

**REPUBLIC OF TURKEY
CAG UNIVERSITY
THE INSTITUTE OF SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE TEACHING**

**TEACHERS' PERCEPTIONS OF RELIABILITY AND VALIDITY
IN TEACHER MADE EXAMINATIONS**

THESIS BY

Ayşe ŞAHİN

SUPERVISOR

Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

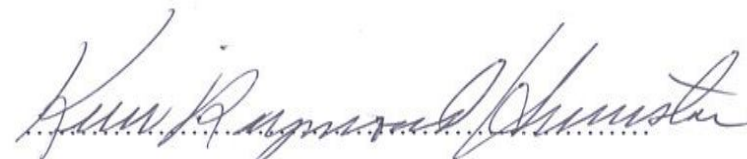
MASTER OF ARTS

MERSİN, DECEMBER 2014

REPUBLIC OF TURKEY
ÇAĞ UNIVERSITY
DIRECTORSHIP OF THE INSTITUTE OF SOCIAL SCIENCES

We certify that this thesis under the title of **TEACHERS' PERCEPTIONS OF RELIABILITY AND VALIDITY IN TEACHER MADE EXAMINATIONS** is satisfactory for the award of the degree of **Master of Arts** in the **Department of English Language Teaching**


.....
Supervisor- Head of examinig committee: Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ


.....
Member of examining committee: Assist. Prof. Dr. Kim Raymond HUMISTON


.....
Member of examining committee: Assist. Prof. Dr. Hülya YUMRU

I certify that this thesis conforms to the formal standards of the Institute of Social Sciences.


30/12/2014

Assist. Prof. Dr. Murat KOÇ
Director of the Institute of Social Sciences

Note: The uncited usage of the reports, charts, figures and photographs in this dissertation, whether original or quoted for mother sources, is subject to the Law of Works of Art and Thought No:5846

ACKNOWLEDGEMENTS

There are many people to whom I would like to express my gratitude. They made this study possible. I would like to first thank to my dear teachers Assist. Prof. Dr. Hülya Yumru and Assist. Prof. Dr. Kim Raymond HUMISTON. I would like to express my deepest gratitude to my supervisor Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ for her contributions, invaluable feedback and patience throughout the completion of this thesis.

I wish to thank my devoted parents, Vechettin and Gülcan Sobacı, for their lasting love, caring and patience. My special thanks go to my dear husband, Ferit, for his deep love, moral support and never-ending trust in me. Without my husband, it would have been impossible for me to survive this year.

30.12.2014

Ayşe ŞAHİN

ÖZET

ÖĞRETMENLERİN ÖĞRETMENLER TARAFINDAN HAZIRLANAN SINAVLARIN GÜVENİLİRLİĞİ VE GEÇERLİLİĞİ İLE İLGİLİ ALGILARI

Ayşe ŞAHİN

Yüksek Lisans Tezi, İngiliz Dili Eğitimi Anabilim Dalı

Tez Danışmanı: Doç. Dr. Şehnaz ŞAHİNKARAKAŞ

Aralık 2014, 78 Sayfa

Bu çalışma, öğretmenlerin öğretmenler tarafından hazırlanan sınavların geçerliliği ve güvenilirliği ile ilgili algılarını incelemiştir. Bu çalışma ayrıca öğretmenlerin sınav hazırlarken güvenilirliği ve geçerliliği sağlamak için nelere dikkate aldıklarını ve neler yaptıklarını da ortaya çıkarmayı amaçlamaktadır.

Bu çalışmada bir kontrol listesi ve yarı yapılandırılmış görüşme kullanılmıştır. İlk olarak, öğretmenlerden kontrol listesindeki soruları cevaplamaları istendi. Bu sorular, terimler kullanılmadan öğretmenlerin güvenilirlik ve geçerlilik ile ilgili algılarını ve güvenilirlik ve geçerliliği sağlamak için hangi yöntemleri kullandıklarını tanımlayan sorulardı. Daha sonra, bu soruları cevaplayan öğretmenlerle güvenilirlik ve geçerlilik terimlerinden bahsedilerek yarı yapılandırılmış sorular soruldu.

Veri analizi, öğretmenlerin güvenilirlik ve geçerlilikle ilgili bilgilerinin birbirine yayık olduğunu ortaya koydu. Sadece birkaç öğretmen güvenilirlik ve geçerliliği akademik literatürde ki terimsel anlamlarıyla ifade etmiştir. Öğretmenlerin kontrol listesindeki sorulara ve görüşme sorularına verdikleri cevaplar arasındaki fark da dikkat çekicidir. Öğretmenlere doğrudan güvenilirlik ve geçerlilik bağlantılı terimler görüşmede sorulduğunda, çoğunu bu terimleri tamamen tanımlayamamıştır. Fakat kontrol listesindeki güvenilirlik ve geçerliliği sağlamak için hangi yöntemleri uyguladıkları sorulduğunda, farkında olmadan güvenilirlik ve geçerliliğin bazı ilkelerini uyguladıkları ortaya çıkmıştır.

Anahtar Kelimeler: Öğretmenlerin hazırladıkları sınavlar, Güvenilirlik, Geçerlilik.

ABSTRACT

TEACHERS' PERCEPTION OF RELIABILITY AND VALIDITY IN TEACHER MADE EXAMINATIONS

Ayşe ŞAHİN

Master of Art Thesis, Department of English Language Teaching

Supervisor: Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

December 2014, 78 Pages

This study aims to reveal teachers' perceptions of reliability and validity in teacher made examinations. It also aims to find out what teachers do and take into consideration to promote reliability and validity in tests while preparing these tests.

The instruments used in this study were a checklist and a semi-structured interview. First, teachers were asked to answer the checklist questions to define teachers' perceptions of reliability and validity and the ways they used to promote reliability and validity without naming the terms "reliability" and "validity. Thereafter, the same teachers were interviewed to find out teachers' perceptions of reliability and validity by mentioning the exact terms "reliability" and "validity".

Analysis of data revealed that teachers' knowledge of reliability and validity are nearly similar. In consideration of academic literature of reliability and validity, the terms "reliability" and "validity" are just defined by few teachers properly. The results showed that there is a difference between teachers' answers to checklist questions and interview questions. When teachers were directly asked to define terms related to reliability and validity in interviews, most of them could not exactly define the terms. However, when they were asked about their practices to promote reliability and validity in tests in the checklist, it was founded that they applied some principles of reliability and validity unknowingly.

Key Words: Teacher made examinations, Reliability, Validity.

LIST OF TABLES

Tablo 1 Teachers' Educational Backgrounds	26
Tablo 2 Teachers' Teaching Experiences	26
Tablo 3 Types of the Schools in Which Teachers Work.....	26
Tablo 4 Teachers' Perceptions of Reliability	31
Tablo 5 Teachers' Perceptions of Validity	34
Table 6 The Ways Teachers Promote Validity	36
Table 7 Teacher's Perceptions of Reliability	37
Table 8 Teacher's Perceptions of Validity	41
Table 9 Reasons for the Importance of Reliability	45
Table 10 Reasons for the Importance of Validity.....	47
Table 11 Suggestions for Promoting Reliability.....	49
Table 12 Suggestions for Promoting Validity	53

TABLE OF CONTENT

COVER	i
APPROVAL PAGE	ii
ACKNOWLEDGEMENTS	iii
ÖZET	iv
ABSTRACT	v
LIST OF TABLES	vi
LIST OF CONTENT	vii

CHAPTER ONE

1. INTRODUCTION	1
1.1 Background of Study	1
1.2. Statement of the Problem	3
1.3. Aim and Significance of the Study	3
1.4. Limitations of the Study	4
1.5 Operational Definitions.....	5
1.6. Conclusion.....	5

CHAPTER TWO

2. LITERATURE REVIEW	6
2.1. Introduction	6
2.2. Use of Language Tests.....	8
2.3. Types of Language Tests.....	10
2.3.1. Placement Tests.....	11
2.3.2. Diagnostic Tests.....	12
2.3.3. Proficiency Tests.....	12
2.3.4. Achievement Tests.....	13

2.3.5. Teacher Made Examinations.....	14
2.4. Principles of Language Tests	15
2.4.1. Practicality.....	16
2.4.2. Authenticity.....	16
2.4.3. Washback.....	17
2.4.4. Reliability.....	17
2.4.4.1. Types of Reliability.....	18
2.4.4.1.1. Test- Retest Method.....	19
2.4.4.1.2. Inter Rater Reliability.....	19
2.4.4.1.3. Parallel Forms (Alternate Forms).....	20
2.4.4.2. Reliability of Teacher Made Examinations	20
2.4.5. Validity.....	21
2.4.5.1. Types of Validity.....	21
2.4.5.1.1. Content Validity.....	22
2.4.5.1.2. Construct Validity.....	23
2.4.5.1.3. Criterion Related Validity.....	23
2.4.5.1.4. Face Validity.....	24
2.4.5.2. Validity of Teacher Made Examinations.....	24

CHAPTER THREE

3. METHODOLOGY	25
3.1. Introduction	25
3.2. Research Design.....	25
3.3. Participants	25
3.4 Instruments	27
3.5. Procedural Details	27
3.5.1. Confirmation and Description of Checklist and Interview Questions....	27

3.5.2. Piloting.....	29
3.5.2.1. Implications for the Main Study.....	29
3.5.2.2. Main Study.....	29
3.6 Data Analysis	30

CHAPTER FOUR

4. DATA ANALYSIS AND RESULTS	31
4.1. Introduction	31
4.2. Findings from Checklists.....	31
4.3. Findings from Interviews.....	37
4.3.1. Question 1: What does reliability remind you?.....	37
4.3.1.1. Consistency.....	37
4.3.1.2. Cheating.....	38
4.3.1.3. Questions.....	39
4.3.1.4. Justice in Evaluation.....	40
4.3.1.5. Environmental Conditions.....	41
4.3.1.6. Nothing.....	41
4.3.2. Question 2: What does validity remind you?.....	41
4.3.2.1. Appropriateness.....	41
4.3.2.2. Cheating.....	42
4.3.2.3. Subjects.....	43
4.3.2.4. Objectivity in Evaluation.....	44
4.3.3. Question 3: To what extent is reliability important in texting? Why?.....	44
4.3.3.1. Students.....	45
4.3.3.2. Tests.....	46
4.3.3.3. Achieving Objectives.....	47
4.3.3.4. Having no idea.....	47

4.3.4. Question 4: To what extent is validity important in texting? Why?.....	47
4.3.4.1. Evaluation.....	47
4.3.4.2. Intelligibility.....	48
4.3.4.3. Evindence for Reliability.....	49
4.3.4.4. Paying no Attention.....	49
4.3.5. Question 5: What procedures do you follow to promote reliability?.....	49
4.3.5.1. Questions.....	50
4.3.5.2. Repetition.....	51
4.3.5.3. Cheating.....	52
4.3.5.4. Do Nothing.....	52
4.3.6. Question 6: What procedures do you follow to promote validity?.....	53
4.3.6.1. Questions.....	53
4.3.6.2. Objectives.....	54
4.3.6.3. Do Nothing.....	55
4.4. Conclusion.....	55

CHAPTER FIVE

5. CONCLUSION	56
5.1. Introduction	56
5.2. Summary of the Study	56
5.3. Discussion.....	56
5.4. Limitations and Implications for Further Study.....	58
6. REFERENCES	59
7. APPENDIXES	63
7.1. Appendix A: Checklist	63
7.2. Appendix B: Interview Questions.....	66
7.3. Appendix C: Piloting the Checklist	67
7.4. Appendix D: Checklist Before Piloting	68

CHAPTER ONE

1. INTRODUCTION

1.1 Background of Study

Testing is an inseparable part of language teaching and learning. It is used as a means of gathering information about national educational standards; students' strengths and weaknesses; what students have learned; what is to be taught to eliminate the deficiencies in learning process and what a teacher has to do in the next step of learning process. In other words, it is of prime importance both to certificate the success and promote learning. Tests are also the most commonly used ways to monitor students' achievements or failures and grade students in order to achieve these goals. Bachman (1990) points out that:

Language tests can be valuable sources of information about the effectiveness of learning and teaching. Language teachers regularly use tests to help diagnose student strengths and weaknesses, to assess student progress, and to assist in evaluating student achievement.(Bachman, 1990, p. 2).

First and foremost, tests are used to provide essential information to make decisions about learners (Bachman, 1990; Fulcher, 2010; Henning, 1987). Learners are tested since a teacher needs to get information about their academic standing to make decisions on their educational progress such as their level, the next subject to be taught in the lesson, pass-fail grading condition, enhancing on-going teaching -learning process. Therefore a wrong decision causes waste of time and effort for both teachers and learners.

