

Deanship of Graduate Studies
Al-Quds University

Evaluation of Tawjihi English Tests Based on Norms
of the Construction and Publication of Good
Achievement Tests.

Eman Mohammad Ahmad Hijazi

M.A. Thesis

Jerusalem- Palestine
1428-2007

Evaluation of Tawjihi English Tests Based on Norms of
the Construction and Publication of Good Achievement
Tests.

By:

Eman Mohmmad Ahmad Hijazi

B.A. in English: Subject Area Teacher English
Bethlehem University-Palestine

Supervisor: Prof. Ahmad Fahem Jaber

A Thesis Submitted in Partial Fulfillment of the
Requirement for the Degree of Master of Arts in
Teaching Methods

Education Department- Al-Quds University
1428-2007



Al-Quds University
Deanship of Graduate Studies
Teaching Methods/ Education Department

Thesis Approval

Evaluation of Tawjihi English Tests Based on Norms of the Construction and Publication of Good Achievement Tests

By: Eman Mohmmad Ahmad Hijazi
Registration No.: 20411630

Supervisor: Prof. Ahmad Fahem Jaber

Master thesis submitted and accepted, date: 29 May,2007

The names and signatures of the examining committee members are as follows:

- | | |
|---|------------------|
| 1. Head of Committee: Prof. Ahmad Fahem Jaber | Signature: ----- |
| 2. Internal Examiner: Dr. Ghassan Sirhan | Signature: ----- |
| 3. External Examiner: Dr. Hana Tushyeh | Signature: ----- |

Jerusalem – Palestine

1428- 2007

Dedication

To The Soul of My Father

To My Mother

To My Husband

To My Brothers and Sisters

To My Daughter "*Sema*"

Declaration:

I certify that this thesis submitted for the degree of Master is the result of my own research, except where otherwise acknowledged, and that this thesis (or any part of the same) has not been submitted to higher degree to any other university or institution.

Signed:

Eman Mohmmad Ahmad Hijazi

Date: 29/ 5/ 2007.

Acknowledgments

I wish to express my special thanks and gratitude to my thesis supervisor Prof. Ahmad Fahem Jaber for his continual, fruitful suggestions, guidance, constructive comments, encouragement and his kind necessary collaboration and commitment in the preparation of this work.

Special thanks also go to Dr. Mohsen Adas for his valuable suggestions, encouragement, and readiness to offer help when needed.

I would like to thank Dr. Ghassan Sirhan and Dr. Hana Tushyeh for having agreed to serve on the discussion committee and whose comments have remarkably contributed to the success of the study.

I also extend my thanks and appreciations to Dr. Salah Shrouf who was a good supporter with his experience and the best advice he gave me during my work on this thesis.

At last, but not least, special thanks to my friends especially "Malak and Ibtihal".

Abstract

The purpose of this study is to evaluate the English as a Second Language Tawjihi tests based on norms of the construction and publication of good achievement tests, in Palestine. These tests were prepared by the Ministry of Education in Palestine to assess the ESL students at the end of academic year as a level test in order to allow students following up their higher education. Tawjihi Teacher's analyzed these tests using an instrument prepared by the researcher depending on the previous studies and the literature review. Also, the researcher evaluated Tawjihi tests from 2000-2006.

The study examined the effects of the independent variables on the teacher's evaluation on each domain (gender, experience and qualification). Moreover, the study aim to find out to what extent are the Tawjihi English tests fulfill the norms of the construction and publication of good achievement tests in Palestine and presented the content in different questions formats. Teachers evaluated the tests; also the researcher evaluated the same tests using the same instrument. Then the researcher compared the two evaluations results with each other.

The population of the study consisted of all Tawjihi English teachers in south Hebron in the Academic year 2006-2007. The purposes of the study were investigated using a referred questionnaire which prepared by the researcher and used by the teachers and researcher. The reliability of the questionnaire was tested using a pilot sample from the Directorate of Education/ North Hebron and Cronbach-alpha turned out to be (0.93). To establish its content validity, the researcher gave it to a panel of judges of ten PhD holders in Bethlehem, University Polytechnic University, and Hebron University. The data collected from the questionnaire and the analysis process was statistically analyzed.

The results of the statistical analysis for both teachers and the researcher have shown that Tawjihi English tests are presented the content of curriculum in different questions formats at a medium level. In addition to that, both teachers and researcher agreed that the content of the test wasn't sufficient evidence in Tawjihi tests. Also, the results revealed that instructions and the face validity were presented at high level in Tawjihi tests.

Moreover, the essay questions were the most frequently used format in the tests. The short answer questions, multiple choice questions and cloze questions were main formats used in Tawjihi tests, although the cloze questions were the least frequently format used. The findings of the study showed that matching questions were never used in Tawjihi tests. In addition to that, speaking and listening skills weren't evident at all in Tawjihi tests.

Furthermore, the results indicated that there is no difference in the ratings of the English teachers due to gender and qualification they agreed that Tawjihi tests fulfill the norms of the construction and publication of good achievement tests.

In order to generalize the results obtained from such a study, the author recommended that the study must be applied to other population and the study must be applied a number of times over different periods of academic years. The

Directorate General of Assessment, Evaluation & Examinations should pay more attention to assessment and evaluation methods especially in the Tawjihi English tests construction and publication. Also, The Directorate General of Assessment, Evaluation & Examinations should include the four skills in the Tawjihi Tests especially listening and speaking. Moreover matching questions is very important format to be included in the Tawjihi tests.

المخلص

الهدف من هذه الدراسة هو تقويم الاختبارات التحصيلية لمادة اللغة الإنجليزية للتوجيهي وفق معايير تصميم و إخراج الاختبار التحصيلي الجيد في فلسطين و التي تعدها وزارة التربية و التعليم العالي لتقييم الطلبة في نهاية العام الدراسي حيث يعد اختبار التوجيهي كامتحن مستوى يؤهل الطلبة من متابعة دراستهم الجامعية. قام معلمي التوجيهي بتقييم هذه الاختبارات باستخدام أداة تم إعدادها من قبل الباحثة بناء على الأدب التربوي السابق و الدراسات السابقة. أيضا قامت الباحثة بتقييم اختبارات التوجيهي المعدة ما بين 2000-2006.

فحصت الدراسة تأثير العوامل المستقلة على تقييم المعلمين لاختبارات التوجيهي (الجنس و الخبرة و المؤهل). أيضا إلى أي مدى تراعي امتحانات التوجيهي معايير تصميم و إخراج الاختبار التحصيلي في فلسطين. و مدى شيوع استخدام أنواع الأسئلة المختلفة في الاختبارات التحصيلية للتوجيهي في مادة اللغة الإنجليزية. حيث قام المعلمين بتقييم امتحانات التوجيهي باستخدام أداة الدراسة و من ثم قامت الباحثة بعملية التقييم و مقارنة نتائجها مع نتائج المعلمين.

تكون مجتمع الدراسة من جميع معلمي التوجيهي في مديرية جنوب الخليل للسنة الدراسية 2000-2006. لتحقيق أهداف الدراسة قامت الباحثة بإعداد استبانة و استخدامها من قبل الباحثة و المعلمين. و قد تم التأكد من صدق الاستبانة عن طريق تطبيقها على عينة اختبارية و عددها عشرون معلم و معلمة من مديرية شمال الخليل حيث وجد أن معامل الاتساق الداخلي (كرونباخ الفا) = (0.93). و لبناء صدق المحتوى تم عرض أداة الدراسة على عشرة محكمين من حملة الدكتوراة في جامعة الخليل و جامعة البوليتكنك و جامعة بيت لحم. قد تم معالجة البيانات التي تم الحصول عليها من الاستبانة.

أظهرت نتائج الدراسة من خلال تقييم المعلمين و الباحثة أن اختبارات التوجيهي عرضت محتوى المنهاج بأنواع من الأسئلة المختلفة. على الرغم من أن هذا المحتوى لم يكن على مستوى عال من الجودة. أيضا أظهرت النتائج أن تعليمات الاختبار و الشكل العام للاختبار كانت على مستوى عال من الجودة.

إضافة إلى ذلك أن الأسئلة المقالية من أكثر أنواع الأسئلة شيوعاً، أيضاً الأسئلة القصيرة و أسئلة الاختيار من متعدد و أسئلة التكميل كانت من أكثر الأنواع استخداماً، على الرغم من ان أسئلة التكميل كانت اقلها نسبة. كشفت الدراسة أن أسئلة المطابقة لم تستخدم قط في اختبارات التوجيهي. إضافة إلى إهمال مهارتي الاستماع و التحدث من اختبار التوجيهي.

أشارت نتائج الدراسة أنه لم يكن هناك فروق في معدل تقييم المعلمين للاختبار التوجيهي وفق المتغيرات المستقلة (الجنس و المؤهل)، واتفقوا أن اختبارات التوجيهي توافق معايير تصميم و إخراج الاختبار الجيد.

من اجل جعل نتائج مثل هذا البحث قابلة للتعميم فان الباحثة تقترح تطبيق هذه الدراسة على مجتمع آخر. و أن تطبق عدة مرات لسنوات دراسية مختلفة. و على وزارة التربية و التعليم العالي، الإدارة العامة للقياس و التقويم و الامتحانات أن تأخذ بعين الاعتبار أساليب التقويم و التقييم للاختبارات و خاصة اختبارات التوجيهي للغة الإنجليزية وفق معايير تصميم و إخراج لاختبار الجيد. أيضاً على الإدارة العامة للقياس و التقويم و الامتحانات ان تضمن امتحان التوجيهي مهارات اللغة الانجليزي الاربعة و خاصة الاستماع و التحدث. اسئلة التوافق يجب ان تضمن في امتحان التوجيهي.

Table of Contents

Title	Page No
Declaration	i
Acknowledgement	ii
Abstract	iii
Abstract in Arabic	v
Table of Contents	vii
List of Tables	xi
List of Appendices	xiii
Glossary	xiv
Chapter 1: Introduction	1
1.1 Introduction	1
1.2 Statement of the Problem	3
1.3 Questions of the Study	3
1.4 Significant of the Study	4
1.5 Purpose of the Study	4
1.6 Limitation of the Study	4
Chapter 2: Review of the Literature	6
2.1 Theoretical Literature	6
2.1.1 The Test Formats	7
2.1.1.1 Essay Tests	9
2.1.1.1.1 Short Answer Tests	11
2.1.1.1.2 Extended Answer Tests	11
2.1.1.2 The Objective Test Format	12
2.1.1.2.1 Matching Test	12
2.1.1.2.2 True False Test	13
2.1.1.2.3 Multiple Choice Tests	13
2.1.1.2.4 Completion Test	14
2.1.1.2.5 Cloze Test	15
2.1.2 The Achievement Test Characteristics	15
2.1.2.1 Reliability	15
2.1.2.1.1 Test Factors	15
2.1.2.1.2 Situational Factors	15
2.1.2.1.3 Individual Factors	16
2.1.2.1.1 Methods for Estimating Reliability	16
2.1.2.2.1.1 Parallel from Reliability	16
2.1.2.2.1.2 Test-Retest Reliability	16
2.1.2.2.1.3 Internal Consistency Reliability	16
2.1.2.2.1.4 Inter and Intrajudge Agreement	17
2.1.2.2 Validity	18
2.1.2.2.1 Content Validity	18
2.1.2.2.2 Criterion Validity	18
2.1.2.2.3 Construct Validity	19
2.1.2.2.4 Systematic Validity	19
2.1.2.2.5 Face Validity	19
2.1.3 Test Objectives	19

2.1.4 Test Length	19
2.1.5 Test layout	22
2.1.6 Test Instructions	22
2.2 Preview Studies	22
2.2.1 Foreign Studies	23
2.2.2 Arabic Studies	33
Chapter 3: Methodology	37
3.1 Population	37
3.1.1 Teachers	37
3.1.2 Tawjihi Tests	37
3.2 Instrument of the Study	37
3.2.1 Validity	38
3.2.1.1 Construct Validity	38
3.2.1.2 Content Validity	38
3.2.2 Reliability	39
3.3 Variables	39
3.3.1 Independent Variables	39
3.3.2 Dependent Variables	39
3.4 Procedures of the Study	39
3.5 The Statistical Analysis	40
Chapter 4: Findings of the Study	41
4.1 The Teachers Evaluation	41
4.2 Findings related to the first Question of the Study	41
4.3 Findings related to the second Question of the Study	43
4.4 Findings related to Each Domain of the Study	44
4.4.1 The Instructions	44
4.4.2 The Content Validity	44
4.4.3 The Face Validity	45
4.4.4 The Essay Questions	45
4.4.5 The Short Answer Questions	46
4.4.6 The Cloze Questions	46
4.4.7 The True False Questions	47
4.4.8 The Multiple Choice Questions	47
4.5 Findings related to the Sequence of Domains	48
4.6 Findings related to the Effect of Independent Variables	49
4.6.1 Findings related to the Effect of Gender	49
4.6.2 Findings related to the Effect of Experience	50
4.6.3 Findings related to the Effect of Qualification	50
4.2 The Researcher's Evaluation	51
4.2.1 Findings related to the first Question of the Study	51
4.2.2 Findings related to the second Question of the Study	54
4.2.3 Findings related to Each Domain of the Study	54
4.2.3.1 The Instructions	54
4.2.3.2 The Content Validity	54
4.2.3.3 The Face Validity	55
4.2.3.4 The Essay Questions	55
4.2.3.5 The Short Answer Questions	56
4.2.3.6 The Cloze Questions	56
4.2.3.7 The True False Questions	57
4.2.3.8 The Multiple Choice Questions	57

