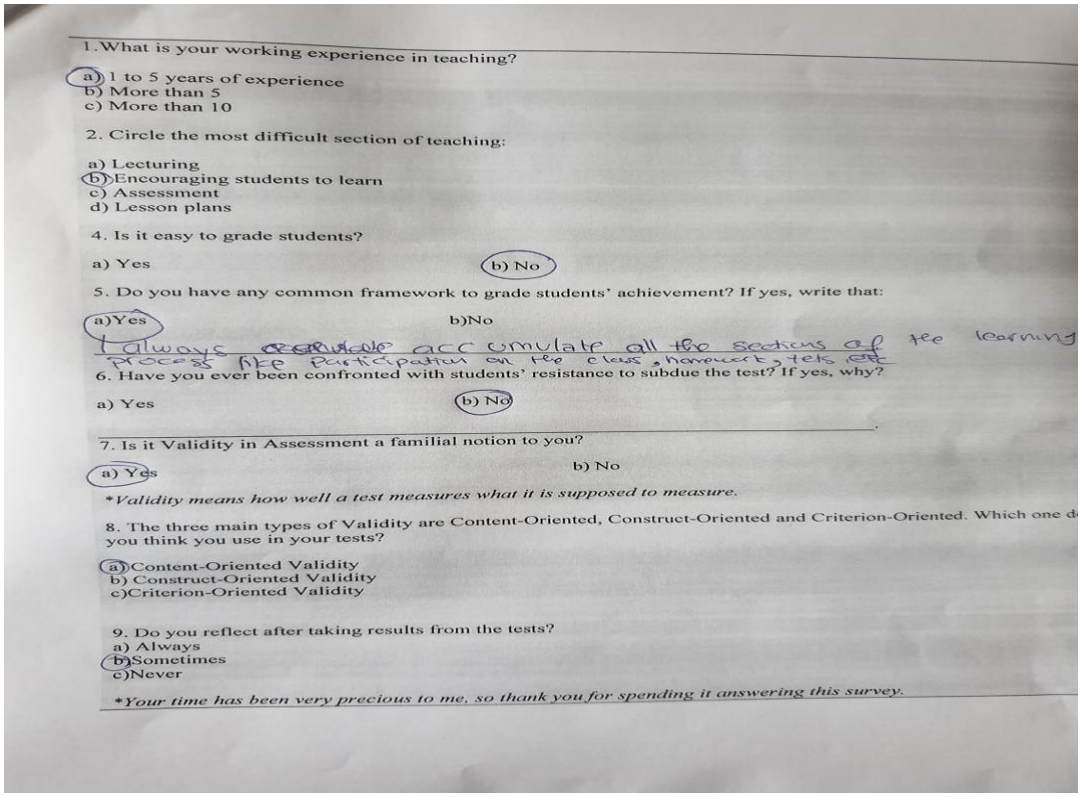
This survey aims the identification of assessment difficulties, in order to offer solutions to all teachers who struggle with the idea of not being enough fair to students when it comes to grading them. Your responses are going to be used for analytical reasons and kept confidentiality.

Appendix IV – Teachers' survey



"Validity Impact in Assessment" Survey for Teachers

This survey aims the identification of assessment difficulties, in order to offer solutions to all teachers who struggle with the idea of not being enough fair to students when it comes to grading them. Your responses are going to be used for analytical reasons and kept confidentiality.



1. What is your working experience in teaching?

a) 1 to 5 years of experience
b) More than 5
c) More than 10

2. Circle the most difficult section of teaching:

a) Lecturing
b) Encouraging students to learn
c) Assessment
d) Lesson plans

4. Is it easy to grade students?

a) Yes b) No

5. Do you have any common framework to grade students' achievement? If yes, write that:

a) Yes b) No

I always accumulate all the sections of the learning process like participation on the class, homework, tests etc

6. Have you ever been confronted with students' resistance to subdue the test? If yes, why?

a) Yes b) No

7. Is it Validity in Assessment a familial notion to you?

a) Yes b) No

**Validity means how well a test measures what it is supposed to measure.*

8. The three main types of Validity are Content-Oriented, Construct-Oriented and Criterion-Oriented. Which one d
you think you use in your tests?

a) Content-Oriented Validity
b) Construct-Oriented Validity
c) Criterion-Oriented Validity

9. Do you reflect after taking results from the tests?

a) Always
b) Sometimes
c) Never

**Your time has been very precious to me, so thank you for spending it answering this survey.*

Appendix V – Focus group

Focus group questions

1. Do you prefer summative or formative assessment?

2. What criteria do you use to grade your students?

3. Do you have ready-made tests or formulate them on your own?

4. Have you ever heard about validity in assessment? If yes, let us know how do you understand it?

Appendix VI – Observation form

Name of teacher:	**Name of observer:**
Focus of observation:	**Date and Time:**
	Level:
	Number of students present:

Observation notes (observer):	
Teacher:	**Observer:**

Appendix VII – Observations

Observation

Name of teacher: Mirlinda Xhelili	**Name of observer:** Arjeta Rexhepi
Focus of observation: **Assessment - Tests**	**Date and Time**: 13th of June 2019, at 2:00 p.m
	Level: Intermediate
	Number of students present: 10

Observation notes (observer):

The teacher entered in the classroom with all her tests and asked the students whether they are ready to be examined. She handed the tests to them and they started working on. By the time the students were working on, I had the opportunity to take a test and the material they went through to analyze it. The test was standardized and it met all the criteria for that specific level, including the assessment of four basic skills, such as: writing, listening, reading and speaking.

The students had a specified time for each section, making everything easier, as they have been announced at the very beginning about how much time they were supposed to have for each section.

The teacher gave me her personal diary as well and I was able to see all the criteria used to grade students at the end of the level. Moreover, they use CEFR about students grade, which is a great facilitation for both teachers and students.

Teacher:	**Observer:**