Due to having so much effect on students' academic lives, tests are to possess several qualities which bring tests to perfection. Among these qualities, validity and reliability are the most important ones. Walt and Steyn (2008, p. 191) state that "validity and reliability have traditionally been regarded as the basic criteria which any language test should satisfy". For efficient and consistent tests, validity and reliability have to be defined in every respect in a detailed manner. Reliability is regularly defined with the question whether a test provides identical or similar results when the same or similar

test is applied to the same student under the similar circumstances in a different time. Brown and Abeywickrama (2010, p. 20) state that “a reliable test is consistent and dependable. If you give the same test to the same students or matched students on two different occasions, the test should yield the similar results”. In other saying, test reliability proves the consistency of students’ marks that they have got from alternative forms of the same tests. Therefore creating a reliable test is not easy, since unreliability may result from some factors such as alternations in students, in scoring, in tester, and in the test itself. Brown and Abeywickrama (2010) state that:

the issue of reliability of a test may best be addressed by considering number of factors that may contribute to the unreliability of a test. Consider the following possibilities: fluctuations in student, in scoring, in test administration and in test itself. (Brown and Abeywickrama, 2010, p. 20)

These factors are categorized as student-related reliability, rater reliability, test administration reliability, test reliability. Beside reliability, validity is another quality that is important to be concerned in testing. Simply, the question whether a test assesses what is taught and meant to be assessed defines validity. Imane (2003, p. 6) cited that “a valid assessment is one which assesses what is set out to assess”. However, assessing the validity of a specific test is troublesome and time consuming process. Hence, to assess validity more easily and systematically, there are types of validity and each type gives an opportunity to collect required data for validating a test. Content validity, criterion-related validity, construct validity, consequential validity, face validity are the types of validity that are mentioned in several research (Brown and Abeywickrama, 2010; Harrison, 1983; Henning, 1987). Despite differences between reliability and validity, these two qualities are complementary as reliability is a prerequisite for validity and reliability and validity are the two essential concepts that lead to good practice in language testing and assessment.

By virtue of having so much importance, quite a lot of research has been dedicated to focus on reliability and validity of tests. However, matters not provided for is teachers’ perceptions about reliability and validity. Teachers’ perception of reliability and validity is a significant subject to be investigated as teachers are a part of testing. Davies (1990) claims that “language testing provides goals for language teaching, and it

monitors, for both teachers and learners, success in reaching those goals” (Davies, 1990, p. 1). In other saying, teachers make benefit of testing to find out how effective their teaching is, what deficiencies in learning- teaching process are, to decide what to do as a next step in educational process and most importantly to decide what to assess next. Through testing, a teacher has a fundamental role both as a tester and supervisor who alternates educational process, makes decisions on teaching process, activities, and subjects.

In the light of the studies, it is momentous to prepare convenient tests that really assess what is meant to be assessed. Qua a tester, teachers should be aware of the qualities of testing because they prepare their own tests which are named teacher –made tests. When a teacher prepares a test, a special attention should be paid to these two qualities; validity and reliability. In order to prepare valid and reliable tests, teachers should gain insight about what reliability and validity are and how important they are for consistent and effective testing. This study attempts to gather information about teachers’ perception of reliability and validity.

1.2. Statement of the Problem

English language teaching and language testing are inseparable components of language acquisition. These two inseparable parts serve each other to get down to a fine art in language acquisition. And a teacher is the most important human being who serves to these inseparable components. S/he is in the centre of language acquisition field. Hence a teacher has to both maintain the language teaching process and assess it. S/he assesses to find out students’ performance degrees and in addition to this s/he assesses to get information to advance the teaching process. To get right information about students’ progress and teaching process, valid and reliable tests are essential. Although the validity and reliability of tests that are implemented to assess English language ability draws considerable attention in literature (Brown and Abeywickrama, 2010; Davies, 1990; Kunnan, 2000), teachers’ perception of validity and reliability does not have the same importance.

1.3. Aim and Significance of the Study

Searching out teachers’ perception of validity and reliability of teacher-made tests is the main purpose of this study. As teachers’ perception generally affect the

teaching and testing process, their perception and attitude gain a big importance. More deeply, this study aims:

- to find out how teachers define validity and reliability
- to search out how much teachers are aware of the validity and reliability
- to show how much importance they attach to validity and reliability of teacher-made exams
- and finally to find out what teachers do to promote validity and reliability of tests.

Researchers make judgments about language tests by assessing their validity and reliability. However, there is a lack of research in the field of English language testing that focuses on teachers' perceptions and attitudes towards validity and reliability. Since teachers' perceptions of validity and reliability affect testing process crucially, research on this field becomes essential to get an idea about what teachers do and respect as a tester. For this reason, the researcher aims to acquire trustworthy data about teachers' perceptions of validity and reliability in teacher made examinations. Therefore this study may contribute to the literature by gaining a different point of view. This study also pays attention to teachers' awareness of validity and reliability as these two qualities are main points of effective and accurate testing. Besides the emphasis is on teacher made examinations which are affected by teachers' attitudes.

As another objective, this study aims to give a light to what teachers do to prepare reliable and valid tests, whether their perception and their test preparing process have a positive or negative relation. The present study aims to give answer to the research questions below:

RQ 1- What do teachers think about the reliability and validity in teacher made examinations?

RQ 2- What do teachers do for reliability and validity in exams which they prepare?

1.4. Limitations of the Study

The major limitation of current study is the number of the subjects. The study is carried out with 23 English language teachers from different schools and backgrounds. Yet, to gain a more general idea, more people should be participated in the study. In

addition to the number of the subjects, in terms of data collection, this study is limited with a checklist and interviews that are used as instruments to collect the necessary data.

1.5 Operational Definitions

Teacher-made Test: Teacher-made criterion-referenced tests evaluate a student's achievement of a teacher's instructional objectives. They are not norm-referenced since their purpose is not to reveal differences among students, but to see what a particular student can do relative to a teacher's instructional objectives (Gage and Berliner, 1979, p. 63).

Validity: Validity in general refers to the appropriateness of a given test or any its component parts as a measure of what it is purported to measure (Henning, 1987, p. 91).

Reliability: Reliability refers to the consistency of test scores, which simply means that a test would offer similar results if it were given at another time (Coombe, Folse, and Hubley, 2007, p. 39)

1.6. Conclusion

This chapter aims to introduce the present study through background of the study, the statement of the problem, the purpose of the study and the research questions and some key words. In the second chapter of the study, there will be a review of literature which includes definition, purposes, uses, types and principles of language testing and a detailed explanation of validity and reliability. In third chapter, settings, participants, methodology, instruments, data collection process and data analysis will be presented. In the fourth chapter, data analysis procedure and findings will be reported. Finally, in the fifth chapter, overview of the study, discussions and conclusion will be presented.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Introduction

This study makes an attempt at searching how aware teachers are of validity and reliability in English language testing. At this chapter, use of language tests is reviewed with various definitions from several researchers. Afterwards, types of language tests are explained and categorized as placement test, proficiency test, diagnostic test, achievement test. Under the term ‘achievement test’, standardized and teacher-made tests are defined briefly.

A general classification of principles of language testing is given as practicality, authenticity, washback, reliability and validity. Due to being main notion of this study, reliability and validity are discussed in detail by defining subtitles of reliability and validity. In addition to this, validity and reliability of teacher-made tests are explained.

Assessment is of a capital importance to education both to certificate the success and promote learning. It is an enormous issue that contains everything from countrywide presentableness tests to teacher’s comments during lesson. As Brown (1996) states it:

is an ongoing process that encompasses a much wider domain. Whenever a student responds to a question, offers a comment, or tries out a new word or structure, the teacher subconsciously makes an assessment of the student's performance (Brown, 1996, p. 15).

In this sense, assessment exists in almost every phase of language education. A good teacher exploits all chances to assess students and use assessment results in his classes. Since assessment is the product of what is taught, an educational process cannot be thought without it.

What and why a teacher assesses is a noteworthy issue to argue about since it has a great importance in education. In the easiest configuration, assessment is a kind of judgement which is professional and an image formed by information about students’ performance. It is a tool used to gather information about whether teaching or learning

processes have been prospering. However teachers, students and policy makers use this information for different purposes. Wiliam (2000) states that:

Educational assessments are also used to provide information with which teachers, educational administrators and politicians can be held accountable to the wider public. ... Within schools, educational assessments are used to determine the route a student takes through the differentiated curricula that are on offer, as well as to report on a student's educational achievement either to the student herself, or to her parents or guardians (William, 2000, p. 1).

Concordantly, the information gathered from assessment may be used for several purposes by various people. An educator's purpose to use this information may be to have an idea about how much his students have improved themselves and what is missing in their teaching and what to do to make learning more effective and successful. On the other hand, a student's purpose may be realizing his weaknesses and strengths in language learning.

These different purposes of assessments end up with a range of approaches. In approaches of assessment, two tendencies emerge. One of them focuses on assessment of learning and the other one focuses on assessment for learning. Assessment for learning that is defined as formative assessment emphasizes on feedback to improve students' learning and performance. Formative assessment is a part of the instructional process. It may be seen as a process assessing both the product and the process. No matter for what reason it is used, it provides the information needed to adjust teaching and learning while they are happening. Both teachers and students can get clear information about what is happening in language education process. As Mulvenon and Dunn (2009) states it is:

as a process used during instruction to provide feedback for the adjustment of on-going teaching and learning for the purposes of improving student achievement related to instructional objectives (Mulvenon and Dunn, 2009, p. 2).

In this respect, formative assessment is formed to monitor student's progress during the teaching and learning process. It evaluates students in the learning process in terms of

their competence and skills with the aim of helping them to go on improving their language learning. It also gives a chance to the teacher to determine next steps during the learning process. Because of these practical purposes, all kinds of informal assessment such as feedback, suggestions, correcting a mistake are formative.

Assessment of learning makes summative judgements which in practice are likely to involve more formal examinations and tests with marks and it is called summative assessment. Summative assessment is defined as a kind of assessment designed to determine students' academic development after a unit and a term. Brown (1996, p. 17) states that it "aims to measure, or summarize, what a student has grasped, and typically occurs at the end of a course or unit of instruction". In this view of educational assessment, this kind of assessment leads to judgements which include everything related to a grade or given point. Therefore, quizzes, periodical unit tests, final and mid-term exams etc. are summative. In other words summative assessment may be defined as language testing which includes all kinds of tests.

2.2. Use of Language Tests

Teaching and testing are interrelated parts of language acquisition. It is hard to separate them from each other since testing provides remarkable information for both teachers and students to find out and correct their deficiencies in the process of teaching and learning. Harrison (1983, p. 1) indicates that "a test is seen as a natural extension of classroom work, providing teachers and students with useful information that can serve each as basis for improvement". Testing also plays the role of guiding students to improve their language, to find out their deficiencies and be aware of their achievements. It is not just grading students; it guides both teachers and student. Each test is a learning setting and a device that contributes to educational programme. In this regard, a teacher may teach students in order to improve their language knowledge and usage, not to test. A teacher who is aware of this quality of testing uses testing as a medium to teach, not as a final target. Siddiek (2010) states that:

Most discussions about the value of educational assessment focus on its use for the purposes of grading the students' marks and ignore other more important role of helping students with their learning. Cliff believes that, assessment must be used to help students. It can operate in two ways: Firstly, any teacher knows that assessment, particularly in the form of

examinations, will determine to a large extent when students will study, what they will study, and the effort they will put into their work. Secondly: when students are provided with a programme of progressive assessment with appropriate information as to their strength and weaknesses, they can take action to correct and improve performance prior to any major or definitive assessment (Siddiek, 2010, p. 133).

Language testing is a method to collect information about quality of teaching- learning process and language abilities from results of tasks that are prepared to supply essential information about students' usage of these language abilities in real contexts. Teachers may also change their teaching methods and process to enhance learning in the light of test results. McNamara (2000, p. 11) states that "testing thus necessarily involves interpretation of the data of test performance as evidence of knowledge or ability of one kind or another". As is stated, testing is a gateway to see whether students succeed in learning new language subjects, achieve objectives of the course, and gather useful information for further teaching. Owen (1997) draws attention to the fact that although necessity for language testing is a matter of discussion among researchers, it is generally accepted to be a useful and practical way to observe learners systematically and give a chance to find out how successful or unsuccessful learners are.

Language testing springs from a need for practicing and applying teaching and learning. It is a significant process through which the success of the educational programme is determined and the educational goals are strengthened. It is an integrated part of language education. Since it is used as an appliance to plan, monitor, guide and promote the educational process, teachers and students get advantages of it. Davies (1990, p.1) claims that "language testing is central to language teaching. It provides a methodology for experiment and investigation both in language teaching and language learning/ acquisition". In other words, it presents the information that is needed to set goals for language teaching, monitors success of the goals and enlightens the process of teaching and learning to improve it. It serves both to the field of language teaching and language acquisition.

Language testing is used as a tool for feedback, classification, and research. It reunites both language and testing. Therefore, it has a role of a researcher for language and language acquisition. In other words, language testing serves to applied linguistics

as a data collection tool. The results that are acquired from the language tests may be used by language researchers no matter which fields they focus on. For example, with the help of a test result, it is possible to determine which method is successful or how successful a method is. Davies (1990, p. 10) points out that “language testing, on the other hand, has a useful role in itself- into the nature of language as well as into language acquisition – and all this in addition to doing what subjects tests typically help explain, the extent of subject learning” .