4.2.5 Findings related to the Sequence of Domains	58
Chapter 5: Discussion of Findings and Recommendation	59
5.1.1 Discussion of findings related to the first Question	59
5.1.2 Discussion of findings related to the Second Question	61
5.1.3 Discussion of findings related to Each Domain of the Study	61
5.1.3.1 The Instructions	61
5.1.3.2 The Content Validity	62
5.1.3.3 The Face Validity	62
5.1.3.4 The Essay Questions	63
5.1.3.5 The Short Answer Questions	63
5.1.3.6 The Cloze Questions	64
5.1.3.7 The True False Questions	65
5.1.3.8 The Multiple Choice Questions	65
5.1.4 Discussion of Findings related to the Sequence of Domains	66
5.1.5 Discussion of Findings related to the Effect of Independent Variables	67
5.1.5.1 Discussion of Findings related to the Effect of Gender	67
5.1.5.2 Discussion of Findings related to o the Effect of Experience	68
5.1.5.3 Discussion of Findings related to the Effect of Qualification	68
5.2 Recommendations of the Study	68
References	70
Arabic References	76
Appendix A	77
Appendix B	80
Appendix C	81
Appendix D	84
Appendix E	85

List of Tables

NO	Table	Page
Table 3-1	The number of teachers in the Directorate of Education /South Hebron.	37
Table 4-1	Means, standard deviations and percentages of teacher's evaluation (The whole questionnaire)	41
Table 4-2	Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Instructions".	44
Table 4-3	Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Content Validity".	45
Table 4-4	Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Face Validity".	45
Table 4-5	Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Essay Questions".	46
Table 4-6	Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Short Answer Questions".	46
Table 4-7	Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Cloze Questions".	47
Table 4-8	Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The True False Questions".	47
Table 4-9	Mean, standard deviation, and percentages for teacher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Multiple Choice Questions"	48
Table 4-10	Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests for each domain in the questionnaire.	48
Table 4-11	Findings related to the effects of the independent variables (gender) on the evaluation of English language teachers of Tawjihi English tests based on the norms of the construction and publication of good achievement tests on each domain in the questionnaire	49
Table 4-12	Findings related to the effects of the independent variables (experience) on the evaluation of English language teachers of Tawjihi English tests based on the norms of the construction and publication of good achievement tests on each domain in the questionnaire.	50
Table 4-13	Findings related to the effects of the independent variables (qualification) on the evaluation of English language teachers of Tawjihi English tests based on the norms of the construction and publication of good achievement tests on each domain in the questionnaire	51
Table	Means, standard deviations and percentages of researcher's evaluation	52

4-14	(The whole questionnaire)	
Table 4-15	Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Instructions".	54
Table 4-16	Means, standard deviations, and percentages for researcher's evaluation of Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Content Validity".	55
Table 4-17	Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Face Validity".	55
Table 4-18	Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Essay Questions".	56
Table 4-19	Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Short Answer Questions".	56
Table 4-20	Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Cloze Questions".	56
Table 4-21	Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The True False Questions".	57
Table 4-22	Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Multiple Choice Questions".	57
Table 4-23	Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests for each domain in the questionnaire.	58
Table 5-1	Discussion of the findings of data analysis related to the First Question of the study. (Teachers and the researcher)	60
Table 5-2	Discussion of the findings of data analysis related to the essay questions. (Teachers and the researcher)	63

Table 5-3	Discussion of the findings of data analysis related to the cloze questions. (Teachers and the researcher)	64
Table 5-4	Discussion of the findings of data analysis related to the multiple choice questions. (Teachers and the researcher	66
Table 5-5	Discussion of the findings of data analysis related to the sequence of domains in the questionnaire. (Teachers and the researcher)	67

List of Appendices

- Appendix A: The Questionnaire before the Judgment.**
- Appendix B: The list of Referees.**
- Appendix C: The Questionnaire after the Judgment.**
- Appendix D: Recommendation letter.**
- Appendix E: The list of Tawjihi Tests (2000-2006).**

Glossary

Test: a device to reinforce learning and to assess students' performance at schools.

Achievement test: A test which aims to establish what has been learned in a course of instruction.

Good achievement test: A test which fulfills the norms of construction and publication good tests like validity, reliability, usability, and different testing formats (T/f questions, multiple choice questions, matching questions, cloze questions, essay questions, and short answer questions)

Tawjihi Test: it is an achievement test prepared by the Directorate of Education in Palestine at the end of twelfth scholastic year as a level test to allow students continuing their higher education.

Norms: international criteria that must be included in the achievement tests.

Chapter 1

Introduction

1.1 Introduction

Testing is a universal feature of human life. Throughout history people have been put to the test to prove their capabilities or to establish their credentials. There are many reasons for developing a critical understanding of the principles and practice of language assessment. Language tests play a powerful role in many people's lives acting as gateways at important transitional moments in education, in employment, and in any field in life. Testing is constructed as a device to reinforce learning and to motivate the students or primarily as a means of assessing the students' performance in the language. A test which sets out to measure students performance as fairly as possible without in any way setting traps for him can be effectively used to motivate the students. A well-constructed classroom test will provide the student with an opportunity to show the recognition and the production of correct forms of the language. Language tests also differ according to their purpose.

In fact, the same form of test may be used for different purposes, although in other cases the purposes may affect the form. The most distinction in terms of test purpose is that between achievement and proficiency tests. Achievement tests are associated with process of instruction. Examples are: end of course tests, portfolio assessments, or observational procedures for recording progress on the basis of classroom work and participation. Achievement tests accumulate evidence during, or at the end of a course of the study in order to see whether and where progress has been made in terms of the goals of learning. Achievement tests should support the teaching to which they relate. Teachers have been critical of the use of multiple choice standardized tests for this purpose, saying that they have a negative effect on classroom as teachers teach to the test, and that there is often a mismatch between the test and the curriculum. Also achievement tests are self-enclosed in the sense that may not bear the direct relationship to language use in the world outside the classroom focusing on knowledge of particular points of grammar or vocabulary. (Gronlund & Linn, 1990)

The evaluation of student's progress and achievement in EFL/ESL classes should be carried out in a manner that doesn't cause anxiety to students. As new EFL/ESL curricula have moved in the direction of developing communicative skills through the integration of language and content as well as language skill integration (listening, reading, writing and speaking), the traditional paper-and-pencil tests no longer cover the variety of activities and tasks that take place in the elementary classroom. The summative form of testing that permeated the traditional curricula wouldn't be fair to students whose studies are based on communicative activities. Fortunately, the field of evaluation has witnessed a major shift from strictly summative testing tools and procedures to a more humanistic approach using informal assessment techniques that stress formative evaluation (O'Neil, 1992).

In all academic settings, evaluation is viewed as closely related to instruction. Evaluation is needed to help teachers and administrators make decisions about students' linguistic abilities, their placement in appropriate levels, and their achievement. The success of any assessment depends on the effective selection and use of appropriate tools and procedures as well as on the proper interpretation of student's performance. Evaluation tools and procedures, in addition to being essential for evaluating students' progress and achievement, also help in evaluating the suitability and effectiveness of the curriculum, the teaching methodology, and the instructional materials. In the past, evaluation tools and procedures were chosen at the level of the Ministry of Education, school district, school administration, or program coordinator. With the advent of learner-centered and communicative teaching methodologies, however, in many settings control over the collection and interpretation of evaluation information has shifted from centralized authority towards the classrooms where assessment occurs on a regular basis. This shift gives the classroom teacher a decisive role in assessing students and makes it necessary for the teacher to look for new assessment techniques to evaluate students achievement and progress. (Fradd and Hudelson, 1995).

The testing tools are characterized by a deliberate move from traditional formal assessment to a less formal, less quantitative framework. The alternative assessment defined as any method of finding out what a student knows or can do that is intended to show growth and inform instruction and is not a standardized or traditional test. Specifically, alternative ways of assessing students take into account variation in students' needs, interests, and learning styles; and they attempt to integrate assessment and learning activities. Also, they indicate successful performance, highlight positive traits, and provide formative rather than summative evaluation. (Pierce and O'Malley, 1992).

Recently the evaluation scene in EFL/ESL classes has been dominated by summative evaluation of learner achievement, focusing on mastery of discrete language points and linguistic accuracy, rather than on communicative competence, with test items typically consisting of matching or gap-filling. Communicative teaching methodology brings with it a considerable emphasis on formative evaluation with more use of descriptive records of learner development in language and learning which language development along with other curricular abilities (Rea-Dickins and Rixon, 1997).

There are some characteristics of evaluation techniques for young learners that they are performance-based and requiring students to perform authentic tasks using oral and written communication skills. These techniques can include traditional classroom activities, such as giving oral reports and writing essays, but they may also involve nontraditional tasks, such as cooperative group work and problem solving. Teachers score the task performances holistically (Shohamy 1995; Wiggins 1998).

In ESL education we should be able to determine the relationship between learner outcomes and the various factors that influence those outcomes, which include

curriculum, classroom instruction, and factors outside the educational setting (e.g., learner personality and learning styles, prior education and life experiences, and opportunities to use English outside the program). This indicates a need for performance evaluation which requires test takers to demonstrate their skills and knowledge in a manner that closely resembles a real life situation or setting. Examples of performance evaluation include oral or written reports, projects, and demonstrations. Performance evaluation isn't easy to develop, administer, score, and validate. For each test developed, we need to know the following:

- Do the test items elicit what learners know and can do?
- Does the test administrator know how to give and score the test?
- Does the interpretation of scores reflect learner knowledge and skills in real-life situations?

Therefore, evaluation becomes a diagnostic tool that provides feedback to the learner and the teacher about the suitability of the curriculum and instructional materials, the effectiveness of the teaching methods, and the strengths and weaknesses of the students. Furthermore, it helps demonstrate to young learners that they are making progress in their linguistic development, which can boost motivation. This encourages students to do more and the teacher to work on refining the process of learning rather than its product. (Katz, 1997).

1.2 Statement of the Problem

Many English teachers in Palestine aren't familiar with the norms of the construction and the publication of good achievement tests in testing English as a second language. They don't know these norms and how to activate them while they design their achievement tests. Therefore, it is so important to make teachers more aware of the presence of these testing norms. This can guide in preparing their tests that suit the testing ESL in the twelfth grade. Also, this study tries to evaluate the Tawjihi tests based on norms of the construction and the publication of good achievement tests in Palestine that are prepared by the Ministry of Education.

1.3 Questions of the Study

The Purpose of the study is to evaluate the Tawjihi English tests based on norms of the construction and publication of good achievement Tests by the English Tawjihi teachers and the researcher (the first two questions only for the researcher, and the five questions for teachers). It also aims to investigate whether teachers' evaluation differs according to gender, teaching experience and qualification. The study attempts to answer the following questions:

1. To what extent do the Tawjihi English tests match the norms of the construction and publication of good achievement tests in Palestine?
2. To what extent do the Tawjihi English tests are presented in different questions formats?
3. How the ratings of female teachers are differ from male teachers?

4. How the ratings of the English teachers are differ due to years of experience?
5. How the ratings of the English teachers are differ due to qualification?

1.4 Significance of the Study

According to the researcher knowledge, this study is the first one that aims to evaluate the ESL achievement tests especially the Tawjihi tests. These tests are prepared by the Ministry of Education in Palestine to assess the ESL students at the end of academic year as a level test in order to allow students continue their high education. This is a new topic in this field. Also, it is very important to evaluate achievement tests in our schools as a main way that evaluates our students by their teachers. Added to this, the importance of the testing process itself which participates in investigating the success in order to develop or enhance it.

In addition to the above, the study reveals the most common testing style used by ESL teachers and gives teaching implications for ESL teachers to use these testing styles. This study is also expected to help teachers to use these norms of the construction and publication of good achievement tests. Moreover, Testing is needed to help teachers and administrators make decisions about student's linguistic abilities, their placement in appropriate levels, and their achievement, in a way that attracts teachers attention and builds students confidence and changes their negative attitudes to English in general and testing in particular to a positive one. This study will lead to further studies dealing with other subjects like Mathematics, Arabic language, History...etc.

Hence, the results of the study are very important to be taken into consideration by the Ministry of Education in constructing the next Tawjihi tests especially after it has developed the new Palestinian curriculum recently.

1.5 Purpose of the Study

This study tries to evaluate Tawjihi English tests based on norms of the construction and the publication of good achievement tests in Palestine that are prepared by the Ministry of Education.