2.3. Types of Language Tests

The importance of language testing as a device for gathering essential information and making decisions has been discussed comprehensively in the previous chapter. However, how a test functions as a medium of gathering information and the teacher’s decisions about it is another subject that is required to be taken into consideration. All tests do not provide information in the same way. Hughes (1989) states that:

We use tests to obtain information. The information we hope to obtain will of course vary from situation to situation. It is possible, nevertheless, to categorize tests according to small kinds of information being sought (Hughes, 1989, p. 9).

Hughes means that types of language tests are results of different purposes and functions. Some tests aim to assess students’ knowledge and abilities about the previous course or subjects and give feedback to both teachers and students about teaching- learning process, while some other tests are used to decide on how to group students.

In this respect, the aim of a test and the way it is constructed divide language tests into different types. Brown and Abeywickrama (2010) refer four main types of language tests which are placement tests, diagnostic tests, proficiency tests and achievement tests.

In this study, these four types of tests will be discussed with different definitions from various researchers. Furthermore, achievement test will be categorized as a standardized test and a teacher-made test with brief information.

2.3.1. Placement Tests

Placement test is designed to gather general information about students' language ability and knowledge in order to group them for their future education. It is a kind of reference for the following learning and teaching process.

In other words, placement test is a test which measures proficiency for placement in a program. Hughes (1989) indicates that:

Placement tests, as their name suggest, are intended to provide information which help to place students at the stage (or in part) of the teaching programme most appropriate to their abilities. Typically they are used to assign students to classes at different levels (Hughes, 1989, p. 14).

Placement test is a general reference to direct future learning so it intends to put students into appropriate teaching groups to provide a qualified and suitable language education. Indeed, this kind of test allows students to be grouped according to their language ability at the beginning of the course (Brown and Abeywickrama, 2010; Heaton, 1990). As a matter of fact, placement test allows the teacher or authority to decide on which level, group or class a learner will learn more effectively. Thus, the teacher takes advantage of this grouping implementation since he/she teaches a language to students who are on the same levels and have similar background of knowledge. Besides, the teacher may benefit from placement tests more efficiently if he designs placement tests which are suitable to syllabus, subjects and objectives of the course. As Davies, Brown, Elder, Hill, Lumley and McNamara (2002) point out:

To be most effective, placement tests should reflect the features of the teaching context (such as the proficiency level of the classes, the

methodology and syllabus type). A grammar placement test, for example, may not be the most suitable test where the syllabus is task oriented (Davies et al., 2002, p. 145).

Finally, teachers might design placement tests in such a way that it will enable teachers to teach according to the syllabus and objectives.

2.3.2. Diagnostic Tests

Diagnostic test is designed to check students' progress during teaching and learning process. As the name itself suggests, diagnostic test defines the deficiencies and problems that occur during teaching and learning process. Harrison (1983, p. 6) states that "a diagnostic test (sometimes called a formative or progress test) checks on students' progress in learning particular elements of the course".

Diagnostic tests may be used during courses or at the end of each unit to find out whether teacher has achieved the objectives of the unit or students has learned intended subjects. By means of diagnostic tests, teachers and students realize students' weaknesses and strengths, and deficiencies of teaching and learning process. In case of lack of knowledge, instant remediation becomes necessary. As Hughes (1989, p. 13) states "they are intend primarily to ascertain what further teaching is necessary".

In other words, it is an immediate feedback which serves teachers and students as an explicit guidance to alter their teaching and learning methods and find ways to advance.

2.3.3. Proficiency Tests

Proficiency test provides a general knowledge that is necessary to enter an institution or get a new position. It assesses students' ability for particular applications of language acquired. It measures student's knowledge, ability, skills compared to other students' and native speakers. It is a test of generalizability. Davies et. al (2002) state that a proficiency test is:

a test which measures how much of a language someone has learned. Unlike an achievement test, a proficiency test is not based on a particular course of instruction. A proficiency test always measures what the

candidate has learned relative to a specific real world purpose, for example, does he/she know enough of target language to follow a lecture, train as an engineer or work as a ski instructor in that medium, or to translate to the requisite standard out of that language (Davies et al., 2002, p. 154).

As stated by Davies et. al, (2002, p. 154) proficiency test is used to answer the question “what has the student learned so far and what can the student do with his learning in real world?”. The content of a proficiency test does not only include the objectives or content of courses which students have taken. It also includes the abilities and knowledge which a student has to acquire in language to be considered a proficient which means he/she has sufficient control of language in a special field. It is prepared to find out student’s general level of language mastery. Bachman and Palmar (1996, p. 4) state that “a good language proficiency test is made up of language tasks that replicate what goes on in the real world”.

2.3.4. Achievement Tests

Achievement test is designed to show to what extent a student has achieved the long-term objectives of the language course. In a sense, it is used at the end of the course and serves as a tool to decide whether a student has gained essential language part or has achieved the goals to pass to next level. Henning (1987) indicates that:

achievement tests are used to measure the extent of learning in a prescribed content domain, often in accordance with explicitly stated objectives of a learning program. These may be used for program evaluation as well as for certification of learned competence (Henning, 1987, p. 6).

Achievement test aims to measure the knowledge, skills, and abilities of the material which are directly taught during the course, at the end of the course, term or unit. Achieving the goals of course is important to set new goals for forward teaching and learning process. This definition of achievement test may be confused with diagnostic test. However, diagnostic test is related to specific short-term objectives and used to give immediate feedback to the teacher and students to change the continuing teaching-learning process. In contrast, the objectives, which are taken into consideration during

the preparation of an achievement test, are related to long-term objectives. Harrison (1983, p. 7) argues that “the conditions for setting an achievement test are that it covers a much wider range of material than a diagnostic test and relates to long-term rather than short-term objectives”.

2.3.5. Teacher Made Examinations

Summative assessment of language teaching and learning or language testing is more commonly used because of several reasons. One of these reasons is that with language testing parents get tangible information about their child's progress. As Ioannou- Georgiou (2003, p. 62) states, parents “often ask to see proof of effectiveness in the traditional form they are accustomed to: test results”. They want to be informed with grades which is more objective and shows them their child’s success level among other students. Another reason is teachers’ familiarity and tendency to this testing process. They assess students in a summative way as they believe that language testing is objective and has a large mechanism to be supported. In addition to this, in a country where a teacher is evaluated with their students’ tests results and success, a teacher has a determined mind to teach what he will assess. By doing so, he gets rid of the stress and the anxiety caused by the feeling of becoming unemployed or getting administrative punishment. Ioannou- Georgiou (2003) states that the teacher’s anxiety:

is, however, unconsciously transmitted to young learners who often find that their fun and exciting lessons have turned into stressful sessions of never-ending practice of exam items. If teachers see test results as evidence of their effectiveness, the unfortunate result will be teaching to the test (Ioannou-Georgiou, 2003, p. 63).

In this sense, being under pressure because of students’ success or failure causes a huge problem affecting learning and teaching process in a negative way. Teachers under this pressure may not be able to teach students the required knowledge but try to teach only the knowledge which will label him as a successful teacher.

The most significant reason for using language tests is that it is thought to be gradable, more interpretable, and reliable. Thus, policy makers want teachers to assess their students by tests. They gather information from these tests results to form or adjust the education policy according to these results.

As testing is an integral part of educational programme, a teacher might have sufficient knowledge of how to teach a language and how to evaluate it. In other words, the teacher's job is not restricted to teaching effectively; but he also assesses learners' abilities and skills. To prepare a good test, a teacher might have enough knowledge of testing, its characteristics, its types, its practice, and the coherence between the objectives of the course and the test and how to use the results of the test more efficiently to promote educational process. Nakarin (2007) states that:

preparing a test, the teacher needs to have a clear conception of how the test, together with the test results are to be functioned and used; and requires prior specification of instructional objectives and decisions regarding the sequence and method of instruction (Nakarin, 2007, p. 71).

Teacher-made tests, as a kind of language tests, are at the centre of testing process by virtue of being prepared by teachers to evaluate the outcome of educational process (classroom instruction) according to specific syllabus, classroom instruction and course objectives. They reveal specific areas of instruction in which students need remedial help. More precisely, students are peerless humans have different abilities, interests, understanding and attitudes towards language and all these differences affect language teaching and learning process in a positive or negative way. Accordingly, a teacher could be aware of these differences and assess students' multiple learning by the help of a comprehensive test which is prepared by considering the objectives of lesson, syllabus, and the instruction process. As Wrightstone (1961) states that:

teacher made tests are a necessary part of evaluation. This is true, first, because day by day classroom appraisal is of equal and greater importance than standardized tests administered usually not more than per year; and second, because standardized tests and published scales are not available for measuring interests, attitudes, critical thinking at the elementary level (Wrightstone, 1961, p. 171).

2.4. Principles of Language Tests

Language tests are useful and essential tools for teachers and authorities to see students' achievements or failures, to find out and reform the deficiencies of the

continuing teaching – learning process, to make decisions on the educational programs. Due to having such great importance in language teaching and learning, language tests should be prepared by taking into consideration some principles. Bachman and Palmer (1996, p. 38) indicate that ‘usefulness’ is the most significant feature of tests, and six test qualities contribute to test usefulness: reliability, validity, authenticity, interactiveness, washback impact and practicality”.

2.4.1. Practicality

Practicality refers to how doable a test is in deference to financial conditions, time, scoring, and equipment. Davies et al., (2002) mention the term ‘practicality’ in a manner as follows:

the term practicality covers a range of issues, such as the cost of development and maintenance, test length, ease of marking, time required to administer the test (individual or group administration), ease of administration (including availability of suitable interviewers and raters, availability of appropriate room or rooms) and equipment required (computers, language laboratory, etc.) (Davies et al., 2002, p. 148).

A good test should be practical in its design, management in terms of learners’ physical and mental efforts, time. Besides, it should be economical with regard to money and time, and easy to score and report its results. Harrison (1983, p. 13) argues that “tests should be as economical as possible in time (preparation, sitting, marking) and in cost (material and hidden costs of time spent)”.

Practicality is a substantial notion which a teacher should not ignore while he/she holds an examination. Since, it is not important how valid or reliable a test is, if it is not practical, it may be difficult to apply it. Harries (1969) indicates that:

a test may be a highly reliable and valid instrument but still be beyond our means of facilities. Thus in the preparation of a new test or the adoption of an existing one, we must keep in mind a number of very practical considerations (Harries, 1969, p. 21).

2.4.2. Authenticity

Authenticity is a significant principle of language tests which has been argued among language testing researchers over the last decades. Authenticity is the answer of

the question how likely the language task given in a test will be performed in the real world. Authenticity is the degree to which test tasks corresponds to real life language use (Brown and Hudson, 2002; Halleck and Moder, 1995; Hoekje and Linnel, 1994; Lewkowicz, 2000).

Authenticity refers to correspondence of features of a language test task to features of real language task. Brown (1996) states that:

to decide how authentic a test is, the below characteristics can be taken into consideration:

1. The language in the tests is as natural as possible.
2. Items are contextualized rather than isolated.
3. Topics are meaningful (relevant, interesting) for the learners.
4. Some thematic organization to items is provided such as a story line or episode.
5. Tasks represent or closely approximate real-world tasks (Brown, 1996, p. 28).

An authentic test includes authentic language, and language tasks replicate real world tasks. For example, in a speaking test, the given task should be similar to real life conversations.

2.4.3. Washback

Washback as a principle of language tests has been discussed in language testing widely as it has a direct impact on students and teachers. Cheng, Watanabe, and Curtis, 2004, p. 13) cited that “washback or backwash, a term now commonly used in applied linguistics, refers to the influence of testing on teaching and learning”. Despite the fact that washback is a commonly used term in applied linguistics today, it is not easy to find the meaning of it. But the word backwash is used in some dictionaries and is defined as the unpleasant after-effects of an event by Oxford Dictionary. This negative connotation of the definition is surprising because washback has positive and negative effect on teaching and learning process.

2.4.4. Reliability

Reliability refers to the consistency of measurement. It means that the scores from a particular student in a different time and under the various testing conditions remain same or equivalent to the same or similar test and similar conditions. Imane (2013, p. 7) cited in his article that reliability is defined as “the degree to which test

scores for a group of test takers are consistent over repeated application of a measurement procedure and hence are inferred to be repeatable for an individual test taker”.

Reliability, as a principle of language testing, gains a big importance. Since the consistency of tests shows the accuracy of testing. If test scores are relatively inconsistent, they may not provide the information about language abilities and knowledge what a teacher wants to measure. However; the reliability of language tests is affected by some factors. Brown (1996) states that:

the issue of reliability of a test may be best be addressed by considering a number of factors that may contribute to the unreliability of a test. Consider following possibilities: fluctuations in the student, in scoring, in test administration and in test itself(Brown, 1996, p. 21).

These factors are categorized and named differently by several researchers. For example; Henning (1987) classifies these factors in four main categories such as fluctuations in the learner, fluctuations in scoring, fluctuation in test administration, and test characteristics affecting reliability.(Henning, 1987, pp. 75-80). On the other hand, Bachman (1990, p. 164) classifies these factors as test methods facets, attributes of test taker and unsystematic or random factors.

Although there are some other suggestions by some researchers as it is mentioned above, Frisbie (1988, p. 27) categorizes these factors in two as “random errors” which are the errors affect each student in a different way, and “systematic errors” which are the errors affecting all students to nearly some extent and cause a high or low score for all students. Random errors occur at one time by chance and change the score in an unpredictable way. Variety of chances in students such as mood, health status, forgetting, inattentiveness may cause these random errors. But, systematic errors may occur because of faulty in test item, scoring problems, interruptions in the test administration. Systematic errors decrease or increase all students’ scores in the same extent.

2.4.4.1. Types of Reliability

Reliability is one of the two essential qualities. Bachman and Palmer (1996, p. 19) argue that “two of the qualities- reliability and validity- are however, critical for

tests, and are sometimes referred to as essential measurement qualities”. Due to having so much importance, reliability can be defined by several methods. Brown (1996, p. 193) classifies these methods into three strategies such as the test-retest, equivalent-forms, and internal consistency strategies. Harries (1969, p. 15) speaks of retest, parallel forms and split-half as types of estimates of reliability. Henning (1987, p. 80-82) categorizes them as test- retest method, inter-rater method, parallel forms reliability, and split-half reliability.

While there are several methods for estimating test reliability, test- retest, inter-rater reliability, and parallel or alternate forms reliability are the basic methods which are mentioned in this study.

2.4.4.1.1. Test- Retest Method

Among four basic reliability methods, test- retest is one of the simplest and appropriate methods to estimate the consistency of tests. Firstly, a test is applied to a single group of students in two different times. Typically, the two separate applications are only a few days or a few weeks apart; the time should be long enough for students to forget the questions and short enough not to learn something new. The connection between the students' scores from the two different applications is estimated to determine how similar the scores are. This type of reliability demonstrates to what extent a test is consistent. Henning (1987) indicates that:

the simplest computational method, test –retest, is calculated by means of product-moment correlation of two sets of scores for the same persons. The same test is administrated to the same people following an interval of no more than two weeks. The purpose of interval limitation is to prevent changes in the examinee’s true score which would cause underestimation of reliability by this method(Henning, 1987, p. 81).

Applying a test two times is not something attractive to the teacher and students. However, this method is the most practical and appropriate method to estimate test reliability.

2.4.4.1.2. Inter Rater Reliability

Inter-rater reliability method seeks the answer of the question whether a student gets the same result if another teacher scores his test. Inter-rater reliability is related to

consistency of scores which is expected among teachers. Brown (1996, p. 203) points out that “interrater reliability is essentially a variation of the equivalent-forms type of reliability in that the scores are usually produced by two raters, the scores”.

2.4.4.1.3. Parallel Forms (Alternate Forms)

Parallel forms are different sorts of the same test which measure the same kind of language ability or knowledge. Parallel forms are equivalent in length, difficulty, time, and format. Parallel forms reliability is computed by administering a test and its parallel forms to the same group of students. The correlation between student's two scores provides the information about consistency of test. Davies et. al. (2002) indicate that:

ideally, alternate forms of a test are two or more tests designed to the same specifications. Each form should correspond in terms of: number of test items, type of item, content of items, difficulty level of item, as well as in instructions, time allowed etc.(Davies et al., 2002, p. 95).

2.4.4.2. Reliability of Teacher Made Examinations

Many studies have been carried about reliability in language testing field and in these studies, the term reliability, threats to test reliability; types of reliability are defined and explained in details by various researchers (Bachman, 1990; Brown, 1996; Harrison, 1983). However; there are a few studies on the reliability of teacher-made exams.

A teacher has two basic roles: instructor and examinant and combines these roles and makes use of each one interchangeably. He uses a test score to make decisions about the next instruction or teaches by taking into consideration the deficiencies. Scores from teacher-made tests are used by teachers and students for different purposes. Scores from the tests, for a class, may show if learning has been completed or if teaching has been effective. For one student, it may give an idea about his/her learning process. Hence, the scores of tests should be consistent and reliable otherwise; inadequate information about learning and teaching process may cause big problems related to language acquisition.

2.4.5. Validity

Validity is another basic principle in language testing and researchers have focused on the importance of validation in language testing since 1960s. Generally speaking, Oxford English Dictionary defines the word validate as “to demonstrate that something is true.” In other words, it is used as a means of verification. However, the definition of validity or validation gains a theoretical and educational background in language testing. American Psychological Association refers to the concept of validity as “appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores”(Kunnan, 2000). In this sense, it is possible to say that validity confirms function and usage of tests in education. It guides teachers, and makes their process of assessment meaningful because the teacher knows what he is testing and why. He is aware of what measures learning process through validity. Ruch (1924) describes the functions of validity as:

by validity is meant the degree to which a test or examination measures what it purports to measure. Validity might also be expressed more simply as the worthwhileness of an examination (Ruch, 1924, p. 13).

Worthwhileness of a test is parallel with what is intended to measure. If the teacher is sure that he is measuring what he taught then he suits the principles of validity. In this sense, validity standardizes tests by excluding any irrelevant criteria that does not suit the aim of assessment. However, American psychologist Samuel Messick changed understanding of validity from a property of a test into interpretation of test results. As he states:

validity in an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or modes of assessment(Messick, 1996, p. 20).

2.4.5.1. Types of Validity

Reliability and validity are closely connected to each other and a test may not have a meaningful reliability without validity. In order to be valid, a test should be reliable. Imane argues that (2013, p. 7) “reliability is essential yet not sufficient to establish a measure’s validity”. In this sense, as a principle of language test validity is prior but relevant to reliability and other principles. Hills (1981, p. 11) states that “there

is nothing more important to be demanding of a test than validity; if a test is not valid, it is worthless”.

Like reliability, in order to establish validity of a test which is a complex procedure, types of validity are defined. Due to having so much importance, reliability can be defined by several methods. Hughes (1989) mentions criterion related validity, content validity, face validity, and construct validity as types of validity. Fulcher and Davidson (2007) categorize validity as construct validity, criterion related validity and content validity. However, Brown (1996, p. 22-27) discusses criterion related evidence, content-related evidence, construct-related evidence, consequential validity, and face validity.

Researchers define types of validity in several ways and in this study, content validity, construct validity, face validity and criterion related validity will be defined to understand validity better.

2.4.5.1.1. Content Validity

As a simplest definition, content validity refers to relationship between the test and the instructional objectives and is based on the knowledge of test content to specify whether the test assesses specific subject areas which are the goals of instruction. Sireci (1998) claims that:

Content validity refers to the degree to which a test measures the content domain it purports to measure. To argue that a test is valid for a particular testing purpose, it must be shown that the items and tasks composing the test are representative of the targeted content domain (Sireci, 1998, p. 229).

A test which involves what is required to assess has content validity. For a content valid test, a test should be abided to curriculum and syllabus. Therefore, comparing test content with syllabus and a content specification is a common way to assess the content validity of a test. Brown (1996) states that:

the most feasible rule of thumb for achieving content validity is to test performance directly, not indirectly. For example, if the test is intended to test learners’ oral production of syllable stress and the given task is to

have learners mark stressed syllables in a list of written words, this is considered as indirect testing of oral proficiency. A direct test of syllable production requires that students actually produce target words orally (Brown, 1996, p. 24).

2.4.5.1.2. Construct Validity

Construct validity refers to a measure of extent which a test assesses a hypothetical or unobservable construct or ability. More generally, a test has construct validity if it is proved that the test just assesses the ability or construct which is meant to be assessed. The term construct means the underlying ability which is hypothesized in a theory of language ability. Buck (2001) points out that:

construct validity is the extent to which the test measures the right construct. It is not the tests that have validity, but decisions based on tests... A test cannot measure the right construct if it is not measuring anything at all. Establishing construct validity is not a simple task (Buck, 2001, p. 195).

Brown (1996) exemplifies construct validity as follows:

Imagine, for example, that you have been given a procedure for conducting an oral interview. The scoring analysis for the interview includes several factors in the final score: pronunciation, fluency, grammatical accuracy, vocabulary use, and sociolinguistic appropriateness. The justification for these five factors lies in the theoretical construct that claims those factors to be major components of oral proficiency. So if you were asked to do an oral proficiency interview that evaluated only pronunciation and grammar, you could be justifiably suspicious about the construct validity of that test (Brown, 1996, p. 25).

2.4.5.1.3. Criterion Related Validity

Criterion related validity is defined as the degree to which the test scores agree with one or more outcome criteria. Davidson and Lynch (2002, p. 133) indicate that “criterion-related validation—to establish that a test agrees with some external criterion which seems to measure the same skill(s) as the test in question”. There are two types of criterion-related validity: predictive validity and concurrent validity (Brown, 1996;

Davies et al., 2002; Weir, 2005). Concurrent validity is established when the criterion manner and the administration of a test occur at the same time, but predictive validity refers to what extent a test can predict student's future performance (Bachman, 1990; Brown, 1996; Hughes, 1989).

2.4.5.1.4. Face Validity

Face validity is a simple type of validity which is concerned with correlation between the appearance of a test and what it aims to assess. In this sense, it refers to what extent a test is valid according to students who take it, the authority who makes judgments on its use, and other untrained individuals (Alderson, Clapham, and Wall, 1995; Anastasi, 1982; Davies et al., 2002). So, a test has face validity if a student, teacher or an individual understands what is assessed in the test by looking at it.

2.4.5.2. Validity of Teacher Made Examinations

Numerous research studies have been conducted on validity of language testing, the types of it, how to prepare a more valid test (Anastasi, 1982; Brualdi, 1999; Rudner, 1994). However, there is little research on the topic of validity of teachers made exams and teachers' perceptions of validity.

This study will not investigate the validity of teacher-made tests. It will emphasize on teachers' perception about reliability and validity.

CHAPTER THREE

3. METHODOLOGY

3.1. Introduction

This chapter centres upon methodology used in this study to investigate the teachers' perceptions of reliability and validity of tests which are prepared by teachers. In the first part of this chapter; research design, description of checklist and interview are explained in details. In the second part, the implementation of checklists and interviews are explained in two parts as pilot study and main study with participants, setting, data collection and data analysis.

3.2. Research Design

In this study, a qualitative study was required to investigate teachers' perceptions of reliability and validity of teacher-made tests because this study aims to examine the teachers' perception about reliability and validity and the influence of their perceptions on administration process of teacher- made tests. Denzin and Lincoln (2011) state that:

Qualitative research is a method of inquiry employed in many different academic disciplines, traditionally in the social sciences, but also in market research and further contexts (Denzin and Lincoln, 2011, p. 7).

3.3. Participants

The current study was carried out in different schools in Siirt, Turkey. There were 23 teachers participated this study with different educational background, teaching experience, working in different kinds of schools. Fourteen of participants were female and nine were male. Their mother tongue was Turkish. Question 1 and 2 in the first section of checklist collected data about teachers' educational background. In sense of their educational background, twenty of participants received bachelor degree and three of them got master's degree. Table 1 demonstrates the distribution of teachers according to the department they have graduated and Table 2 shows their teaching experiences.

Tablo 1 Teachers' Educational Backgrounds

Degree	Number of Teachers
Department of English Language Teaching	15
Department of English Language and Literature	6
Others	2
Total	23

Question 3 in the first section of checklist collected data about teachers' teaching experience, and Table 2 shows the number of teachers according to their teaching experience.

Tablo 2 Teachers' Teaching Experiences

Years of Teaching Experience	Number of Teachers
1 to 4 years	13
5 to 8 years	4
9 to 12 years	1
More than 13 years	5
Total	23

Question 4 sought for an answer to how many of teachers had taken a course on language testing. According to the answer of this question, nineteen of the teachers had a course on language testing and four of them didn't have a language testing course.

Question 5 in the first section of checklist was asked to teacher to get information about the types of the school where these teachers had been working. Table 3 shows the results of this question in a detailed way.

Tablo 3 Types of the Schools in Which Teachers Work

Type of the school	Number of the teacher
State primary school	3
State secondary school	7
State high school	11
Private primary school	2
Total	23

3.4 Instruments

Two data collection tools were used in this study. The first data collection tool was a checklist with three parts and thirty-three questions. This checklist is given in Appendix A. The aim of this checklist is to study out what teachers do during testing procedures for reliability and validity.

The second data collection tool is a semi-structured interview. The interview questions are given in Appendix B. All of the teachers who answered the checklists were interviewed. The aim of interviewing is to find out whether teachers know the terms “reliability” and “validity”, what teachers do to promote reliability and validity. How the instruments were formed and used is explained in details in the next chapter;

3.5 Procedural Details.

3.5. Procedural Details

The current study was conducted in several schools; private primary schools, state primary schools, state secondary schools, and state high schools in Siirt, Turkey. During two weeks, teachers filled the checklist and answered the interview questions in teachers’ rooms and sometimes in managers’ rooms. Before the interview, the researcher asked participants’ permission to record the answer. During the interview sessions, the answers of sixteen participants were recorded and the rest were written down as the other participants did not want their interviews to be recorded. Only one of the teachers wanted to interview in English, the others asked to answer and be asked the questions in Turkish and said that it would be easier, more understandable and less threatening to talk in Turkish.