1.6 Limitation of the Study

The study has the following limitations:

1. Generalization of the results will be limited to Tawjihi English achievement tests (2000-2006) the literary stream which belong to the Ministry of Education in Palestine. The results of the study will not be generalized out of these borders.
2. The results of the study will generalize only to the ESL teachers who teach the Tawjihi students in Palestine/ West Bank.
3. This study will confine itself to investigate the Tawjihi English tests based on the norms of the construction and publication of good achievement tests, through asking teachers to do this job.
4. The study will be conducted by using an instrument containing a list of norms in the construction and publication of good achievement tests which was developed by the current researcher.

5. Other limiting factors to be taken into consideration are the concepts, statistical analysis and the procedures of the study.
6. This study is limited to the teachers of English employed by the Ministry of Education who teach Tawjihi in the public schools in south Hebron districts in the scholastic year 2006/ 2007. So, care should be taken in generalizing the findings of this study to other teachers in other school districts.

Testing is needed to help teachers make decisions about the students' abilities, their placement in appropriate levels, and their achievement. This study evaluated the Tawjihi tests based on the norms of construction and publication of good achievement tests. It attempts to find out to what extent do the Tawjihi English tests match the norms of the construction and publication of good achievement tests in Palestine and presented in different questions formats. Also, how the ratings of teachers are differ due to gender, qualification and experience. This study is limited to Tawjihi English achievement tests, and to the ESL teachers who teach Tawjihi in the academic year 2006/2007. This chapter is followed by chapter two which is about the review of the literature.

Chapter 2

Review of the Literature

This chapter aims at demonstrating the available relevant literature to the study. The researcher surveyed the literature and the previous foreign and Arabic studies that discussed in this field. So this chapter is divided in two parts; the first part deals with theoretical literature and the second part deals with the foreign and Arabic studies.

2.1 Theoretical literature

Achievement tests are frequently the major basis for evaluating student's progress in schools. One would have difficulty in conceptualizing an educational system where student isn't exposed to tests. Although the specific purposes of the tests and the intended use of the results may vary from one school to another or from one teacher to another, it is essential that we recognize the value that test results can play a main role in the life of the students, parents, counselor, and other educators.

Wiggins (1998) has used the term 'educative evaluation' to describe techniques and issues that educators should consider when they design and use evaluation. His message is that the nature of evaluation influences what is learned and the degree of meaningful engagement by students in the learning process. Wiggins contends that evaluation should be authentic, with feedback and opportunities for revision to improve rather than simply audit learning, including the following principles:

1. How different assessments affect students?
2. Will students be more engaged if assessment tasks are problem-based?
3. How do students study when they know the test consists of multiple-choice items?
4. What is the nature of feedback, and when is it given to students?
How does evaluation affect student effort?

Answers to such questions help teachers and administrators understand that evaluation has powerful effects on motivation and learning and enhances student achievement. (Black & Wiliam, 1998)

Evaluation impacts student learning and motivation. It also influences the nature of instruction in the classroom. There has been considerable recent literature that has promoted evaluation as something that is integrated with instruction and not an activity that merely audits learning. When evaluation is integrated with instruction it informs teachers about what activities and evaluation will be most useful, what level of teaching is most appropriate, and how summative evaluation provides diagnostic information. For instance, during instruction activities informal, formative evaluation, helps teachers know when to move on, when to ask more questions, when to give more examples, and what responses to student questions

are most appropriate. Standardized test scores, when used appropriately, help teachers understand student strengths and weaknesses to target further instruction. (Shepard, 2000).

Using formative evaluation can help to decrease the level of anxiety generated by concentration on linguistic accuracy and increase student's comfort zone and feeling of success by stressing communicative fluency. Some teachers and researchers call for allowing students to have a say not only in deciding the format of the test but also in deciding its content and the way it is administered. Thus, Mayerhof (1992) suggests allowing students to discuss questions during the test quietly as long as each writes his own answers; she is referring to subjective types of questions. Friel (1989) recommends involving students in suggesting topics for the test or in generating some questions. (Mayerhof, 1992), (Friel, 1989)

Four major skills in communicating through language are often broadly defined as listening, speaking, reading and writing. It is important for the teacher to include those skills in testing students which are:

1. **Listening:** it is the comprehension skill, in which single utterances dialogues, talks and lectures are given to the testee.
2. **Speaking:** the ability, that usually in the form of an interview, a picture description and reading aloud.
3. **Reading comprehension:** which its questions are set to test the students understanding of a written text.
4. **Writing ability:** this skill usually is in the form of essays, letters and reports.

Test can assess integrated reading and writing, reading, writing and listening, or separate them carefully and the test deliberately included input from reading test in a writing task. Teachers find out that there is a problem in doing this. But so far as the distinction into four discrete skills is thought to be either invalid or at least limited and possibly distorting its view of language use. So it is the test constructor's task to assess the relative importance of these skills at the various levels and to devise an accurate means of measuring the student's success in developing these skills.

Most teachers wish to evaluate individual performance, the aim of the classroom test is different from external examination, and so good classroom test will also help to locate the precise difficulties encountered by the class or by individual students. A well constructed test will provide the students with an opportunity to show their ability to recognize and produce correct forms of the language. The purpose of testing in the second language skills is that the students will be able to master some of the required skills in the first language and no guarantee at all that he will be able to transfer those skills to another language. (Heaton, 1997).

2.1.1 The Test Formats

After planning the content and cognitive objectives for the test, teachers must decide on the best way to measure students that is, they decide on the test format.

The format refers to whether the test will be objective (multiple choice, true-false, matching, etc.) or essay. What factors do faculty consider when deciding on the format of the test? Teachers should choose the format that is most appropriate for measuring the cognitive objectives of the test.

Class size is often an important factor influencing the decision about test format. It is very difficult to give essay tests when there is large number of students in the class because the scoring time is prohibitive. A survey of 1100 professors, Cross (1990) showed that class size is the factor that professors consider most important when they decide what test format to use. Two-thirds of the faculty surveyed said they preferred the essay format but could not use it because of the size of their classes. They used essay tests only in small classes. (Cross, 1990)

It is very important for teachers to use a variety of types of alternative testing, especially non-threatening informal techniques, with young EFL/ESL learners. However, there is no claim that these types of testing are without shortcomings. Brown and Hudson(1988) point out that performance evaluation is relatively difficult to produce and relatively time-consuming to administer. Reliability may be problematic because of rater inconsistencies, limited number of observations, and subjectivity in the scoring process. For example, in self-assessment, accuracy of perceptions varies from one student to another and is usually affected by language proficiency (Blanche, 1988).

Regardless of whether the choice is productive response, utilizing a one-word, short-answer , or extended answer, or whether it is a selection response, utilizing a multiple choice, true-false, or matching, knowledge test must be formatted to yield valid assessments of what students know .So to develop an effective test and efficient measure of achievement, one should follow certain strategies :

1. Phrase the item in a clear and understandable manner, with simple and direct wording.
2. Items should be clearly independent of each other to avoid inter item clues.
3. Reading difficulty should be appropriate and below the reading ability of the group of students taking the test. The test item should assess student's knowledge, not their reading skill.
4. Questions and answers should be closely related to objectives from instruction or material used during instruction.
5. Avoid using statements and phrases from the text verbatim in the test item, which should be paraphrased or summarized following three steps:
 - a. Identification of important information.
 - b. Translation of the information into a thought unit
 - c. Establishment of a task or intellectual operation for assessing understanding.
6. Phrase item so the students know what information to include and how to format their answers. The item should be clearly stated to

them in order to compare and contrast present supportive findings, review what is known, explain different views, .etc.

7. Construct the test so the students have enough time to answer the questions
8. Require examinees to respond to all items. The test shouldn't offer a range of items from which students can pick one or two questions because it will provide noncomparable results.
9. Examine the test for item difficulty, discrimination, test reliability, and test validity following its administration.
10. Provide examinees with clues regarding scoring criteria at the time they take the test.

Once individual test items are generated, they should be arranged on the test in systematic manner that is conducive to generating optimal scores from examinees. Following are suggestions from psychometric experts:

1. The group items with the same format in one place on the test. Ideally, all the multiple choice questions should be grouped together, the true-false, as the matching, and so forth.
2. Within each section, group together items of similar content, thereby allowing students to focus on one area of knowledge before moving to another.
3. The intellectual operations and response types depend on the content and objectives being tested.
4. The test should proceed from easy to more difficult items.
5. A very important issue is the provision of directions to examinees. It is absolutely critical that students know how to answer each item. The directions should be both written at the beginning of the test and within each major section. (Tindal & Marston, 1990).

So the test must set out to measure a student's performance as fairly as possible without setting traps for him and can be effectively asked to motivate the students. So there are different formats in testing the skills of English as a second language.

2.1.1.1 Essay Tests

The essay type question requires the examinee to read the question, formulate his response, and write the response. The person scoring the response must be knowledgeable in the area being measured. This type of question can be used to measure many processes: it can require the examinee to make comparisons, to supply definitions, to make interpretations, to make evaluations, or to explain relationships.

There are many uses for essay tests. Teachers prefer to use essay tests because they emphasize the whole subject being measured. Another way to justify their use is the fact that they require the students to supply the response and also essay tests can be used to measure educational objectives which can't otherwise be measured, such as attitudes, creativity, and the ability to organize materials. Also teachers use them to

measure knowledge of facts and principles. Knowledge can be more effectively measured with more objective types of examinations. Essay examinations are also used in assessing the quality of an examinee's higher order mental process: application, analysis, synthesis, and evaluation.

When measuring achievement with items prepared in the essay format, several rather complicated problems are encountered which might be thought of as weakness of this type. The problem can be grouped into three broad categories, the lack of content validity, the lack of scoring economy, and the scorer unreliability. (Tindal & Marston, 1990).

Essay tests enable teachers to judge student's abilities to organize, integrate, interpret material, and express themselves in their own words. Research indicates that students study more efficiently for essay-type examinations than for selection (multiple-choice) tests. Students preparing for essay tests focus on broad issues, general concepts, and interrelationships rather than on specific details, and this studying results in somewhat better student performance regardless of the type of exam they are given. Essay tests also give teachers an opportunity to comment on student's progress, the quality of their thinking, the depth of their understanding, and the difficulties they may be having. However, because essay tests pose only a few questions, their content validity may be low. In addition, the reliability of essay tests is compromised by subjectivity or inconsistencies in grading. (McKeachie, 1986). There are many points of strength to essay items which are:

1. Essay items are an effective way to measure higher level cognitive objectives. They are unique in measuring students' ability to select content, organize and integrate it, and present it in logical prose.
2. They are less time-consuming to construct.
3. They have a good effect on students' learning and students do not memorize facts, but try to get a broad understanding of complex ideas, to see relationships, etc.
4. They present a more realistic task to the student. In real life, questions will not be presented in a multiple-choice format, but will require students to organize and communicate their thoughts.

Although essay items have these point of strengths there are other limitations for this format:

1. They require more time to read and score.
2. They are difficult to score objectively and reliably. Research shows that a number of factors can bias the scoring:
 - A. Different scores assigned by different readers or by the same reader at different times.
 - B. A context effect operates; an essay preceded by a top quality essay receives lower marks than when preceded by a poor quality essay.

- C. Papers that have strong answers to items appearing early in the test and weaker answers later will be better than papers with the weaker answers appearing first.
- D. Scores are influenced by the expectations that the reader has for the student's performance. If the reader has high expectations, a higher score is assigned than if the reader has low expectations. If we have a good impression of the student, we tend to give him/her the benefit of the doubt.
- E. Scores are influenced by quality of handwriting, neatness, spelling, grammar, vocabulary, etc. (Cross, 1990). The essay tests have two types:

2.1.1.1.1 Short- Answer Test

An essay question is considered to be a short – answer question when it contains only one central idea and can be answered in one or two sentences. Items requiring students to supply definitions or short explanations of concepts and relationships fall within this category. Short-answer items employ answers that range from a phrase or sentence to a short paragraph. The stem is divided into two parts: The first establishes the content area and knowledge to be addressed, and the second directs the format and structure of the response. (Bordonaro, 2006).

Short-answer test items have four advantages. First, they can assess higher levels of knowledge or intellectual operations than single word item. More, scoring of responses is easier and likely to be completed with more consistency than for extended answers. Moreover, they can be completed in enough time to include several items of this type. Finally their production format allows a range of variation that probably provides a more accurate reflection of student differences in learning.

Added to the above, there are two major disadvantages. First, it is difficult to write good short answers that delimit the question enough to avoid confusing the students. It is also, difficult to create a clear and objective scoring system. (Tindal & Marston, 1990).

2.1.1.1.2 Extended- Answer Test

The response to an extended answer essay question may be forming one-half page to several pages long. Because of the time required to respond extensively, this type of question should be used only for measuring a student's ability to deal with complex relationships, comparisons, and evaluations.