Data was collected in this study through checklists and interviews. In the first step of this study, the researcher used a checklist on reliability and validity in order to search out the teachers’ knowledge and perceptions of reliability and validity. In the second step of this study, teachers were interviewed and aimed to find out the answer of what reliability and validity are, which producers teachers use to provide reliability and validity of teacher-made tests.

3.5.1 Confirmation and Description of Checklist and Interview Questions

The checklist and interview questions were confirmed after a detailed investigation of literature and writing literature review. Before applying the pilot study,

the researcher asked four English language teachers' opinions about the questions and terms in checklist and interview questions. These teachers were asked to answer several questions related to checklist and interview questions to modify checklist and interview questions. (see Appendix C) In deference to their ideas and suggestions, essential modifications were made and first drafts of checklist and interview questions were laid out.

The checklist applied in pilot study had three sections. First section was aimed to inform the researcher about participants' personal information. Second section had eight questions about reliability and the last section had sixteen questions. After the pilot study, the researcher discovered that the second and the third sections were needed to be revised, and she modified the items and the number of items in the checklist. There were four questions which were asked to participant through the interviews. These questions were also changed and six questions were decided to be asked in the interviews in main study.

The questions in checklist and interview questions were prepared according to research questions in order to find out teachers' knowledge and perceptions of reliability and validity of teacher-made tests. The checklist has three sections first of which is introduction section. The first section includes five questions related to teachers' personal information. In these sections, questions were about teachers' educational background, teaching experience, and the school they had been working.

Forming the checklist and interview questions considering the pilot study, there were ten questions in the second section related to knowledge about reliability. Five of these questions were open- ended while the others were yes-no questions. By favour of this section, it is aimed to find out what teachers take into consideration to provide reliability while administrating teacher-made test. (see Appendix A for checklist) The table below describes distributions of the factors which result in unreliability.

In the third section, there were eighteen questions relevant to validity. Thirteen of these questions are yes- no questions and the rest is open-ended. By way of these questions, this section aims to be informed about teachers' attitudes towards validity and what they do in testing process for validity. (see Appendix A for checklist) The divisions of the items according to types of validity are shown in the table below.

In the second step of this study, a semi-structured interview with six questions was made with the teachers. Three of these questions were related to reliability while the other three questions were relevant to validity. (see Appendix B)

3.5.2 Piloting

Due to the deficiencies in studies about validity and reliability of teacher-made tests, and teachers' perceptions of validity and reliability, the researcher decided to conduct a pilot study. It was possible to confront with various problems since the current study was held with several teachers with different educational backgrounds and field information. With the help of pilot study, researcher aimed to find out the deficiencies and possible problems in her study and modify them before conducting the main study.

The pilot study was conducted with six English language teachers who had been working in different schools and had different work experience and educational background. Three of these participants were graduated from English language teaching department, two were graduated from English language and literature department. The other participant were graduated from German language teaching department and worked as an English teacher in a state high school. As a first step, a checklist with twenty questions was given to participants. (see Appendix D) Then these participants were interviewed.

3.5.2.1 Implications for the Main Study

The researcher made use of the analysis of the pilot study as it provided essential information about the lacks of the checklist and interview questions. At first, it was found out the number and the content of items was not sufficient and explanatory to get required information about reliability and validity. In order to fill this deficiency, the researcher added some more items related to reliability and validity and used more illustrative items. Secondly, it was understood that interview questions were explanatory enough to find out teachers' perceptions and knowledge of reliability and validity, their suggestions to provide reliability and validity. Therefore, by the help of pilot study, the researcher modified her questions.

3.5.2.2 Main Study

By dint of piloting and getting rid of deficiencies, the researcher had an opportunity to conduct the main study with too little problems and limitations. The main

study lasted from 26 May to 6 June covering two weeks as the researcher visited several schools and distributed checklists to teachers and interviewed with them. The researcher visited nine schools and distributed the checklist to and interviewed with twenty-three teachers with different educational and academic background. Teachers filled the checklist in the teachers' room without any inferences but interviewed one by one and alone.

3.6 Data Analysis

The researcher analyzed the results of the checklist and identified what percentage of teachers had knowledge of reliability and validity and what procedures they took into consideration while administering a test. The researcher put out down teachers' answers given in the interviews and categorized to interpret the results more easily and trustworthily. The categories were used to explain the results and find out if the participants used the key words of reliability and validity.

CHAPTER FOUR

4. DATA ANALYSIS AND RESULTS

4.1. Introduction

In this chapter, the data analysis results of instruments used in current research are presented. In the light of the objectives of this study, the researcher used a checklist and an interview consisting of six questions. The findings of checklists and interviews were identified and categorized in tables.

4.2. Findings from Checklists

The researcher analyzed the results of checklist and identified the findings of close-ended questions in tables item by item with their percentages. Open ended questions in the checklist were interpreted subsequently. The results are presented as follows:

Tablo 4 Teachers' Perceptions of Reliability

Question	Yes	No
1. Do you look at the students' names while grading the exam papers?	10 43,4%	13 56,6%
2. Do you have an answer key for each exam paper?	17 73,9%	6 23,1%
3. Do you take time and the length into consideration while preparing the exam questions?	20 86,9%	3 13,1%
4. Do you grade all exam papers in a day?	8 34,7%	15 65,3%
5. While grading exam papers, do you take students' class performance into consideration?	10 43,4%	13 56,6%
6. (If you have a listening section in your exams) do you take care of the conditions such as noise, tape recorder volume etc.?	16 69,5%	7 30,5%
7. Do you prepare the exam questions that help to identify differences in achievement among students?	20 86,9%	3 13,1%
Total Number	23	

As seen in the Table 4, percentages vary in responses. These variances were generally consistent with categorization but sometimes were not. For instance, it is observed that Q3, Q6 and Q7 which may be categorized under the heading of test and test administration reliability had almost same percentage value when the results were analyzed. However, Q1, Q4 and Q5 might be categorized under the same title; rater reliability but the results of the questions have shown that teachers do not pay attention to these items equally. To be more precise, as a result answers given to Q1 and Q5, almost 57 % of teachers paid attention to promote reliability in tests without their interference to grades by avoiding looking at students' names and not taking their class performances into consideration, however, as observed in Q4, just nearly 35% of teachers graded exam papers in different days which was an alternative way to promote reliability.

The two other questions in this part are open ended questions. The 8th question which was "if you have a rubric to assess the productive skills such as writing, to what extent do you use it?" was answered by twelve teachers as: "I don't use a rubric." One teacher said that he usually used it. The rest did not answer this question. The last question was asked to get information about what teachers did to promote reliability of a test. Seven teachers said that they had no idea about what might promote reliability in tests while the others had different suggestions. Duration of the exam and school rules were considered to be a way to promote reliability by eight teachers. On the other hand, five teachers suggested separating successful and unsuccessful students. Besides, three teachers advised to take students' attitudes and level into consideration for reliability. Another teacher laid stress on teaching style as a criterion which affected reliability. According to seven teachers, the most important criterion for promoting reliability was preparing questions which were related to course subjects.

It can be concluded that nearly all teachers were aware of test reliability in some degree and tried to prepare reliable tests. However, when they are specifically asked what reliability reminded them and what they understood from reliability, surprisingly it is seen that only five of the teachers recalled and defined the term reliability. To make it clear, as a result of analysis of checklists and interviews, it is observed that T2, T4, T12,

T19, and T20 gave consistent answers to checklist questions and interview questions about what reliability reminded them. To exemplify, it is understood that these five teachers had the knowledge of reliability according to the analysis of the checklist and their answers to Q1. The following statements set examples for teachers' comprehension of test reliability.

T19: "Reliability is the degree to which an assessment tool produces stable and consistent results."

T20: "If you give the same test to same students or compare students on different occasions, the test should give similar results. This is reliability."

Furthermore, Q8 in the checklist and Q5 of the interview gather information about what teachers do to promote test reliability. However, throughout the analyses, it is found out that almost all of the teachers' answers to these questions are incoherent. To illustrate, T4 stated that she took school rules and students' class performances into consideration as a reply to Q8 while she replied Q5 as she did nothing to promote reliability.

Tablo 5 Teachers' Perceptions of Validity

Question	Yes	No
1.Do you pay attention to the instructions to be understandable and clear?	22 95,65 %	1 4,34 %
2. Do you sometime have legibility problems in your exam paper because of breakdown of the photocopier or other technical problems?	16 69,56 %	7 30,43 %
3. Do you pay attention to spelling mistakes while preparing the exam questions?	20 86,95 %	3 13,04 %
4. Do you think your students can understand easily what is being asked?	18 76,26 %	5 21,73 %
5. Do you use various types of questions in your exams?	23 100 %	0
6. Do you pay attention to use a question type that easily measures what you want to measure?	20 86,95 %	3 13,04 %
7. Do you have different sections for each skill and grammar? For example grammar section, writing section, vocabulary section etc.	21 91.30 %	2 8.7 %
8. Do you have a speaking exam? (answer the question if you have speaking lessons)	22 99,65 %	1 4,34 %
9. Do you measure all the subjects you have taught in your exams?	20 86,95 %	3 13,04 %
10. Do you use other teachers' written exams?	7 30,43 %	16 69,56 %
11. Do you use past exams that you have prepared for other classes?	14 60,9%	9 39,1%
12. (If you have writing section in the exam) Do you decide some criteria before grading this part?	18 78,27%	5 21,73
13. Do you pay attention to the lesson objectives when you prepare the exam questions?	23 100%	0
Total Number	23	

The questions which are analyzed as shown in Table 5 are related to validity. As observed in Table 5, there are a few changes in the percentages of answers. These changes were generally consistent with categorization but sometimes were not. For example, it is observed that Q3, Q5, Q7, Q8 and Q12 which may be categorized under the title of construct validity have almost same percentage value. However, Q6, Q10 and Q11 might be collected under the same heading; criterion-related validity but the analyses of the questions have shown that teachers do not put emphasis on this kind of validity evenly. To be more precise, in the light of answers given to Q6, 86.95 % of teachers paid attention to promote criterion-related validity in tests by using types of questions that easily measured what they wanted to measure, meanwhile, 69.56 % of teachers did not answer Q10 which was an issue that might be considered to promote validity. Besides as seen in Q11, just almost 39.1 % of teachers did not use the previous exams that were prepared for other classes as an alternative way to promote validity.

The other five questions in this part are open ended questions. The four of these questions gathered information about what teachers took into consideration while grading grammar, reading, listening, and vocabulary sections. Q14 which was “What do you take into consideration in grammar section while grading the exam papers? (spelling mistakes, grammar mistakes, etc.)” was answered by eleven teachers as “Grammar mistakes are important for me while grading.” Seven teachers said that they paid attention both to grammar and spelling mistakes. “What do you take into consideration in reading section while grading the exam papers? (spelling mistakes, grammar mistakes, comprehensibility etc.)” was Q15. Six teachers answered the question as “comprehensibility is an issue that is important in reading section.” Eight teachers stated that comprehensibility and spelling mistakes were significant criteria in reading section. Five teachers expressed that they took spelling mistakes into considerations. Four teachers gave no answer to this question. Q16 was “What do you take into consideration in listening section while grading the exam papers? (spelling mistakes, grammar mistakes, comprehensibility etc.)” Eight teachers mentioned that there were no listening sections in their tests. Three teachers indicated that they paid attention to understanding of listening text while grading. Twelve teachers represented that spelling and grammar mistakes were important. Q17 which was “What do you take into consideration in vocabulary section while grading the exam papers? (spelling

mistakes, grammar mistakes, comprehensibility etc.)” was answered as ‘Spelling is an inseparable part of vocabulary’ by seven teachers. Nine of the teachers stated that meaning was the only issue that they attached importance to. One teacher said that using a word in a sentence was the criterion for grading vocabulary. The rest did not answer the question. The last question was asked to gather information about what teachers did to promote validity of a test. Teachers gave more than one suggestion and their answers to Q18 are shown in Table 6.

Table 6 The Ways Teachers Promote Validity

Answers	N
Time	1
Multiple Choice Questions	4
Using Different Types of Questions	4
Having an Answer key	2
Measuring All Subjects	7
Not answered	8

Teachers’ partial awareness of test validity and their efforts to prepare valid tests are concluded from their answers to the checklist. However, when they are specifically asked about their knowledge on validity, none of the teachers could clearly define the term validity although nine of the teachers could describe the term content validity correctly. To exemplify, throughout the analyses of checklist questions and interview questions, nine teachers made comments on content validity as follows:

T5: “For a valid test, the test questions should contain course subjects.”

T9: “When I think of validity, the first thing that comes to my mind is the necessity of asking questions related to course subjects and the syllabus.”

T18: “Preparing a test that involves all course subjects is validity.”

Additionally, question 18 in the checklist and question 6 in the interview were asked to gather information about what procedures teachers followed to promote validity, but the results of analyses showed that nearly all of the teachers gave different answers to these

two questions. For example, T6 replied Q18 as “to promote validity, topics, duration of the exam and school rules are important” while she stated that she paid attention to difficulty level in her answer to Q6.

4.3. Findings from Interviews

A semi structured interview with six questions about reliability and validity was applied to twenty three teachers with different educational background and from different schools. At the first stage of analysis, the teachers’ answers to each interview question were coded and then some categories and subcategories were obtained according to the codes. The findings of each interview question were presented.

4.3.1. Question 1: What does reliability remind you?

According to analysis of answers to this question, six categories were obtained. Table 7 gives the categories and number of teachers who gave responses that fell into each category. The total number of teachers is more than twenty three because some of the teachers gave more than one response.

Table 7 Teacher’s Perceptions of Reliability

Categories	Subcategories	N
Consistency	-in exam results	5
	-between class performance and exam results	2
Cheating		8
Questions	-clearness	4
	-course subjects	3
Justice in Evaluation		4
Environmental Conditions		1
Nothing		1

4.3.1.1. Consistency

Consistency was one of the features that came to teachers’ mind when reliability was mentioned. As a general opinion, seven teachers stated that consistency was the first word that they recalled. The researcher subcategorized consistency in two as consistency in exam results and consistency between student’s class performance and grades. Five teachers stated that same results which were gotten from a test done in different times and occasions were an important part of reliability. Therefore the

researcher put these answers under the subcategory named as consistency in exam results and in terms of this subcategory, two teachers referred consistency in exam results and emphasized on different occasions as followed:

T20: "...If you give same test to the same student or compare students on different occasions, the test results should be same or approximately similar to the previous one. If the results are same or similar then we can say that the exam is a reliable one."

T12: "When I think of reliability, the first thing that I can say is exam questions that are always answered in the same or similar way in every condition."

Another teacher mentioned consistency in exam results by emphasizing the term of exam time.

T2: "Reliability reminds me the same results of a test. In other saying, it is a term that is related to the same results of a test that has been done in different times."

The other subcategory was consistency between class performance and exam results. Two teachers indicated this similarity as a sign of reliability. The common point in those statements was that class performance might be a clue to guess the exam result otherwise the test might not be reliable. To ensample, the teachers uttered their statements:

T3: "First of all, there must a similarity between students' performances in the class and the grades they get from an exam."

T23: "A test only can be reliable if results of this test show similarity with students' performance. In other words, if a student is unsuccessful in the class, his grades also should be bad. Then we can talk about reliability."

4.3.1.2. Cheating

During the interviews, some teachers complained about unfair test results because of students' attitudes during the test process. These attitudes were referred as not studying for the test and trying to get help from their classmates and stated to be a cause for unreliability. Eight teachers reported that cheating was the most important factor which caused unreliability in testing. For instance, some of these eight teachers stated that:

T8: "Reliability reminds me the situation that students answer the exam questions by using their own knowledge. I am speaking of cheating. Answering the exam questions on their own proves that the test is a reliable one."

T9: "Student's own answer is reliability for me."

T6: "In a sentence, if students answer the test on their own without any help, this test is reliable."

T13: "Not allowing students to cheat is a sign of reliability."

Although all teachers agreed that not cheating was a criterion which might be taken into account, one of these teachers also mentioned of distinction and fairness between successful and unsuccessful students in exam results and their performances. The teacher reported that:

T17: "... A lazy student must not get a higher mark than a successful student. If a lazy student's mark is higher, we cannot talk about reliability. The student who gets a good mark should have a good class performance."

4.3.1.3 Questions

This category pointed out the teachers' ideas about how questions might affect the reliability. Seven teachers indicated that reliability was related with questions in a test. In terms of teachers' answers to interview questions, the researcher subcategorized questions into two as course subjects and clearness. Four of these teachers stated that they took notice of course subjects when they prepared the exam questions. For instance, teachers expressed that:

T4: "... I prepare exam questions according to subjects I have thought. The lesson subjects gain importance when reliability is mentioned. I think reliability has a strong relationship with lessons subjects."

T7: "When I am asked what I can say about reliability, the first thing I can say is the subjects that have been taught are also the subjects of exam questions."

T16: "The only sentence I can say is that assessing what students have learned is reliability. For example a test which does not assess a lesson subject that has been taught is not reliable."

Clearness was the second subcategory and defined as comprehensibleness of exam questions which might affect reliability of a test. Three teachers revealed that ambiguity was a threat for reliability and coping with this ambiguity was required to have a reliable test. The following statements set examples for this subcategory:

T1: "The clearness and understandability of exam questions are very important issues for me. I always check the exam questions one by one to find out whether they are understandable or not. ..."

T19: "There must not be an ambiguity in exam questions. The questions should be clear, and understandable."

T22: "Reliability is word that reminds me unambiguous questions. According to me, a test should be understandable otherwise we cannot talk about reliability."

4.3.1.4. Justice in Evaluation

Justice in evaluation was another category obtained from the analysis of the interview data. Justice in evaluation was defined as being objective and not considering class performance or emotional ties while evaluating exam papers. Four of the teachers reported that justice in evaluation was a significant issue that might be considered while grading the exam questions. To illustrate, teachers expressed the following statements:

T3: "... A teacher has to pay attention not to bestow a privilege on a student. I try not to check students' names during evaluation."

T5: "... The evaluation should be equitable."

T21: "... In my opinion, reliability is accuracy for me. Thus, I make effort to be fair, grade equally and use an answer keys."

4.3.1.5. Environmental Conditions

One of the teachers indicated that environmental condition was one of the factors that caused unreliability and a considerable issue to take into account in testing process. The following statement set examples for this category:

T11: “The conditions in the class should be available for testing. Namely each student has to feel comfortable. There must not be noise, too much sunshine in the class.”

4.3.1.6. Nothing

One of the teachers mentioned that she knew nothing about reliability and could not say even a word for this term. To exemplify, she stated that:

T15: “Nothing. I have no idea about what reliability is in testing. I just want to think of it.”

4.3.2 Question 2: What does validity remind you?

In the analysis of this interview question, the researcher obtained four categories and some subcategories as shown in the Table 8.

Table 8 Teacher’s Perceptions of Validity

Categories	Subcategories	N
Appropriateness	-level	2
	-objectives	5
Cheating		2
Subjects	-subjects taught in class	10
	-equality among course subjects	5
Objectivity in Evaluation		4

4.3.2.1. Appropriateness

Almost two out of three teachers who were interviewed called attention to suitability of questions to students’ levels and goals of the lesson. The researcher separated this category into two as appropriateness to level and appropriateness to

lesson objectives. Two teachers put emphasize on being aware of students' knowledge and language ability or their language level. In the following sentences, teachers said:

T2: "As far as I am concerned, a valid test is prepared by considering students' level ad language knowledge."

T20: "When I think of validity, level is the first word that comes to my mind. Level is a factor that affects the validity positively or negatively."

The other subcategory was objectives and it was identified as appropriateness of test questions to the syllabus and goals. Five teachers spoke of objectives differentially. Three teachers stated that assessing what was aimed to assess was a sign of validity. The following statements were examples for this subcategory:

T2: "Validity is assessing the target knowledge and ability."

T19: "Validity refers to how well a test measures what is purported to measure."

T20: "If the extent made from assessment results is meaningful, appropriate and useful in terms of the purpose of the assessment, it means that the test is valid."

However, two teachers accepted objectives as a guide in order to prepare exam questions. As the following statements reported:

T9: "The curriculum is in the centre of testing. For a valid test, the exam questions should be prepared according to curriculum."

T23: "The syllabus is like my holly book in testing. The exam questions are always related to the objectives of the lesson as a valid test is also a test that assess the objectives of a lesson."

4.3.2.2. Cheating

In the analysis process, the researcher defined cheating as a criterion that affected validity in negative ways and caused a problem for it. Two of the teachers mentioned that cheating was a factor that might be considered not to have an invalid test. To illustrate, the following statements reported:

T11: "... In order to have a valid exam, cheating should be annihilated. Cheating changes the results of a test in an unfair way so as a teacher we have to be very careful of not letting the students to cheat. Otherwise; our testing process will be invalid."

T6: "Validity is a sub-heading of testing and essential to assess students properly. Thus, I never tolerate cheating since cheating has a bad effect on test results and ruins the accuracy and validity of the test."

4.3.2.3. Subjects

Throughout the analysis of interview data, fifteen teachers stated the importance of course subjects and their impact on validity. Therefore, the researcher collected these statements under the same category named subjects. However, their ideas differentiated in some way. While two out of three of these teachers called attention to the topics that were taught in the lesson, the rest just spoke of the equality among lesson subjects in testing process. Because of this differentiation, two subcategories were formed. One of these subcategories called subjects taught in the class and ten teachers expressed that asking students what was taught was a cue of validity. The following statements set examples for this subcategory:

T5: "... The exam questions should contain the topics that have been taught in the lesson. At that case, we can talk about validity of that test."

T7: "Testing is a tool to see what students have learned and consequently it must involve all the subjects that the teacher has taught."

T8: "Lesson topics are the first things to be considered while preparing exam questions. If we do not ask all the subjects that are taught, how valid the test could be."

T14: "A valid test involves all the subjects that have been taught in the lesson."

The second subcategory was equality among subjects and it was mentioned by five teachers as a must for validity. Equality among subjects might be defined as uniformity

of course subjects or emphasizing equally each course subject and asking same number of questions for each subject. To illustrate, teachers uttered that:

T15: “The distribution of exam questions according to lesson subjects should be equal. Each subject has to gain the same importance in the test. Then, this test may be valid.”

T4: “Validation refers to including lesson subjects equally in a test.”

T21: “A valid test is a test that contains all the lesson subjects in the test coequally in number.”

4.3.2.4. Objectivity in evaluation

In the analysis process, the researcher defined objectivity in evaluations as teachers’ objectivity in grading the exam papers and not favouring a student because of his/ her class performance. As the following statement reported:

T9: “A teacher cannot speak of validity if s/he does not treat to all students coequally during the evaluation of exam paper. The same or similar answers are to be marked same.”

T12: “I think, validity refers to objectivity in evaluation.”

4.3.3. Question 3: To what extent is reliability important in testing? Why?

Data obtained from question 3 was thoroughly examined and categorized as shown in Table 9. Almost all of the teachers agreed on the great importance of reliability in testing. Only one teacher stated that she had no idea about the importance of reliability.

Table 9 Reasons for the Importance of Reliability

Categories	Subcategories	N
Students	- level	5
	- differentiation	6
Tests	- trusting to the test	4
	- Reflecting the truth	6
Achieving objectives		3
Having no Idea		1

4.3.3.1. Students

This category pointed out the importance of reliability in terms of students. Substantially, teachers uttered that reliability was a capital importance since a test was used to get information about students' levels, their knowledge and ability and distinguish them from each other in terms of their abilities and academic successes. In this context, the researcher formed three subcategories in difference to the common words expressed by teachers. The first subcategory was named as level, since two teachers attributed the importance of reliability to use a test as a tool to decide on students' levels. They uttered sentences similar to following:

T2: "It is very important because testing is used to determine the students' level and determining students' level properly has a big role for the education in the future."

T23: "It is very important of course. The importance of reliability cannot be ignored as testing is a tool to specify students' level."

In the analysis process, the researcher defined differentiation as finding out students' disparateness such as their success, their domain of interest, their knowledge and their abilities. To illustrate, teachers reflected their thoughts as the following utterances:

T11: "Reliability is extremely important because testing is a key to distinguish students from each other in way of their success. A teacher can easily

understand which student is successful and which one is not under the favor of testing. Only a reliable test can be trustworthy.”

T18: “It is very important. If a test does not tell the difference among students with regard to their achievement, how reliable it can be.”

T20: “It is important. A reliable test refers to a reliable differentiation in class.”

4.3.3.2. Tests

Teachers attached importance to tests in terms of reliability. Ten teachers used the word ‘test’ in common when they were asked why reliability was important. Thus the researcher formed a category as tests. However; two subcategories were obtained since teachers indicated that a reliable test reflected the correct results and give confidence to both teachers and students. Trusting the test was the first subcategory that was mentioned by teachers as followed:

T9: “Reliability is extremely important. An unreliable test does not evaluate truthfully and may not give confidence to both teachers and students. In order to feel comfortable, I have to rely on exam results. Thus, reliability gains a big importance in testing for me.”

T12: “It is very important. A test should have correct results that students can trust it.”

T21: “Absolutely, reliability is important because reliable test is also a trustworthy test.”

Reflecting correct results was the common usage of six teachers. They repeatedly used the word “correct results” and its importance for reliability. The below statements set examples for this category:

T4: “Reliability has a big importance in testing as it reflects the same and correct results. In other words, if a test gives different results in each implication, it is not possible to speak of correct results and unreliability occurs as a result of incorrect exam results.”