This item contains two parts, a brief description of an issue, position, problem, or event, and a directive for the student to respond in some manner. The first part of the item should provide both background context and specific information that is being addressed. The second part, should tell students how to structure their responses, it should contain a specific and active verb. (Bordonaro, 2006).

Extended answer responses have several advantages. First, they require students to produce their own answers, rather than recognize a correct answer. This eliminates blind guessing and prevents students from taking advantage of other clues embedded in the test. A second advantage to extended answer responses is that they are more appropriate for assessing complex intellectual operations, synthesizing, organizing, and sequencing large amounts of diverse information. These advantages, however, must be considered in the light of several limitations, because essay tests allow students to produce their own answers and usually include only a few items. First, scoring can be difficult and unreliable. Second, a great deal of scoring time may be needed to get the job done correctly. This disadvantage can be avoided by limiting or directing the type of responses required or by providing a clear scoring key. Finally, it assesses only a relatively small range of behavior to determine student's knowledge. (Tindal & Marston, 1990)

2.1.2 The Objective Test Formats

2.1.2.1 Matching Test

Matching tests typically consist of a list of questions or problems to be answered along with a list of responses. The examinee is required to make an association between each question and a response. Matching tests can be used to measure the lower levels of the cognitive domain. Vocabulary, dates, events, and simple relationships can be efficiently and effectively measured with these items. Furthermore, they may be scored rapidly, accurately, and objectively by individuals who are unqualified to teach in the subject area being examined. (Cross, 1990)

The matching type of items isn't particularly applicable to measurement at the higher levels of the cognitive domain. It is extremely difficult to develop a set of premises for a matching exercise that will measure at the higher levels of the cognitive domain and at the same time, share alternates. In order for a set of matching items to function properly, it must contain homogeneous premises. Otherwise, the differences among premises will provide clues to the correct response. It is difficult to find enough important and homogeneous ideas to form a matching set. Moreover, the construction of a homogeneous set of matching items often places an overemphasis on a rather small portion of the content area to be tested. This may result in failure to conform to the table of specifications and thus, cause a bias in content sampling. Matching tests can be in different forms:

2.1.2.1.1 Word Matching

The students are required to draw a line under the word which is the same as the word on the left. The students in the lower levels can be tested in this way because they like visual things like pictures.

2.1.2.1.2 Sentence Matching

It is similar to the word matching item. The testee is required to recognize as quickly as possible sentences which consist of the same words in the same order. He reads a sentence followed by four similar sentences only one of which is exactly as the previous one.

2.1.2.1.2 Picture and Sentence Matching

Students will concentrate on word and sentence comprehension using pictures test different skill.

Matching items have several advantages. They are easy to produce and can be employed with a wide range of tasks. They also allow for generation of a large number of items since they involve little reading and require few concepts to generate multiple answers. It is possible to generate a great number of items, potentially increasing the amount of behavior sampled on the test. Add to this that the scoring is easy and efficient. There is also a biggest disadvantage that these tasks may be limited to reiteration and summarization of content, thereby reflecting an emphasis on lower rather than higher levels of intellectual operations. (Tindal & Marston, 1990).

2.1.2.2 True/False Tests

The true /false is one of the most widely used to test language. The scores obtained by the testees can be reliable. True /false tests can be constructed easily and quickly allowing the teacher more time for his many other tasks. A true/false test has two main disadvantages: Firstly, it can encourage guessing, since a testee has 50% chance of giving a correct answer for each item. Secondly, as the base score is the 50% the test may fail to discriminate widely enough among the testees unless there are a lot of items. Many teachers argue that true false items encourage students to guess, students guess only when a test is so difficult that they have simply no idea what an answer should be. (McNamara, 2000).

True-false items are particularly appropriate for factual recall information, which is a high priority among achievement testers. However, this format is less useful for assessing more complex intellectual operations and may suffer from the lack of validity because students can guess the correct answer. True-false items require only the presence of statements that are phrased in unequivocal terms, with which examinees must agree or disagree.

True-false items have several advantages and disadvantages in addition to those listed in the more general form of multiple choice formats. They are short and concise, and they are particularly useful for factual information that is central to understanding and they can be constructed quickly, with less attention to the creation of distractors that have an equivalent plausibility as correct answers. Nevertheless, the disadvantages include the limitation to factual information and the high probability of student guessing, given only one foil. (Tindal & Marston, 1990).

2.1.2.3 Multiple-Choice Test

Multiple choice items offer a good way of testing student's language. However it is usually extremely difficult to write four good options (one correct answer and three answers incorrect) for each multiple choice item (McNamara, 2000).

It is common device for testing student's text comprehension. They allow testers to control the range of possible answers to questions and to some extent to control the

students through processes when responding. However, the value of multiple choice questions has been questioned. By virtue of the distractors they may present students with possibilities, they may not otherwise have. The ability to answer multiple choice questions is a separate ability different from the reading, writing, listening and speaking ability. Students can learn how to answer multiple choice questions by eliminating improbable distractors or by various forms of logical analysis of the structure of the question (Alderson, 2000).

The multiple choice item has few weaknesses. Although critics claim that it can be used only to measure factual knowledge, many items have been developed to measure understanding, application of principles analysis, synthesis and evaluation. It is also criticized for not being adaptable to measuring creativity. Even though this is quite true, it is unlikely that items can be written in any test format which will accurately measure this dimension and also measure the acquisition of instructional objectives. The teacher should be skillful enough to acquire considerable time facility in item writing. Compared with true-false item, multiple-choice items need more time to answer. The multiple-choice test is the most flexible and versatile of all selection-type examinations. It may be used to measure instructional objectives at all levels of the cognitive domain: knowledge, application, analysis, synthesis, and evaluation. (Wang, 2000).

Also, it is the most popular item type to appear in modern testing. It predominates in nearly all forms of testing from published norm-referenced to curriculum embedded achievement tests. It is relatively easy to construct, flexible, adaptable to all types and levels of knowledge, capable of generating many items, easy to score, and it has potential of generating reliable results. (Tindal & Marston, 1990).

2.1.2.4. Completion Test

Completion items are useful in testing a student's ability to recall information. They can range from one word completion answers to the completion of a sentence. The completion items which consist of sentence must be have single words missing. Completion tests require the testee to supply a word or a short phrase which measures recall rather than recognition although such items are supply type items and thus similar in many ways to open –ended questions in tests they are often regarded as belonging more to the objectives category of test items (McNarma, 2000).

The completion item is a written statement which requires the examinee to supply the correct word or short phrase in response to an incomplete sentence, a question, or a word association. The completion question has been widely used in workbooks and tests accompanying textbooks. Consequently, familiarity with this items type is available to classroom teacher. Several weaknesses are associated with the completion tests. Perhaps the most serious one is found in the kinds of material which can be used. It is extremely difficult to construct a supply item to measure analysis, synthesis, or evaluation skills which can be answered in only one word or a short phrase. Another problem is that the structuring of completion items so that they have one and only one correct response. Completion items are designed to

require the examinee to supply the correct word or phrase. There are other strengths for constructing completion test. They are easy to construct, simplify the item development task and reduce the amount of time needed for item construction. The completion tests minimize the chance of guessing the correct answer, and when the item is constructed to yield only one correct response. It is simple to make a scoring key. It measures the recall of information rather than recognition. (Rodriguez, 2002)

2.1.2.5 Cloze Test

Cloze tests are similar in appearance to the completion items. Cloze tests shouldn't be confused with simple blank filling tests. In cloze tests the words are deleted systematically. Thus once the actual text has been chosen the construction of cloze test is quite objective. Every word is deleted by teachers is usually between 5th and 10th words. The cloze test which was originally intended to measure language difficulty has been applied to first language testing. The testee should be required to fill each blank in the text itself, or on a separate answer sheet or list. (Milanovic, 1999).

Cloze tests measure students understanding of certain features of language. Several tests specialists argue that cloze tests measure general abilities. It is very useful for assessing language proficiency in a short time and can be used for selection and proficiency purposes. It is important to let students see the first sentence or two without any blanks. This will give them an opportunity to get used to the topic and style of the passage. (McNarma, 2000).

2.1.2 The Achievement Test Characteristics

The achievement test has specific characteristics that teachers have to consider to evaluate the test they use:

2.1.2.1 Reliability

The ability of a test concerns its precision as a measuring instrument. Reliability asks whether an assessment instrument administered to the same respondents a second time would yield the same results. Three different types of factors contribute to the reliability of language assessment instruments:

2.1.2.1.1 Test Factors

Test factors include the extent of the objectives, the degree of ambiguity of the items, the clarity and explicitness of the instructions, the quality of the lay out, the familiarity that the respondents have with the format, the length of the total test with longer tests being more reliable (Hughes, 1989) (Bachman, 1990).

2.1.2.1.2 Situational Factors

Along with test factors, teachers need to be mindful of situational factors such as the manner in which the examiner presents the instruction, the characteristics of the room (e.g. comfort, lighting), and outside noises. These factors may contribute to the lack of consistency of responses from the test takers.

2.1.2.1.3 Individual Factors

These include the physical health and psychological state of the respondent, the mechanical skill, I.Q. ability to use English, and experience with similar tests. (Hughes, 1990). (Bachman, 1990).

2.1.2.1.1 Methods for Estimating Reliability

There are various methods for estimating the reliability of tests which are the following:

2.1.2.1.1.1 Parallel Form Reliability

Parallel form reliability is also known as alternate form or equivalent form reliability. It measures the equivalence of items sampled from the same domain and represents the correlation coefficient between the scores obtained on two forms of the same test. It can be obtained by having two forms of the same test. This kind of reliability requires the following:

1. Develop or adopt a test with at least two forms that use the same sampling plan from the same domain and unique items.
2. Give each test to a group of students in successive testing situations.
3. Correlate the results of the test.

2.1.2.1.1.2 Test-Retest Reliability

This reliability is the correlation between scores on the same test administered twice, separated by a brief period of time. This measure reflects the stability of individual scores between testing and retesting using the same questions or items. This form of reliability is particularly sensitive to variation in student's responses and test administration doesn't change from one administration to another. The test-retest approach used to determine the variation of test content. There are procedures for developing test-retest reliability:

1. Develop or adopt a test with a broad range of items.
2. Test a group of students the first time.
3. Give them the same test within approximately two weeks.
4. Correlate the results of the two tests.

Tindal & Marston (1990) proposed three indices to establish the reliability of criterion-referenced tests. First, we can calculate the number and the percentage of individual items that have been answered the same upon two administrations of the same test separated by at least a week. Second, we can calculate the percentage of the students scoring at various percentages of discrepancy between the two administrations. Third, we can calculate the difference between observed and chance portions of agreement in mastery decisions for each of two administrations of the same test. Tindal & Marston (1990) reported on these reliability estimates for three criterion-referenced tests commonly used in public schools. They found that traditional estimates of reliability and the corrected proportions of examinees

scoring the same across two administrations of the same test were in agreement and generally quite high. An interesting finding was that these tests were not reliable for making mastery decisions, with a range of from 15% to 33% of the total test scores. (Tindal & Marston, 1990)

2.1.2.1.1.3 Parallel Form and Test-Retest

Another form of reliability is based on parallel form and test-retest reliability, which provides the most stringent assessment of consistency. The procedure involves the administration of two unique forms of tests at two times. As with parallel form reliability, two tests are administered, as with test-retest reliability, these two administrations are separated by one or two weeks. This form of reliability is rarely used, in part because it is so stringent, and adequate levels are difficult to achieve. Another reason for its infrequent use probably related to the type of application for which it is most appropriate: measuring is growth on a large domain over a long period of time. We perform the following procedures to determine this type of reliability:

1. Develop or adopt two forms of the same test. Each form should have unique items that have been sampled from the same domain.
2. Test students in the following manner: give test one to half of students and test two to the other half.
3. Wait for two weeks and switch the tests given to each group.
4. Correlate the results.

2.1.2.1.1.4 Internal Consistency Reliability

Internal Consistency Reliability also referred to as odd-even and split-half reliability is based on an analysis of items that make up a test. This form is useful if we want our test items related. Basically, internal consistency reliability measures the degree of interrelationship between items on the same test and is based upon the average correlation among items within a test. Internal consistency reliability is important to establish when defining domains, either because of the diverse nature of the items that the domain includes or because we want to make inferences about students generalized performance in a domain or skill area. This form of reliability is probably the most common type in published achievement tests because of three related issues:

- a. Most of these achievement tests sample items from very broad domains.
- b. Many of these tests establish analysis that includes only those items representing middle levels of difficulty.
- c. Documenting this type of reliability is the cheapest and the most convenient option.

There are procedures for determining internal consistency reliability follow:

1. Develop or adopt one form of a test that contains a substantial number of items. Since the test will be split in half, each half must have an adequate number of items to provide an estimate of reliability.
2. Administer the entire test to a group of students.
3. Divide the test into two halves using odd-even items or splitting the test into a first and second half.
4. Correlate the two sets of scores.