T22: “Of course, it is very important. An unreliable test will result in misevaluation.”

4.3.3.3. Achieving objectives

Achieving objectives was another category obtained from interview data. A teacher stated that reliability was an inseparable part of testing and very important because of being a sign of achieving the targets of lessons. Namely, a reliable test referred to a test that measured all objectives. For instance, teachers mentioned that:

T1: “Reliability is important in terms of assessing students according to the objectives. A reliable test achieves to assess the lesson objectives.”

T13: “... A reliable test succeeded in testing the objectives.”

4.3.3.4. Having no idea

One of the teachers indicated that she had no idea about to what extent reliability was important. The following statements exemplified this category:

T15: “I have no idea about this question. It may be important or not.”

4.3.4. Question 4: To what extent is validity important in testing? Why?

In the light of the analysis, data from question 4 was categorized and subcategorized as in Table 10

Table 10 Reasons for the Importance of Validity

Categories	Subcategories	N
Evaluation	- objectives	6
	- teachers’ performance	5
Intelligibility		10
Evidence for Reliability		5
Paying no Attention		1

4.3.4.1. Evaluation

Evaluation was created as a category which described teachers’ ideas about evaluation of their performances and the lesson objectives. Teachers attached importance to validity in testing because of giving current and accurate information

about teachers' performance and to what extent the lesson objects were achieved. Six of the teachers indicated that validity of a test had importance since a valid test referred to assess the students in terms of lesson objectives. They stated sentences similar to following:

T1: "Validity is significant with regards to understanding which objectives have been accomplished and which ones are to be achieved."

T3: "Validity is crucial since it gives a chance to find out whether the objectives have been achieved or not."

Evaluation of teachers' performances was the second subcategory. Five teachers indicated the significance of validity in terms of evaluating what teachers did in the class. The common point in those utterances was how the evaluation of teachers' performances is related to validity. For instance, teachers stated that:

T9: "Validity is substantial for teachers. A teacher evaluates his/her own performance by the help of testing. He/she can find out his/her deficiencies in teaching process. Therefore, a valid test will give this teacher the essential and accurate information."

T15: "Validity is very important in terms of evaluating our own teaching process by testing and finding out the deficiencies."

4.3.4.2. Intelligibility

This category pointed out the importance of validity in terms of the understandability of topics. Teachers stated that they paid attention to validity in testing as they obtained the essential accurate information about whether students learned course subjects or not. The following statements set examples for this category.

T6: "I give importance to validity. If I did not pay attention to validity, I would not be able to find out which lesson subjects were learned."

T18: "It is crucial to get information about students' learning process, and what they have learned so far."

T8: "A valid test should test the entire lesson topics that have been learned to show the teachers and students which ones has been learned."

4.3.4.3. Evidence for reliability

Evidence for reliability was another category obtained from the analysis of question 4. Teachers mentioned that validity was very important for reliability and a valid test was reliable too. The below utterances exemplified this category:

T7: "Reliability and validity are related to each other. A valid test is reliable too."

T19: "While reliability is necessary, it is not sufficient alone. For a test to be reliable, it also needs to be valid. ..."

T20: "Without validity, a test cannot be reliable."

4.3.4.4. Paying no attention

A teacher expressed that he knew nothing about validity and also did not realize the term before answering the checklist questions. Thus, he thought that validity was not important. To illustrate, the teacher reported that:

T21: "I have never paid attention to validity. Indeed, I had not noticed the term till you mentioned it. For me, it is not important because I know nothing about it."

4.3.5. Question 5: What procedures do you follow to promote reliability?

Data obtained from question 5 examined and categorized as shown in Table 7.5. Some categories subcategorized in order to examine the data in detail.

Table 11 Suggestions for Promoting Reliability

Categories	Subcategories	N
Questions	- understandability	3
	- different types	4
	- difficulty	2
	- number	1
	- answer key	1
Repetition	- course subjects	2
	- exam	2
Cheating		7
Do Nothing		3

4.3.5.1. Questions

This category represented teachers' choices of questions in the way of clearness, number, difficulty, type and using an answer key. The researcher formed five subcategories in order to make the analysis more understandable and clear. The first subcategory was named understandability and defined as the clearness of the exam questions which had a notable effect on students in testing process. Teachers indicated that ambiguity was a danger for reliability in testing and because of that they took cognizance of clearness and understandability of exam questions. For instance, teachers reported that:

T10: "I prepare the exam questions by myself. Therefore I check the questions a few times to see if they are understandable or not."

T19: "I do not write ambiguous items. I try questions to be understandable and clear for students. As understanding the question is the first and most important step of answering the exam questions."

Teachers expressed that using different types of questions in tests was a way to promote reliability. Asking different questions for the same course subject was referred to be a good way for testing students' language knowledge and ability as it increased the reliability of a test. The following statements were examples for this subcategory:

T3: "I use different types of questions such as open-ended questions; fill in the blanks questions etc. to promote reliability. I think if a student answers all or almost all questions that are related to the same lesson subjects but in different types, the test is reliable."

T20: "I prepare a test that has too many questions and in this test, there are multiple choice, yes/ no questions, open ended questions, true false questions and the other types that I cannot remember now. I think I have a reliable test."

Difficulty was described as another procedure that was followed by teachers to have a more reliable test. Two teachers pointed out that too difficult or too easy questions cause a big problem for reliability. They drew attention to the level of difficulty as an issue that might be taken into consideration during the testing process. These teachers reported similar statements as in the following:

T6: "The level of difficulty is significant for me. A question should be difficult enough to assess the target knowledge."

T12: "I consider important the students' levels. The exam questions should not be too easy or too difficult for students."

Number was another subcategory that might be defined as having too many questions or a few questions in a test. One teacher stated that he used too many questions to prepare a reliable test. He stated a sentence similar to following:

T2: "In testing process, number of the questions is an issue that I always attach importance to. I do my best to decide how many questions will be enough both in terms of testing the knowledge and time."

One of the teachers reported that he used an answer key to promote reliability. He stated that:

T3: "... I always use an answer key for reliable results."

4.3.5.2. Repetition

Repetition was another category obtained from the analysis of question 5. Repetition was defined as teaching the lesson subjects again and again in case of ambiguity and re-examination. Four of the teachers reported that repetition was sometimes crucial and essential as students either did not understand the lesson topic or got a very bad mark from the test. As a subcategory, repetition of lesson subjects was used by two teachers as a procedure to promote reliability. To illustrate, teachers expressed the following statements

T20: "I always teach the lesson subjects several times and before each exam I repeat the lesson subjects. I think this repetition promotes the reliability of my tests."

T23: "Teaching over and over is the way I use for a reliable test. Students learn better and get higher marks."

Two teachers expressed that re-examination was an alternative way for them to promote the reliability of their tests. For instance, they stated that:

T22: "If I realize that there is unfairness, I make another exam. Otherwise how I can speak of a reliable test."

T17: "I sometimes make another exam to be sure of reliability."

4.3.5.3. Cheating

The researcher defined cheating as a factor that caused unreliability in testing. Seven teachers revealed that cheating was as a threat for reliability and coping with it was a necessity to promote a reliable test. The following statements set examples for cheating:

T5: "For a reliable test, getting rid of cheating is a must. Thus, I change students' places to avoid cheating."

T11: "Cheating results in unreliability. Thus, in my exams, successful and unsuccessful students sit on different places. So it becomes hard to cheat from successful students."

T13: "I change the successful student's place to get rid of cheating. Since cheating is a criterion that affect the reliability of a test in a bad way."

4.3.5.4. Do nothing

Three teachers indicated that they did nothing to promote reliability of a test. For example, they reported as followed:

T4: "I do not do anything for a more reliable test."

T16: "I never follow a procedure to have a reliable test. I just prepare an exam and make it."

T21: "I never do anything."

4.3.6. Question 6: What procedures do you follow to promote validity?

The researcher examined and categorized the data obtained from question 6 as seen in Table 7.6.

Table 12 Suggestions for Promoting Validity

Categories	Subcategories	N
Questions	- understandability	2
	- variety	6
Objectives		11
Do Nothing		2

4.3.6.1. Questions

All along t the analysis of interview data, eight teachers mentioned the importance of questions and their impact on promoting validity. Therefore, the researcher collected these data under the same category named questions. However, teachers' ideas differ in some ways. Two of the teachers pay attention to the understandability of questions, while the other six spoke of the importance of variety of questions in order to promote validity. Because of this disparateness, the researcher formed two subcategories. One of these subcategories named understandability and teachers talked about understandability as a way to promote validity of a test that they made. The following statements set examples for this subcategory:

T1 : "I pay great attentions to prepare questions that can be understood easily by students. As in my opinion, for a more valid test, understandable questions are a must."

T8: "Understandable tests are also valid tests. Therefore I try to make my questions understandable and clear. If a question is clear enough, we can talk about a valid test."

The second subcategory is variety that is defined as different kinds of questions on the same subject in a test. Six teachers expressed that variety was a very important criterion for a valid test. The teachers reported that statements as examples:

T19: "I use different questions in my exams. There are all kinds of questions such as multiple choice, comprehension questions, writing part etc."

T21: "Variety is one of the issues that I take into consideration while preparing exam questions."

T23: "To promote validity, I use various questions in my exams. Different kinds of questions are a key for validity."

4.3.6.2. Objectives

This category points out the noteworthiness of objectives as another way to promote validity. Eleven teachers set forth that objective was an issue that they took into consideration during preparing test questions as objectives specified what to teach in the lessons and teachers taught according to them. The following statements were the examples of this category:

T3: "I make an effort to ask questions that include all lesson objectives. Because I make my lesson plans and teach by taking these objectives into consideration."

T7: "Preparing the exam questions according to yearly plan is a way of promoting validity."

T22: "Curriculum has the main importance in promoting validity. If exam questions are appropriate the curricula and objectives, then it is valid."

T2: "I try to prepare exam questions that contain lesson subjects that I have taught and done exercises about."

T11: "Validity is important so I think in order to promote validity, I should ask questions only that I teach in the lesson. Otherwise; the exam will be invalid."

T20: "I especially take the content validity into consideration. I measure all the materials I have taught."

T23: "I aim to measure all the items I have taught. I ask questions from all units."

4.3.6.3. Do Nothing

Two teachers explained that they did nothing to promote validity. For instance, they stated that:

T16: "To speak honestly, I do nothing for that. I don't think validity is a significant issue for tests."

T18: "Nothing. I never follow a procedure or do something to promote validity."

4.4. Conclusion

In this chapter, the data obtained from the checklists and the interviews were analyzed and presented in two parts. In the first part, the data consisted of checklists was analyzed under the headings teachers' perceptions of reliability and teachers' perceptions of validity. In the second part, the data was analyzed and some categories and subcategories obtained and given in details in tables. Then the results were discussed.

CHAPTER FIVE

5. CONCLUSION

5.1. Introduction

This chapter summarizes the study and then it discusses data collected from interviews and checklists by elaborating on teachers' perception of validity and reliability that are common terms crucial in language testing. Finally, suggestions for language instructors and further studies are provided in the scope of this thesis

5.2. Summary of the Study

The researcher aimed to find out teachers' perceptions of validity and reliability of teacher-made tests. With this study, what teachers did to enhance the validity and reliability of teacher-made tests was discovered.

A qualitative research method was used in this study. After a detailed revision of the related literature in the field of testing, the researcher prepared research questions. A checklist and an interview with six questions were prepared and piloted in order to collect data. Furthermore, some English Language Teachers were consulted during the pilot study and with reference to their ideas and results of the pilot study; the checklist and interview questions were modified.

The checklist and the interview questions were replied by 23 English Language teachers in Siirt, Turkey from different schools and with different educational backgrounds. These teachers were aware of lesson curriculum and experienced about testing. These questions were analyzed and the results were shown in Tables. The results of interview questions were coded and explained in a detailed way.

5.3. Discussion

The terms validity and reliability, which are the focal points of this study, are two basic concepts that are supposed to be known by English Language teachers. Language teaching cannot be taken into consideration without testing and testing itself remains incomplete without a comprehensive knowledge of validity and reliability. Although many studies have been carried out about these two terms and they are widely mentioned in Testing and Evaluation courses in English Language Teaching and related

departments at universities, this study digs out surprising findings about teachers' understanding of reliability and validity.

In the light of research questions, “what do teachers think about reliability and validity in teacher made examinations?” and “what do teachers for reliability and validity in exams which they prepare?” following findings were obtained. First of all, this study demonstrates that teachers' knowledge of reliability and validity do not differ much. Considering theoretical literature of reliability and validity, few of the teachers participated in this study could define both validity and reliability properly. To clarify, just five of the teachers mention that reliability is related to the consistency in exam results but also they add the terms cheating, class performance, course subjects which are not included in the definition of reliability. Moreover, five of the teachers state that validity is the extent to which a test assesses what it aims to assess. However, none of the teachers speak of the terms related to types of reliability while only one teacher uses the term content validity which is one of the types of validity.