2.1.2.1.5 Inter and Intrajudge Agreement

Interjudge or intrajudge agreement denotes the agreement among judges or within judges over time. This form of reliability is sensitive only to variation in scoring and isn't likely to pick up any of the other sources of variation noted earlier. Interjudge or intrajudge reliability is important when subjective factors affect test scoring. It is also important for teacher-made testing when student performance is hand-scored and the results tallied individually. This form of reliability has little bearing on most published achievement testing, since all student protocols are machine scored. The procedures for calculating interjudge or intrajudge agreement are simple:

1. Test a group of students.
2. Independently score the protocols using either two sets of judges or one judge at two times.
3. Correlate the results. (Bachman, 1990)

2.1.2.2 Validity

Validity refers to whether the assessment instrument actually measures what it is constructed to measure. The assessment of test validity is conducted in different ways:

2.1.2.2.1 Content Validity

It is determined by checking the adequacy with which the test samples the content or objectives of the course or area being assessed. (Hughes, 1990).

Also, it is the degree to which a test has an explicit domain represented by items in the test and specific procedures describing how they were selected. (Tindal & Marston, 1990).

Content validity is very important. First, the greater the test has content validity, the more likely it is to be an accurate measure of what it is supposed to measure. Secondly, such a test is likely to have a harmful backwash effect. Areas which aren't tested are likely to become areas ignored in teaching and learning. (Hughes, 1990).

2.1.2.2.2 Criterion–Related Validity

Calls for determining how closely respondent's performance on specific sets of objectives on a given assessment instrument parallels their performance on another instrument, or criterion, which is thought to measure the same or similar activities. (Hughes, 1990). (Bachman,1990)

Criterion-related validity is established by comparing performance on an accepted standard or criterion. If the testers are interested in how well the performance is, we need to examine its criterion predictive validity. In other words knowing how well a student performs on the curriculum based measures, I can predict how well she or he will perform now or in the future on many published achievement tests. (Bachman, 1990)

2.12.2.3 Construct Validity

Construct validity refers to the degree to which scores on an assessment permits inferences about underlying traits. In other words, it examines whether the instrument is a true reflection of the theory of the trait being measured: language, in this case.

2.12.2.4 Systemic Validity

It is one that induces in the education system curriculum and instructional changes that foster the development of the cognitive skills that the test is deigned to measure.

2.12.2.5 Face Validity

It refers to whether the test looks as if it is measuring what it is supposed to measure. (Hughes, 1990). There are factors affect the validity of tests:

1. Purpose of the test: Is it for grades? For assessing program effectiveness only? For planning individual activities? Or something else?
2. Weight of test: This is important if it is to contribute to students final grades.
3. Time: This is available or required for taking the test. Number of the parts or blocks
4. The procedures: for recording their answers or responses, for example is it on a sheet paper, is it on a tape recorder. (Genesee& Upshur, 1998). (Bachman,1990)

2.1.2.3 Test Objectives

If the teacher wants to have an adequate test before the test can be constructed, the objectives must be clearly in mind and should be stated in writing in behavioral terms. This statement must be followed by a discussion of ultimate and instructional objectives, and suggested systems for their classification. Since the primary purpose

of classroom testing is to obtain individual measures for evaluating students with regard to their acquisition of the instructional objectives, a blueprint for selecting appropriate test items must be developed. This blueprint is called a Table of Specifications. Then the teacher has to choose an applicable blueprint to the objective and to identify the complexity of the intellectual or affective activity involved.

Baker (2002) recently noted" the operational limits of the target domain of learning" is a necessary condition for using evaluation effectively for both accountability and for school improvement. But if evaluation is to direct reform, the achievement targets that constitute the domain of each of these tests must be:

(a) be a legitimate domain of achievement targets.

(b) be sufficiently described to be communicated effectively to others, especially instructional personnel.

(c) be reliably sampled by the test (i.e., not only does the test sample the domain well, but also, teachers believe it will sample the domain well).

These principles apply to any test originating from outside the classroom and intended to effect change in the classroom.

Educators work more effectively if they believe their goals are worthy. In education, that means the value of the targets of instruction is apparent. While teachers can and often do make judgments according to our own beliefs about any set of curricular goals (i.e., learning targets), harnessing the efforts of schools, districts, and an entire state requires a shared belief in the worth of the goal. In our democratic society, that requires a process, usually political, in order to attain a consensus. For example, in developing goals in some content area, a process that effectively includes representation of teachers will be better accepted by educators than one that does not. The state school board, representing a broad constituency of stakeholders, can be an accepted authority to approve both the process and the product. (Baker, 2002)

As Popham (2006) has noted, curricular goals are often too broad to be covered completely by teachers. There typically are too many curricular goals either to cover comprehensively or to be represented completely in an assessment program. Popham concluded that curricular goals must be reduced, both in order to focus teaching on a manageable set of learning targets that are of most importance, and in order to allow valid evaluation of student achievement across that domain.

Popham (2006) suggested three approaches to reduce the domain of instructional targets: new standards, coalesced standards, or derivative evaluation frameworks. New content standards would be time-consuming to develop. By coalesced standards, Popham described a hierarchical structure in which current targets are subsumed under fewer, but nevertheless measurable goals, but that, too,

would be time-consuming and perhaps difficult to achieve. Given these drawbacks, Popham concluded that the third option, derivative evaluation frameworks, is most likely to succeed.

A test is supposed to assess what students are to know and are able to do with what they know. Actually, though, every achievement test item prompts both these elements because it requires a student to do something with something. Classroom assessment textbooks such as Nitko (2001) almost universally recommend a table of specifications as a device for describing test items in terms of the content and process dimensions. That is, what a student is expected to know and what he or she is expected to do with that knowledge are described by combinations of content (e.g., rows) and process (e.g., columns) in a table of specifications. Even a highly motivated educator cannot attain a goal that is unclear. Some way is needed to clarify the domain of each test so it can communicate unambiguous targets in combinations of both content and process dimensions. Everyone agrees that the domain of any assessment should be sampled representatively on each test form. However, teachers who have worked with mandated assessments often do not feel the test covers what they have been teaching, even when they honestly believe they have represented their district's curriculum instructionally. Perhaps they are often right. The connection between the tested domain and the educators' learning targets needs to be established at the start of the appropriate instructional sequence. Not only must the educator understand the domain, but he or she must also believe the test will sample it appropriately. Otherwise, the test will be marginalized as irrelevant and teacher motivation expected as a result of the assessment and accountability program could be lost.

In summary, a testing program is effective as a guide to instructional goals to the extent that it covers a publicly accepted learning domain that is described in terms of both content and cognition. (Nitko, 2001).

Any test item should be related to Bloom's Taxonomy of educational objectives. The taxonomy had seven levels:

1. Memory: which requires the students to recognize or recall information.
2. Translation: translation thinking is quiet literal and doesn't require students to discover intricate relationships implications or suitable meaning. The student changes information into different symbolic form of language.
3. Interpretation: The student relates facts generalization, definitions, values, and skills.
4. Application : the student solves a problem that requires the identification and skills .There are three main characteristics of the questions in the application categories:
 - a. These questions deal with knowledge.
 - b. They deal with the whole of ideas and skills rather than solely with parts.
 - c. Application questions include a minimum of directions or instructions.

5. Analysis: it requires solutions of problems in the light of conscious knowledge of the parts and process of reasoning.
6. Synthesis: synthesis questions encourage students to engage imaginative original thinking. The student solves a problem that requires original creative thinking.
7. Evaluation: evaluation questions are easy to compose and are frequently used in class discussion .Skill in evaluation requires knowledge of the nature of values. (Bloom, 1956).

2.1.2.4 Test Length

The test must be of sufficient length to yield reliable scores. Usually, the longer the test, the more reliable the results. If the Table of Specifications is carefully followed and the item pool is adequately sampled, the test should be valid if it is reliable. Consequently, all that is now required is to construct a test of sufficient length. (Tindal&Marston, 1990).

2.1.2.5 Test layout

The arrangement of the test items within the examination influences the speed and accuracy of the examinee. The best layout is one which utilizes the space available while retaining readability. In most cases it is wise to avoid a layout which results in one-line questions spanning an eight-inch-wide page. Some educators recommended that as a means reducing test anxiety, test items be arranged in order from the easiest to the most difficult items. (Tindal&Marston, 1990).

2.1.2.6 Test Instructions

Any test should contain instructions to let students what they are supposed to do. We want to discover what students have learned, now how well they can understand test instructions. General instructions for test provided students with general information that will orient them to the task ahead.

If the examinees don't clearly understand the question format the test may be measuring only their understanding of item types rather than their acquisition of the instructional objectives. Although the instructions may be oral, a combination of written and oral is probably desirable except with young students.

The instructions should be clear, concise, and explicit. After the examinees have read the instructions they should be encouraged to ask questions. In most cases it is advantageous to use the instructions as a cover page since this permits the directions to be separated from the body of the test. (Tindal & Marston, 1990).

2.2 Second: Previous Studies

The review of related literature has a great number of studies that have been conducted concerning this topic. The review of the studies is divided into two parts. The first deals with the foreign studies. The second deals with the Arabic studies.

2.2.1 Foreign Studies

Coulson & Silberman (1960) found that there are no significant differences between teachers trained on multiple choice programmed material and those using constructed response mode.

A study reported by Scannell & Marshall, (1966) were analyzed an essay exam which was used by twelve American history teachers, these 12 forms of essay questions were analyzed, and the researchers found that; the forms of these tests were constructed in a way which contained various numbers of spelling, grammar and punctuation errors. As a result of this study the twelve high school English teachers participated in training courses to avoid these errors in their exams. (Cited in Marshall & Hales,1971)

Marshall (1967) in his study analyzed thirteen forms of the essay exams, finding out that these forms containing many errors. He found 18 spelling errors, and 18 grammatical errors. He also found out that the exam is fairly neatly written, because it is handwritten essay exam. In the same time the exam has clear instructions. So the results of this study indicated that teachers have no experience in writing the exams appropriately even when they are giving explicit instructions to students. (Cited in Marshall & Hales)

Marshall & Hales (1971), analyzed a sample of essay examination questions used by classroom teachers. He found that these questions fell into three categories: simple-recall questions, short-answer questions, and discussion questions. He reported that across the first twelve grade levels 35% of the questions were simple-recall questions, 35% were short-answer questions, and 30% were discussion questions. Elementary and secondary school examinations were different in that. 42%-18% of these three types were found in the elementary school examinations and 25%, 29%, and 46%, respectively, of these three types were found in the secondary school examination.

Victor (1972) applied an achievement test administered to students in grades nine, ten, eleven, and twelve in fourteen schools in nine different states in USA with over 22000 students participating in forty five minutes experimental test. The test covered composition, literature, math, science and reading. The primary goal of the study was to study the techniques and the formats of the test. They found out that 93 of the items were easy for grade twelve, but they were difficult for students in all three grades. Also, they discovered that large numbers of students have incorrect responses in the multiple choice items. Because distractors were written to be unclear to students, they haven't developed in an appropriate level of understanding and have only partial or inaccurate information about the concepts. Distractors should generally become less attractive and clearer to students.

In his study Newman (1981) aimed to identify the cognitive levels of the teacher's tests due to educational qualification, and experience. He took (294) teachers as a sample of all teachers in different classes and in different states in America. He used

two instruments for the in-service teachers and the second an evaluation list to evaluate teacher-made tests. The findings showed that teachers have good knowledge in constructing their tests according to the main norms and this knowledge could be increased if there is an increase in teacher's knowledge and qualification. Furthermore, he found that teachers have problems in constructing multiple choice questions especially using poor distractors and specific determiners such as (always, may, none, never, all,...) that lead to the correct answer.

Madsen (1982) mentioned in his study, that unclear or inaccurate instructions and inadequate time allocation as source of test anxiety, all have an influence on the test performance.

Fleming and Chambers (1983) conducted a study investigating the cognitive abilities that teacher-made tests measured. The results indicated that teachers construct tests that test knowledge and facts. Also, he found out that 80% of the questions are facts, idioms, rules and principles, 94% of the questions are knowledge used in evaluating the elementary levels and 69% of these questions are knowledge used in evaluating the secondary levels.

William and others (1984) in their study aimed to identify the formats that test must contain (multiple choice, short questions, T/F....) and the questions in the achievement tests that were made by different teachers in different faculties in different universities. The sample of the study contains (1220) tests. He used an instrument to evaluate the test using ten categories .He found out that 44% of these questions were essay questions and multiple choice questions that contain the keying words which lead directly to the correct answers by having the negative forms.(Cited in Al-janzrah,1999)

Kirby & Oescher (1987) conducted a study to determine characteristics of teacher-composed classroom tests, with emphasis placed on describing the levels of knowledge addressed by the test items. In this preliminary investigation, 19 mathematics and 16 science teachers working in 4 high schools in a mixed suburban/rural school district were asked to:

- (1) Complete a brief instrument describing the format, objectives, analysis, and uses of their tests as well as their level of confidence in their testing skills.
- (2) Supply the researchers with their most recently administered unit or quarter examination.

A rating form was devised to analyze a sample of teacher-composed tests. Interrater agreement for a sample of the tests ranged from 90 to 100 percent. Teacher's perceptions of the levels of knowledge addressed by their test items were compared to the researcher's actual ratings by means of t-tests or mean differences with the alpha levels. Results indicated that there are major weaknesses discovered in constructing and the objectives of the tests aimed to test low cognitive levels. Moreover, there were flaws in construction of individual test items, and inadequate instructions.

Another study (Marso & Pigge, 1988) analyzed a group of tests taking into consideration the methodology of applying the tests and the cognitive skills that required to design them. Also, the study aimed to identify the mistakes that teachers made in designing tests. This study analyzed the teacher's reports and (175) tests those teachers designed. The study aimed to analyze the nature and the type of the test that teacher had in the governmental schools. The researcher found out that teachers used multiple choice, matching and short questions only and rarely used the essay questions.

Kirby, (1988) in his study aimed to develop an objective instrument to assess a teacher's perceived engagement in reflective practice. A Reflective Teaching Instrument (RTI) was developed around three dimensions of reflective practice in teaching:

- (1) Diagnosis (problem setting).
- (2) Testing
- (3) Personal causation.

Indicators of each of the dimensions were compiled from a review of the literature to generate an instrument. Four educators comprised an expert panel that assessed the face validity of each item. The pilot instrument of 48 Likert-format items was administered to 40 practicing teachers enrolled in graduate classes. A field test was subsequently conducted with 102 public elementary and junior high school teachers, representing a response rate of 94%. At the end the researcher has an objective instrument to assess a teacher's perceived engagement in reflective practice.

Gentry, (1989) presented a paper at the annual meeting of the mid-south educational research association in USA suggesting the necessity for comprehending and knowing the objectives and the table of specification because they play an important role in developing the tests. Also, he gave instructions for constructing tests in its different formats.

Nair-Venugopal, (1991) examined the oral communication courses for English majors at the National University of Malaysia including tests designed by faculty and coordinated with the curriculum. This study aims to discover the ideas that a teacher who has been actively involved in curriculum design is in a good position to design a test for that curriculum, and that teacher-made tests have a beneficial backwash effect on student learning. The course features have two levels of instruction, each taught over two consecutive semesters. Final tests for both have levels sample global communicative ability. Because the approach is communicative, the examinations are series of tests administered throughout the semester, allowing for continuous feedback to aid instruction. At level 1, the tests focus on three speaking tasks: extended, impromptu speech; group discussion; and an end-of-semester project. The tasks test three modes of speech: talking about oneself, others, experiences; narrating and describing events; and expressing and justifying opinions. At level 2, tests focus on group discussion, public speaking, debating, and an end-of-semester project. Rating scales have been constructed for all tests based on the types of communicative ability required. The researcher

discovered that continuous testing has reduced test anxiety, developed student's oral communication. Teachers also aren't in good position to design tests due to the lack of experience, so the researcher recommended that teachers should have practice sessions in teacher-made tests.

Talmir, (1991) conducted a study investigating how multiple-choice items can be designed and used as an effective diagnostic tool by avoiding their pitfalls and by taking advantage of their potential benefits. He found out that teachers have a problem in designing them; they use the positive or negative words which led to guessing the answer.

In other experiment Beaton (1992) found out that neither teacher-made tests nor the Scholastic Aptitude Test are appropriate for measuring the educational goals.

Boothroyd (1992) carried out a study. This study assesses teacher's measurement training and the extent to which their measurement knowledge is adequate to develop quality classroom tests. Forty-one 7th- and 8th-grade science and mathematics teachers were assessed using 65-item multiple-choice tests. Participants were asked to identify violations of item writing principles in 32 multiple-choice and completion items. Three questions were addressed:

- (1) What was the nature and extent of measurement training?
- (2) What measurement knowledge and skills did these teachers possess?
- (3) What teacher characteristics are related to their measurement knowledge?

Results indicated that teachers' knowledge of measurement was insufficient, probably at least partially due to inadequate training; and that teachers frequently tested students with their own tests and placed more weight on students' scores on these tests when assigning end-of-course grades than on other forms of assessment.

Hynie, (1992) in his study examined the effectiveness of test's questions that were used by teachers, and the effects of experience, qualification, and their training sessions in testing. The researcher analyzed (993) questions that were constructed and developed by (15) teachers who were chosen by supervisors to carry out this job. These questions were investigated according to nine criteria:

1. The spelling mistakes.
2. The punctuation mistakes.
3. The use of key words.
4. The usability of tests.
5. The reliability of tests.
6. The clarity of tests.
7. The cognitive levels.
8. The effectiveness of questions.
9. The distractors.

The researcher investigated the effects of the demographic variables, which were qualification, experience and training programs that they had in constructing tests.

The study showed that teachers who had less than eight years of experience commit spelling and punctuation mistakes very often. Also, the clarity of tests was absent. Moreover, teachers who had BA and higher degrees along side with participation in the training programs were the best in constructing their tests.

Dereshiwsky (1993) aimed in his study to present a procedure for developing and refining teacher-made surveys and tests, which would be valid and reliable for meeting local needs. First, a brief rationale is given for teachers producing their own instrumentation. Next, an easy-to-apply process for developing and pilot-testing one's surveys and tests is presented, a process that requires no computers or statistics, but rather depends on open sharing, discussion, and communication with colleagues. To illustrate these procedures, an actual example of a survey used to evaluate the 1992 Arizona Leadership Academy is provided. The researcher discovered that classroom teachers have the best possible vantage point for constructing locally appropriate surveys and tests.

Pigge & Marso (1993) investigated in their study the teacher's attitudes towards teacher-made tests. They present a summary of findings from a review of approximately 225 studies of K-12 classroom teachers' attitudes toward teachers-made tests, other educator's evaluation and support of teacher-made tests and testing practices. The findings indicated that classroom teachers have a positive evaluation toward teacher-made tests and regard these tests as having a far more positive impact upon their day-to-day instruction than do other types of tests. Further, teacher's positive evaluation regard for these tests is reflected in their heavy reliance upon and frequent use of these self-constructed tests in their classrooms.

Childs, (1998) in his study examined the steps of test construction and presented suggestions for interpreting the outcomes of the achievement tests. The first steps involved identifying what the students should have learned and designing the test. The learning objectives emphasized determine the material to include and the form the test will take. Once the objectives have been designed, the second step is writing the questions. General principles of test construction are reviewed. Guidelines for construction of multiple-choice tests, probably the most difficult to construct, are also given. He discovered that a carefully constructed achievement test can help the educators teach more effectively and the student masters more of the objectives because tests didn't measure the content objectives.

Another study conducted by Daniel and King, (1998) aimed to identify the elementary and secondary classes teachers knowledge in constructing tests, and their educational knowledge in measurement and testing. The population of the study consisted of (95) male and female teachers. The study indicated that teachers don't have enough knowledge in the concepts and principals of testing, although they used their little knowledge of testing in constructing their tests. Moreover, there was no significant difference between male and female teachers who teach elementary and secondary classes in constructing their tests.

Mills (1998) investigated the tests developed by elementary foreign language teachers of French, Japanese, and Spanish in a school district in South Carolina. The tests were designed to determine the level of end-of-year student learning and to provide a basis for evaluating the curriculum of each of the three languages. The French and Spanish tests contained tests of listening and comprehension, vocabulary, and reading, and the Japanese test contained tests of listening, complex listening skills, and vocabulary. The tests were analyzed in terms of item difficulty, high-low discrimination indices, and distributions patterns. The subtests were also analyzed, highlighting the tendency of teacher-made tests toward the measurement of skills. The study provides descriptive statistics for all parts of the tests and the total test results. Analysis indicates that, in general, all three tests had too low a level of difficulty, with few questions to challenge the more able students and developing listening and reading skills avoiding the writing and speaking skills. These results are a contribution toward the improved design of foreign language tests for elementary school students, for whom foreign language study is still relatively rare.

Three research studies conducted at Thai universities involving language testing and courses in General English and English for Academic Purposes (EAP) are discussed. The first study examined the predictive validity of different types of language tests on academic achievement in General English and EAP courses. It was found out that the test format, as shown in the matching and cloze test has a significant role in predicting future academic achievement, and the content of language tests may play a role in academic achievement for each type of language program. The second study showed the direct and indirect relationships between subskills of General English and EAP tests. It was found out that all language subskills, regardless of content, are significantly related. The third study examined the underlying relationships between General English and EAP tests. It was found that EAP tests may predict achievement in EAP programs better than General English tests; the formats associated with each discipline tend to predict academic success in science better than those that are not related to a specific discipline; and there is a common factor shared by the EAP tests, General English tests, and knowledge of the subject matter represented by student grade point average. (American Association of Colleges for Teacher Education, 2000).

Kane, (2000) carried out a research to find out the current status of teacher testing practices and materials in the public schools of the District of Columbia. This study found out a high degree of readiness within the district for the use of subject matter examinations as a criterion for teacher certification. Issues were examined by a policy analysis tool known as the convening process. The researcher recommended for a teacher testing policy:

- (1) Subject matter knowledge testing for teacher certification or licensure.
- (2) Requirement of a specified certification score.
- (3) Certification testing for all teachers regardless of other certification.
- (4) Limitations on temporary certification.
- (5) Analysis of processes used in hiring, promotion, and tenure decisions.
- (6) Review of tests and development of new tests.

(7) Writing test for diagnostic and prescriptive purposes for all new employees.

Specific recommendations were made to improve the current validity procedures and test development process, including the establishment of a research and measurement unit.

The committee used its evaluation framework to evaluate a sample of five widely used tests produced by the Educational Testing Service. The tests the committee reviewed met most of its criteria for technical quality, although there were some areas for improvement. The committee also attempted to review a sample of National Evaluation Systems tests. The findings showed that on all of the tests that the committee reviewed, minority candidates had lower passing rates than nonminority candidates on their initial testing attempts. The committee concludes its evaluation of current tests by reiterating the following: The profession's standards for educational testing say that information sufficient to evaluate the appropriateness and technical adequacy of tests should be made available to potential test users and other interested parties. (American Association of Colleges for Teacher Education, 2000).

Karabenick, (2000) designed a survey to take approximately 15 minutes to complete, and administered it to 1,656 elementary school teachers in Michigan to obtain information on a variety of topics related to student assessment and mandated state testing. Most of the teachers were employed in small suburban or urban schools, and 88% were employed in the public schools. Teachers apparently placed very little value on the mandated tests as a way to evaluate a student's progress, and only 36% said that they used the state tests for this purpose. The assessment measures that teachers found valuable were those that provided timely and useful information about individual children. Most teachers recognized a role for mandated tests as diagnostic tools, but most did not agree that such tests should be used to test students or school accountability purposes.

Kopriva,(2001) work with English language learners examined assessments using think aloud methods. She recommended that all test designers use think aloud methods to better understand test design and its effects on student test-taking processes. According to Kopriva, verbalizations used for think aloud data provide valuable insights into the following:

- Student understanding of constructs.
- Student skill level.
- Relevance of items to student life experience.
- Relevance of items to content taught.

Ediger, (2001) conducted a study aimed to evaluate a science test. He found out that teacher-developed tests can be more valid and reliable than standardized tests in evaluating student achievement in science. Many teachers, however, are not acquainted with the norms to use in writing tests and to design good tests. Also, science portfolio is a good way to evaluate the everyday science achievement

students using Multiple Intelligences Theory. When teachers want to use written tests, there are criteria that should be applied to the construction of test items. Teacher's observation is another important aspect of assessment in science. It is also important to consider metacognition skills when evaluating student achievement in science.

(Haladyna, 2002) stated in her study that teachers can also minimize construct-irrelevant variance by adhering to effective design strategies. Such design features may increase the content validity of information that can be gleaned from test data.

A study conducted by The National Research Council, (2002) to determine, and note similarities and differences in the cognitive objectives of examinations used in ninth grade courses in a junior high school. These examinations are prepared by individual teachers and teachers as members of committees. The researcher analyzed the test items according to Bloom's taxonomy. The Item frequencies were tabulated and percentages were calculated. The courses covered were Civics, History, Math, Biology, science, French, English, Economics, and Business. The researcher found out that half of the questions only assess memory level. There was lack of concern for the areas of analysis, synthesis, and evaluation.

Thompson, (2002) conducted a study focusing on the Think Aloud Method (Cognitive Laboratory) research methodology to detect design issues in large-scale tests, based on a framework of universal design. They described the methodology in general and evaluated its effectiveness for finding design issues in tests for students with disabilities, English language learners, and English proficient students without disabilities. Finally, they discussed limitations and future directions for this methodology, particularly for students with disabilities with whom this methodology has not been used extensively before. They found issues related to unclearly defined constructs, inaccessibility of items, unclear instructions, incomprehensible language, and illegible text and graphics. To this end, think aloud methods appear to be a useful strategy in the design and refinement of large-scale assessments. Think aloud methods appear to be an effective way to determine the effects of item design for a wide variety of students. According to Thompson's study (2002) there are some elements that should be included as universally designed assessments:

- (1) Inclusive test population.
- (2) Precisely defined constructs.
- (3) Accessible, non-biased items.
- (4) Simple, clear, and intuitive instructions and procedures.
- (5) Maximum readability.
- (6) Comprehensibility of content.