In fact most of the teachers' answers to the checklist questions show that they are aware of factors determining reliability and validity but even those teachers do not know that those factors are component to promote reliability and validity. For example, according to the answers given to the checklist questions, 73,9 % of participants, which is 17 out of 23, mention that they use an answer key for each exam paper they prepare. However, when they are asked about what they do to promote reliability in the exams they prepare in the interview and open ended questions in the checklist, only 1 person says that he uses an answer key to prepare a more reliable test. Similarly, participants' answers for promotion of validity shows the same inconsistency. When they are asked whether they pay attention to instructions to be understandable and clear, 95,65 %, which is 22 out of 23, of the participants state that they pay attention. Nonetheless, only two of the teachers mention about understandability of questions when they were asked what they did to promote validity. This inconsistency demonstrates that most of the teachers are not aware theoretical literature while they assess but they unconsciously provide reliability and validity.

While grading different skills, teachers are supposed to take construct validity into consideration. To provide construct validity, which is a type of validity, teachers

assess each language skill with its required assessment criteria, For example, in a listening test, elements such as comprehension, vocabulary usage are required to be taken into consideration while grading. However, 12 participants in this study mentioned that that pay attention to spelling and grammar mistakes while grading listening exam paper. This shows that participants some implementations while grading different skills may sometimes affect validity negatively.

As a result, in the light of this study, it is possible to say that teachers' perception of reliability and validity seems to be inadequate to prepare reliable and valid tests. This inadequacy not only affects testing process but also learning and teaching process itself because test results are used by teachers and students to find out the deficiencies of learning and teaching process, to plan next step in teaching and improve the process. Some suggestions for annihilating this inadequacy may be applying talent development programs and having more courses related to testing and evaluations in English Language Teaching Department.

5.4. Limitations and Implications for Further Study

This study set forwards new studies concerning assessment in language teaching process. Further studies can be listed as the follow:

- The researcher of this study used only checklist and interview to collect data from language teachers. Similar studies with other means of collecting data can be carried out as well.
- Number of the participants of this study is limited to 23 people and only one district of the country, this limitation can be widened with more people and some other different parts of the country.
- If integration of all skills and areas in terms of lessons and exams are provided, further studies can be conducted more effectively.

6. REFERENCES

- Alderson, J. C., Clapham, C., & Wall, D. (1995): *Language Test Construction and Assesment*. Cambridge: Cambridge University Press.
- Anastasi, A. (1982): *Psychological testing* (5th ed.). New York
- Bachman, L. F. (1990): *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996): *Language testing in practice : designing and developing useful language tests*. Oxford: Oxford University Press.
- Brown, H. D., & Abeywickrama, P. (2010): *Language assessment : principles and classroom practices*. White Plains, NY: Pearson Education.
- Brown, J. D. (1996): *Testing in language programs*. Upper Saddle River, N.J.: Prentice Hall Regents.
- Brown, J. D., & Hudson, T. (2002): *Criterion-referenced language testing*. Cambridge, UK ; New York: Cambridge University Press.
- Brualdi, A. (1999): Traditional and modern concepts of validity. *ERIC Document Reproduction Service*. doi: ED435714)
- Buck, G. (2001): *Assessing listening*. Cambridge ; New York: Cambridge University Press.
- Cheng, L., Watanabe, Y. J., & Curtis, A. (2004): *Washback in language testing : research contexts and methods*. Mahwah, N.J.: Lawrence Erlbaum.
- Coombe, C. A., Folse, K. S., & Hubley, N. J. (2007): *A practical guide to assessing English laugage learners*. Ann Arbor, Mich.: University of Michigan.
- Davidson, F., & Lynch, B. K. (2002): *Testcraft : a teacher's guide to writing and using language test specifications*. New Haven ; London: Yale University Press.
- Davies, A. (1990). *Principles of language testing*. Oxford: Basil Blackwell.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (2002): *Dictionary of Language Testing*. Cambridge: Cambridge University Press.
- Denzin, N. K., & Lincoln, Y. S. (2011): *The Sage handbook of qualitative research* (4th ed.). Thousand Oaks, Calif.: Sage.

- Fulcher, G., & Davidson, F. (2007): *Language Testing and Assessment an Advanced Resource Book*. New York: Routledge.
- Frisbie, D. A. (1988): Reliability of Scores From Teacher-Made Tests. *Instructional Topics in Educational Measurement*, 16(28), 25-35.
- Fulcher, G. (2010): *Practical language testing*. London: Hodder Education.
- Gage, N. L., & Berliner, D. C. (1979): *Educational psychology* (2nd ed.). Chicago, IL: Rand McNally.
- Halleck, G. B., & Moder, C. L. (1995): Testing language and teaching skills of international teaching assistants: The limits of compensatory strategies. *TESOL Quarterly*, 29, 733-758.
- Harries, D. P. (1969): *Testing English as a Second Language* New York-USA: McGraw-Hill.
- Harrison, A. (1983): *A language testing handbook*. London: Macmillan.
- Heaton, J. B. (1990): *Classroom testing*. London: Longman.
- Henning, G. (1987): *A guide to language testing : development, evaluation, research*. Cambridge, Mass.: Newberry House Publishers.
- Hills, J. R. (1981): *Measurement and evaluation in the classroom* (2nd ed.). Columbus: Merrill.
- Hoekje, B., & Linnel, K. (1994): "Authenticity" in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly*, 28, 103-123.
- Hughes, A. (1989): *Testing for language teachers*. Cambridge Cambridgeshire ; New York: Cambridge University Press.
- Imane, B. N. E. (2013): *Conceptualizing Essay Tests' Reliability and Validity: From Research to Theory*. Huzhong Normal University. China.
- Ioannou-Georgiou, S. (2003): *Assesing Young Learners*. Oxford: Oxford University Press.

- Kunnan, A. J. (2000): *Fairness and validation in language assessment : selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida*. Cambridge: Cambridge University Press.
- Lewkowicz, J. (2000): Authenticity in language testing: Some outstanding questions. *Language Testing*, 17, 43-64.
- McNamara, T. F. (2000): *Language testing*. Oxford: Oxford University Press.
- Messick, S. (1996): Validity and Washback in Language Testing (pp. 1-20). Princeton, New Jersey: Educational Testing Service.
- Mulvenon, S. W., & Dunn, K. E. (2009): A Critical Review of Research on Formative Assessment: The Limited Scientific Evidence of the Impact of Formative Assessment in Education. *Practical Assessment, Research & Evaluation*, 14(7), 1-11.
- Nakarin, P. (2007): Building Effective English Teacher Made Tests: A Major KPI for Thai ESP Teachers. *Hatyai Journal*, 5(10), 71-79.
- Owen, C. (1997): *Testing*. Birmingham: Center for English Language Studies.
- Ruch, G. M. (1924): *The improvement of the written examination*. Chicago, Atlanta etc.: Scott.
- Rudner, L. M. (1994): Questions to ask when evaluating tests. *ERIC Document Reproduction Service*. doi: ED385607)
- Siddiek, A. G. (2010): The Impact of Test Content Validity on Language Teaching and Learning. *Asian Social Science*, 6(12), 133-143.
- Sireci, S. G. (1998): Gathering and Analyzing Content Validity Data. *ERIC*. doi: EJ588889
- Walt, J. L. v. d., & Steyn, F. (2008): The validation of language tests. *Stellenbosch Papers in Linguistics*, 38, 191-204.
- Weir, C. J. (2005): *Language testing and validation : an evidence-based approach*. Basingstoke: Palgrave Macmillan.

William, D. (2000): *Integrating Summative and Formative Functions of Assessment*.
Paper presented at the European Association for Educational Assessment,
Prague: Czech Republic.

Wrightstone, J. W. (1961): Teacher Made Tests and Techniques. *Journal of Association
for Supervision and Curriculum Development*, 19(42), 170-200.

7. APPENDIXES

7.1. Appendix A: Checklist

Part 1

Please, tick () the suitable answer for you.

1. Which department did you graduate from?

Department of English Language Teaching () Department of English Language and Literature ()

Department of Translation and Interpreting () Others ()

2. Which program did you last graduate from?

B.A. degree () M.A. degree ()

3.. How long have you been teaching totally?

1 to 4 years () 9 to 12 years ()

5 to 8 years () more than 13 years ()

4.. Have you taken any courses on testing

Yes () No ()

5. What kind of school do you work?

State Primary School () Private Primary School ()

State Secondary School () Private secondary School ()

State High School () Private High School ()

Part 2

		YES	NO
1.	Do you look at the students' names while grading the exam papers?		
2.	Do you have an answer key for each exam paper ?		
3.	Do you take the time and the length into consideration while preparing the exam questions?		
4.	Do you grade all exam papers in a day?		
5.	If you have a rubric to assess the productive skills such as writing, to what extent do you use it?		
6.	(If you have listening section in your exams) do you take care of the conditions such as noise, tape recorder volume etc.?		
7.	Do you prepare the exam questions that help to identify differences in achievement among students?		
8.	Do you grade the exam papers question by question or student by student?		
9.	While grading the exam papers, do you take students' class performance into consideration?		
10.	What promotes reliability of your exams while you are preparing exams in your opinion ? Your comments (such as time, school rules, other teachers' attitudes etc.)		

Part 3

		YES	NO
1.	Do you pay attention to the instructions to be understandable and clear?		
2.	Do you sometimes have legibility problems in your exam papers because of breakdown of photocopier or other technical problems?		
3.	Do you pay attention to spelling mistakes while preparing the exam questions?		
4.	Do you think your students can understand easily what is being asked?		
5.	Do you use various types of questions in your exams?		
6.*	Do you pay attention to use a question type that easily measures what you want to measure?		
7.	Do you have different sections for each skill and grammar? For example grammar section, writing section, vocabulary section etc.		
8.	Do you have a speaking exam? (answer the question if you have speaking lessons)		
9.	Do you measure all the subjects you have taught in your exams?		
10.	Do you use other teachers' written exams?		
11.	Do you use past exams that you have prepared for other classes?		
12.	(If you have writing section in the exam) Do you decide some criteria before grading this part?		
13.	Do you pay attention to the lesson objectives when you prepare the exam questions?		
14.	What do you take into consideration in grammar section while grading the exam papers? (Spelling mistakes, grammar mistakes, etc.)		
15.	What do you take into consideration in reading section while grading the exam papers? (Spelling mistakes, grammar mistakes, comprehensibility etc.)		
16.	What do you take into consideration in listening section while grading the exam papers? (Spelling mistakes, grammar mistakes, etc.)		
17.	What do you take into consideration in vocabulary section while grading the exam papers?		
18.	What may promote validity of the exams in your opinion ? your suggestions		

*For example using multiple choice question for measuring grammar, or wh- questions in reading section etc.

7.2. Appendix B: Interview Questions

1. What do you understand from reliability?
2. What do you understand from validity?
3. To what extent, is validity important in assessment?
4. What procedure do you follow for the validity of your assessment?
5. To what extent, is reliability in assessment?
6. What procedure do you follow for the reliability of your assessment?

7.3. Appendix C: Piloting the Checklist

1. How long did it take to complete the checklist?

2. Were there any unclear instructions in the checklist ? If so, please state it or them.

3. Were any of interview questions / sentences ambiguous? If so, can you write which and why?

4. Were there any terms or phrases that need explanation? If yes, please write them.

5. Did you object to answering any questions or parts? If so, please state them.

6. Do you think any topics should be omitted from the checklist or interview or included in it? If yes, can you please write it or them?

7. Was the layout of the checklist clear/attractive?

8. Would you like to add any comments?

7.4. Appendix D: Checklist Before Piloting

Part 1

		YES	NO
1.	Do you look at the students' names while grading the exam papers?		
2.	Do you have an answer key for each exam paper ?		
3.	Do you take the time and the length into consideration while preparing the exam questions?		
4.	Do you grade all exam papers in a day?		
5.	(If you have writing section in the exam)Do you decide some criteria before grading this part?		
6.	(If you have listening section in your exams) do you take care of the conditions such as noise, tape recorder volume etc.?		
7.	Do you grade the exam papers question by question or student by student?		
8.	How many questions do you ask in an exam?		

Part 2

		YES	NO
1.	Do you pay attention to the instructions to be understandable and clear?		
2.	Do you sometimes have legibility problems in your exam papers because of breakdown of photocopier or other technical problems?		
3.	Do you pay attention to spelling mistakes while preparing the exam questions?		
4.	Do you think your students can understand easily what is being asked?		
5.	Do you use various types of questions in your exams?		
6.	Do you pay attention to use a question type that easily measures what you want to measure?		
7.	Do you have different sections for each skill and grammar? For example grammar section, writing section, vocabulary section etc.		
8.	Do you have a speaking exam?		
9.	Do you measure all the subjects you have taught in your exams?		
10.	Do you use other teachers' written exams?		
11.	Do you use past exams that you have prepared for other classes?		
12.	Do you use some criteria while grading writing section? Such as comprehensibility, rhetorical discourse elements, organization of ideas?		