(7) Maximum eligibility.

Johnstone, (2003) indicated that teachers can create assessments that are more accessible to diverse students by designing items using elements of universal design. He studied the content of the tests specifically vocabulary tests. The average of the content in the tests was moderate 35%. Johnston was able to divide his vocabulary test in vocabulary and general vocabulary. The researcher found that vocabulary was less related to the content of the passage that students studied. The correlation of specific vocabulary with performance on the reading tests was 39%.

Lerkanen, (2004) conducted a study that aimed to investigate prospective relationships between reading and writing performance during the first grade of primary school. The data was collected from 83 Finnish-speaking children who were examined four times on reading, spelling, and productive writing skills during the first grade. The results showed that all the testing tools concentrate on testing the reading skill and avoid the writing skill. So the researcher recommended that it may be important to emphasize the compositional writing which may lead to the development of reading skill.

O'Neil, (2004) asked expert science teachers to evaluate the content and cognitive characteristics of the science test items for the 10th grade science. The results indicated the content area representation was fairly consistent across years and the proportion of items measuring the different cognitive skill areas was also consistent. However, the experts identified important cognitive distinctions among the test items that were not captured in the test specifications.

Posner, (2004) conducted a study that revealed that intense pressure that teachers have causes them to devote virtually all classroom time and resources to prepare students for the standardized test only. This phenomenon is called "teaching to the test. The tests measure success in teaching the curriculum and so "teaching for the test" is "teaching for the curriculum" without paying attention to the content objectives and formats because this test is a recall test. So problems that can appear on a standardized test are, of course, quite limited in form and complexity, as the student is allocated only a minute or two to complete each one. The intellectual processes required to solve a really complicated problem are not essentially the same as those required to solve these simpler problems. Then a student prepared only to solve standardized test problems could lack the mental preparation required to attack really hard problems and all these tests were recall tests.

(South Dakota Department of Education, 2004). The South Dakota Assessment System provides information for teachers at schools on how to use different methods in evaluating their teaching and curriculum as well as allowing parents to monitor their child's progress. The assessment is used to determine individual, school-level, district-level, and statewide achievement in reaching the goal of proficiency of the state's essential core reading and math content standards.

Dolan, (2005), and Johnstone, (2003) have attempted to clarify design issues by demonstrating how specific designs play a main role in the improvements of tests which can affect student performance.

Dolan, (2005) conducted a recent research on test design which stated that test design should be accessible and understandable to a wide variety of students (including students with disabilities and English language learners).

Erkaya, (2005) conducted a study that aimed to familiarize English as a Foreign Language (EFL) instructors with the effectiveness of using literature in language instruction. Literature must be integrated in the curricula because it adds a new dimension to the teaching of EFL. Short stories, for example, help students to learn the four skills: listening, speaking, reading and writing more effectively because of the motivational benefit embedded in the stories. So the researcher found out that short stories can teach literary, cultural, and higher-order thinking benefits. Using short stories in the EFL classes, helped students in learning the four skills (listening, speaking, reading and writing).

Nicosia (2005) aimed to improve the basic skills in reading, writing, speaking and listening because students need to succeed in the business world and for transfer to a senior college. So he evaluated the activities and the assessment tools that teachers use to improve the English language skills which are reading, writing, speaking and listening. The researcher found out that those teachers concentrate only on the writing skill rather than on reading, listening and speaking.

Popham, (2005) found out in his study that for the last four decades, student's scores on standardized tests have increasingly been regarded as the most meaningful evidence for evaluating U.S.A schools tests. Most Americans, indeed, believe students' standardized test performances are the only legitimate indicator of a school's instructional effectiveness due to the follow up evaluation for these tests that the Ministry of Education did.

Abbott, (2006) in his study analyzed 32 questions in a test. He found out that all these questions were testing the reading skill strategies like breaking a word into smaller parts, scanning, paraphrasing, matching skimming, connecting, and inferring.

Johnstone, (2006) stated in their study that think aloud methods, as they designed them, were not effective for students with cognitive disabilities. Students had great difficulty in producing the language needed to explain problem-solving processes. Think aloud methods also did not produce informative data for very difficult mathematics items because students had difficulty verbalizing their thoughts while solving problems. Also, he stated that think aloud method appears to be an effective way to determine the effects of item design for a wide variety of students (with the exception of students with cognitive disabilities) and for items with low to moderate difficulty levels.

Levacic, (2006) presented a paper in England. He encouraged secondary school teachers to be more diverse by becoming specialists in constructing their tests. This paper estimates the relative effectiveness of specialist teachers in testing student's achievement taking into consideration the student's gender, age, curriculum context, and the different cognitive levels. This will lead teachers to have a test that matches the international norms of good achievement test.

Cohen, (2006) conducted a study, to examine the contribution of phonological and nonphonological language skills among students with and without disabilities aged 10-13. The results showed that tests evaluate the reading skill while other skills were avoided.

Kjellin, (2006) aimed in his study to evaluate the classroom activities and testing tools that deals with the four language skills. Results showed that the instruction was concerned more with the practice of basic skills in reading and writing than practice of the language listening and speaking, so all the testing tools were constructed to test reading and writing while avoiding listening and speaking skills.

2.2.2 Arabic Studies

Garadat (1988) conducted a study to identify the science teachers' knowledge who teach the elementary classes in Jordan. He investigated the way they construct their achievement tests, and how they use them. He investigated the effect of experience, qualification and gender of teachers. The researcher applied an instrument containing the norms of publishing the test to (298) teachers. The researcher analyzed these tests that teachers made according to the instrument. The results revealed that the teacher's knowledge according to these norms weren't enough and there was a significant difference to the BA and MA teachers and for those who have short period of experience. So according to these results he recommended that more attention be paid by increasing teachers' participation in training sessions during working in order to develop their knowledge in constructing their tests.

Al-Omar, (1989) conducted a study which aimed to reveal the testing methods and formats that elementary teachers used in evaluating their students in Jordan. It aimed also to evaluate the effectiveness of achievement tests according to the use of cognitive levels and to what extent they match the norms of good tests. The researcher used two instruments: the first consisted of two questionnaires: one for the teachers to investigate the testing methods they use and the other questionnaire for students. The second instrument was a list of norms that the researcher constructed and used to analyze (202) tests prepared by male and female teachers. The findings revealed that the essay tests were the most frequent format and teachers rarely used the (True\ False) questions. Also, the science and math teachers used the multiple choice, matching and cloze questions mostly, whereas the Arabic and English language teachers like to use the essay tests. In addition to that, these tests were full of spelling and printing mistakes. Moreover, the tests lacked instructions, and the teacher's gender and qualification have no effects on constructing tests.

Abu-Taleb, (1991) conducted a study which aimed to identify the testing formats that were used by teachers in the governmental, private and the UNRWA schools in Jordan. The researcher used two instruments that deal with the testing formats, a questionnaire and a list containing the categories for good achievement tests and they were applied on (44) tests including (320) questions. The sample consisted of (215) female and male teachers. The researcher found out that teachers used the test which required students to explain their answers, discussions, projects and observation. Also, they used the short questions and completion only.

Al-Aga, (1994) conducted a study in Gaza. The researcher aimed to construct a list of norms to help teachers in constructing their tests. The researcher analyzed the ninth grade tests in science according to the instrument he prepared. The findings of the study showed that there were many norms to construct tests, the researcher had chosen the most important and practical norms and they were (36), 10 for the instructions, 10 for the essay and objective questions. The researcher found out that 95% of the test's questions were knowledge and facts and 5% of them were about the understanding and application levels.

In 1995 the Ministry of Education in Jordan held a number of workshops. Supervisors in evaluation and measurement and supervisors from Britain participated in these workshops in the Directorate of Testing. These workshops aimed to develop Tawjihi test according to the measurements methods and to what extent the test reflects the objectives of curriculum. Also, to what extent Tawjihi test measures high cognitive skills. Supervisors analyzed Tawjihi curricula according to the content, skills, objectives, and building the table of specification. Then supervisors in 1996 held other workshops for teachers to practice designing tests. Then the ministry applied experimental tests randomly. The results of these tests were analyzed and a questionnaire was given to teachers and supervisors to give feedback. As a result of these workshops Tawjihi test has now its new face validity, measure different and high cognitive skills, and cover the curriculum objectives.

Another study conducted in (1995) by the supervisors committee at the Ministry of Education in Jordan to measure, to what extent do the questions of tests match the behavioral objectives. The researchers used an instrument that aimed to investigate the criteria of good tests, to check teacher's knowledge in designing tests, and to know the different attitudes between teachers in different classes in designing tests. Researchers found out that teachers haven't got enough experience to design good tests according to objective domains.

The Administration for Education Research at the Ministry of Education in Jordan (1995) analyzed the fifth class tests in all subjects. The study aimed to reveal the levels of cognitive objectives that the tests measure and to what extent tests are aware of individual differences and the norms of construction. The study chose randomly 25% of the fifth class tests. The researchers revealed that:

1. The norms of construction which were used in designing the test were 90%
2. Unclear handwriting tests, crowded tests, and poor readability.
3. The essay tests were poorly constructed.
4. Tests reflect 87% of curriculum objectives.
5. Tests items measured facts and knowledge objectives, without paying attention to the individual differences between students.

Al-Omarie, (1997) conducted a study which aimed to evaluate the teacher made-tests in the governmental schools in Jordan. Also, it aimed to develop the constructional norms in designing tests. The researcher analyzed 200 tests using a list of norms which were chosen randomly. The researcher found out that teachers have poor knowledge of designing tests; problems face them in constructing tests. Also, there was lack of general instructions and the clarity of the evaluated tests.

Al-Ekbaty, (1998) conducted a study which aimed to find out the evaluation methods that teachers (female and male) and supervisors used and the effectiveness of these tests in Yemen. The population of the study consisted of all teachers; (female (82), male (120)) and (6) supervisors in AL-HededeH District, also (44) achievement tests in the academic semester 1996-1997. The researcher used two instruments: a questionnaire to measure the methods used by the teachers and supervisors and a list of norms that should be found in good achievement tests to analyze them. The study results showed that, the essay tests, cloze tests and oral tests were the most frequent formats used by teachers. The evaluation methods that were used were the individuals and groups projects. Also, the tests have no general instructions; the percentage of tests that have general instructions was 36.4%.

Al-Janazrah's study (1999) aimed to evaluate teacher-made tests with respect to specific norms of constructing and publishing good achievement tests, to explore the influence of demographic variables (teacher's gender, experience, and educational qualifications) on the quality of these tests, and to explore the currency of the different types of test items. The population of the study consisted of all achievement tests (the final test in chemistry) prepared by teachers in Hebron and Bethlehem provinces. The researcher used an instrument containing a list of norms for the construction and publication of good achievement test. The results of the study showed that most of tests lack test general instructions although instructions contain marks for each question. Moreover, most of the essay type questions measure the lower cognitive levels. Also, test items covered less than half of the curricula content, and half of the test papers contained either spelling mistakes or punctuation errors. The teachers also, have a problem mostly in writing distractors and the presence of key words. In addition, the findings of the study showed that the tests written by less experienced teachers had better overall quality, where as more experienced teachers are better. Male teachers holding educational qualifications were better in the dimensions of writing test constructions and the test publication dimension; whereas female teachers who don't hold educational qualifications were better in writing test items. The results indicated that essay questions were the most frequently used in these tests while matching and cloze questions weren't used at all. The researcher recommended that more attention

should be paid to assessment and evaluation courses in pre and in- service teacher training programs. He also recommended teachers to follow up the several norms in constructing their test and to conduct other studies on different populations and different subjects area according to their different demographic levels.

Achievement tests are frequently the major basis for evaluating students' progress in schools which can be constructed in different formats like essay questions, short answer questions, matching questions, true false questions, completion questions, and multiple choice questions. Reliability and validity are main characteristics that teachers have to consider in constructing their tests. The literature has great number of studies that have been conducted concerning the thesis topic, Arabic and foreign .Some of these studies agreed with the study results and others disagreed. Methodology is very important part in any study, so the next chapter is dealing with methodology and the procedures of the study.

Chapter 3

Methodology

This chapter includes a description of the population and methodology of the study. It also includes a description of the process of preparing the study instrument and means needed to ensure its validity and reliability. The variables of the study, the procedures of application and the statistical analysis, are also described and explained in this chapter. The researcher used the descriptive method because she took all the members of population.

This study aims to evaluate the ESL Tawjihi tests based on norms of the construction and publication of good achievement tests, in Palestine. These tests were prepared by the Ministry of Education in Palestine to assess the ESL students at the end of academic year as a level test in order to allow students to continue their higher education. Tawjihi Teachers analyzed these tests using an instrument prepared by the researcher depending on the studies and literature review.

3.1 Population of the study

3.1.1 Teachers

The population of the study included all English Tawjihi teachers who were (50) teachers; (26) female and (24) male teachers in (40) schools at the governmental schools in South Hebron district in the first Semester of the academic year 2006-2007.

Table 3.1 The number of teachers in the Directorate of Education /South Hebron

Female	Male
26	24
50	

3.1.2 Tests

Also, the population of the study consisted of all English Tawjihi achievement tests(literary stream) prepared by the Ministry of Education in Palestine from (2000- 2006).

3.2 Instrument

A questionnaire was constructed as a major tool for obtaining the needed information for this study. The researcher reviewed the previous literature to find out a suitable instrument to use in the study, but the researcher discovered that the literature lacked an instrument for such a study. So the researcher has to construct one by herself, depending on the previous studies and literature that deal with

testing like (Gronlund & Linn, 1990), (Tindal & Marston, 1990), (Cross,1990), (Al-Ekbaty, 1998), (Al-Janazrah,1999) and (Thompson, 2002).

To accomplish the aims of the study, an instrument containing a list of norms in the construction and publication of good achievement tests was developed and used by the teachers and current researcher to evaluate the achievement test.

The preliminary form of the questionnaire included seventy norms and nine domains. The researcher revised the questionnaire in light of the feedback received from the jury members. Also, the researchers omitted a main domain with its items which is the matching questions because it isn't used in the English Tawjihi tests. (see appendix A)

The instrument used Likert scale (poor, fair, good, very good, and excellent). It contained fifty items and eight domains of evaluation the English Tawjihi tests based on norms of the construction and publication of good achievement tests and which are:

1. The Instructions.
2. The Content Validity.
3. The Face Validity.
4. The Essay Questions.
5. The Short Answer Questions.
6. The Cloze Questions.
7. The True and False Questions.
8. The Multiple Choice Questions.

3.3 Validity

To ensure the validity of the instrument the researcher used two kinds of validity:

3.3.1 Construct Validity

The researcher referred to and reviewed many resources in evaluation and measurement in constructing the achievement tests and to the results of the previous studies in (Gronlund & Linn, 1990), (Tindal & Marston, 1990), (Cross,1990), (Al-Ekbaty, 1998), (Al-Janazrah,1999) and (Thompson, 2002).

3.3.2 Content Validity

The instrument was prepared with the help of thesis supervisor. To establish its content validity, the researcher gave it to a panel of judges of ten PhD holders in Bethlehem University, Polytechnic University and Hebron University (see appendix B). The jury members were requested to read the items and to indicate whether such items can evaluate Tawjihi achievement tests. In light of their recommendations, suggestions, and comments, the instrument was reviewed and modified. (see appendix C)

3.4 Reliability

To establish the reliability of instrument, it was randomly distributed to twenty teachers as a pilot study at the governmental schools in North Hebron district in the first Semester of the academic year 2006-2007. So, twenty questionnaires were considered and analyzed. Cronbach-alpha procedures were applied. Cronbach-alpha coefficient was calculated for the instrument and it was (0.93).

3.5 Variables

3.5.1 Independent Variables

1. The teacher's gender (female and male).
2. The teacher's qualification.
3. The teacher's experience.
4. The Bloom's Taxonomy objectives.
5. The different question formats.

3.5.2 Dependent Variable

Tawjihi tests that fulfill the norms of the construction and publication of good achievement tests in Palestine.

3.6 Procedures of the Study

The following steps were followed by the researcher:

1. After the instrument of the study was prepared, the researcher contacted his respondents in schools. The purpose of the study and its importance were explained to respondents. They were assured that their responses would be used for academic purposes only. In addition, each teacher was to fill out the questionnaire in person and that their responses will be confidential. Moreover, the researcher attached each questionnaire with a copy of English Tawjihi test in order to assist these teachers to fill out the questionnaire objectively.
2. The researcher got a recommendation letter from the department in Al-Quds University, in order to have permission of the Directorate of Education /South Hebron to facilitate the work at schools.
3. The researcher distributed twenty questionnaires as a pilot study at the governmental schools in North Hebron district in the first Semester of the academic year 2006-2007 to ensure the reliability of the instrument and Cronbach-alpha procedures were applied. Cronbach-alpha coefficient was calculated for the instrument and it was (0.93).
4. The researcher distributed the questionnaires to fifty (50) female and male teachers at (40) forty secondary schools.
5. Then the researcher collected, computed and analyzed the questionnaires.
6. The researcher also evaluated Tawjihi tests using the same instrument to give the thesis more consistent by counting the number of the repeating norms in the tests.
7. Finally, the researcher compared her results with the teacher's results.

3.7 Statistical Analysis

Data were obtained from the teachers and the researcher responses to the questionnaire. Then descriptive statistics were used, which included the mean, the standard deviation and percentages were used calculated at the item level and then at the domain level. The researcher adopted the following grading scale, based on the review of the literature (Al-Janazrah, 1999):

- | | |
|-------------------------------------|------------|
| 1. (4.50) / (90% - 100%) | Excellent. |
| 2. (4.49 - 3.50) / (89% -70%) | very good. |
| 3. (3.49 - 2.50) / (69% - 50%) | good. |
| 4. (2.49- 1.50) / (49% - 30%) | fair. |
| 5. (Less than 1.50) (Less than 30%) | poor. |

The questions of the study were answered by distributing to fifty teachers at forty schools in the Directorate of Education in south Hebron. The researcher assured the reliability and the validity of the instrument as main step in constructing it. After the procedures of the study were applied, the researcher found out the results which appeared in the chapter four.

Chapter 4

Findings of the Study

This study aimed at evaluating the Twajjhi English Tests based on norms of the construction and publication of good achievement tests by the Twajjhi teachers in Hebron school districts. This chapter is divided in two parts. Part one presents the statistical analysis of the data which was provided by the respondents (Tawjjihi English language teachers). The responses of the subjects were fed on a five-Likert-scale which included fifty items. The responses are presented in many tables. Part two presents the researcher's statistical analysis of the data.

4.1 The Teachers Evaluation 'Tawjjihi English Tests Based on the Norms of the Construction and Publication of Good Achievement Tests.

The findings of the teachers' evaluation are presented in the following order: first, findings related to the evaluation of Tawjjihi English tests based on the norms of the construction and publication of good achievement tests for the whole questionnaire (question one, two and three), then on each domain. Second, findings related to the effect of the independent variables (gender, qualification, and experience) on the evaluation of English language teachers of Tawjjihi English tests based on the norms of the construction and publication of good achievement tests on each domain in the questionnaire.

4.2 Findings related to the First Question of the study

1. To what extent do the Tawjjihi English tests match the norms of the construction and publication of good achievement tests in Palestine?

Table 4.1 Means, standard deviations and percentages of teacher's evaluation (The whole questionnaire)

The list of Required Norms of Evaluation of the ESL Tawjjihi Tests Based on the Norms of the Construction and Publication of Good Achievement Tests.	Mean	Standard Deviation	Percentage %
1. The instructions contain marks for each question.	4.20	0.97	84
2. Each question is provided by its own instructions.	4.04	1.14	80.8
3. The instructions contain the number of the questions.	3.92	1.05	78.2
4. The instructions contain allotted time for the test.	3.82	1.04	76.2
5. Copies of the test are clear.	3.76	0.97	75.2
6. Questions are presented in different formats (essay questions, matching, true or false, multiple choices, cloze).	3.72	1.21	74.4
7. The instructions are simple, clear, and definite.	3.70	0.97	74
8. The test contains a suitable space between the instructions and the questions.	3.64	0.98	72.8
9. There is a suitable space between each question and the following one.	3.60	1.07	72
10. Questions are sequenced from the beginning till the end of the test.	3.58	1.01	71.6
11. Content of the questions reflects the textbook objectives.	3.56	1.15	71.2
12. Statements are concise and clear.	3.56	0.97	71.4
13. Consists of a single word or short phrase.	3.52	1.05	70.4

14. The stem is written in simple and understandable language.	3.50	0.93	70
15. Reading skill is adequately assessed.	3.50	1.21	70
16. Questions are phrased so that the task is clearly defined for the student.	3.48	0.97	69.6
17. Charts, tables, or figures are printed clearly and labeled correctly.	3.44	1.23	68.8
18. The questions present the number of words and paragraphs needed for the answer.	3.44	1.26	68.8
19. Questions avoid making the correct answer markedly longer or shorter than the other distractors.	3.38	1.08	67.6
20. The test has optional questions	3.36	1.10	67.2
21. The test is free from spelling, printing, and language mistakes.	3.34	1.24	66.8
22. Statements are arranged so that there is no pattern of answers (such as A,B,C ,A,B,C and C,B,C,C,BC) .	3.32	1.35	66.4
23. Content of questions assesses different cognitive levels. (Bloom's Taxonomy).	3.32	1.06	66.4
24. Blanks are arranged to make answers easy.	3.28	0.93	65.4
25. Distractors are free from double negatives.	3.26	1.00	65.2
26. Statements include a single major idea in each one.	3.24	0.94	64.8
27. Test items use from three to four distractors.	3.24	1.12	64.8
28. Statements are arranged so that there is no discernible pattern of answers (such as T, F, T, F, T, F and T, T, F, F, T, T, F, F) for true and false statements.	3.24	1.32	64.8
29. Questions don't use more than two blanks within an item.	3.22	0.91	64.4
30. Statements are free from double negatives.	3.20	1.03	64
31. Writing skill is adequately assessed.	3.20	1.30	64
32. Questions are free from ambiguities.	3.20	1.04	64
33. Literature is adequately assessed.	3.18	1.11	63.6
34. There is only one correct or best distractor.	3.18	1.10	63.6
35. Test items have a single correct answer.	3.16	1.05	63.2
36. Statements are constructed in a language that is at a lower level of difficulty than the text.	3.14	0.86	62.8
37. All distractors are approximately homogeneous in content, form, and grammatical structure.	3.12	0.94	62.4
38. Questions avoid using all-of-the-above and none-of-the-above distractors.	3.12	1.00	62.4
39. The language of the stem and response distractors is as simple as possible to avoid skill overlap.	3.08	0.89	61.6
40. Questions are proceeding from easy to more difficult items.	3.06	1.19	61.2
41. Questions are phrased so there is only one possible answer.	3.04	1.06	60.8
42. The test contains general instructions.	2.98	1.26	59.6
43. Blanks are either in the middle or at the end of statements rather than at the beginning.	2.94	0.96	58.8
44. Statements are free of specific determiners such as (always, may be, none, never, all, usually, generally, typically, sometimes).	2.90	1.03	58
45. Distractors are free of the words that give verbal clues to the correct answer, such as (always, may be, none, never, all, usually, generally, typically, sometime).	2.80	0.99	56
46. True statements are about the same length as false	2.72	1.12	54.4

statements.			
47. Sgnificant words are omitted from the statement.	2.66	0.98	53.2
48. True, false identification are placed before the statements.	2.46	1.21	49.2
49. Speaking skill is adequately assessed.	1.00	0.00	20
50. Listening skill is adequately assessed.	1.00	0.00	20
Total	3.23	0.51	64.4

In order to answer the first question above, means, standard deviations and percentages were calculated at the item level and then at the domain level.

Accordingly, items of a mean value of more than (2.50) were considered as a prevalent norm of evaluation in the Tawjihi tests. The results of evaluation are presented according to the results of teacher's evaluation of the Tawjihi tests on each item and domain in the questionnaire.

The study instrument included fifty items which were categorized under eight domains. These domains are: instructions, content validity, face validity, essay questions, short-answer questions, cloze questions, true-false questions, and multiple choice questions.

Table (4.1) shows the results of teacher's evaluation of Tawjihi achievement tests at the level of each item in the questionnaire. According to table (4.1), the calculated mean of teacher's evaluation for the English Tawjihi achievement tests ranged between 4.20 and 1.00, and percentages 84%-20%. There is no calculated mean and percentages were less than the criterion adopted by the researcher. It includes the means, the standard deviation, and the percentages for each item in the questionnaire. The calculated means and percentages of responses show that the sixth item in the questionnaire which is the instructions contains marks for each question is 4.20 was very prevalent item in the tests at a percentage of 84% and this is a high percentage. Also, the calculated means of the seventh item in the questionnaire each question is provided by its own instructions is 4.04; at the percentage of 80.8% was very prevalent item in the evaluated tests.

Listening and speaking skills according to table above weren't prevalent at all in the Tawjihi tests. The lowest calculated mean is 1.00 and the percentage is 20%. Teachers' percentage of the response on this domain was 64.4% and the calculated mean was 3.23.

4.3 Findings related to the Second Question of the study

To what extent are the Tawjihi English tests presented in different questions formats?

The table above (4.1) showed that Tawjihi English tests presented the content of curriculum in different questions formats at a percentage of 74.4% and a calculated mean 3.72. Also the essay questions were the most frequently format used in the tests at the calculated mean of 3.42 and percentage of 68.4% and the cloze

questions were the least frequently format used in the tests at the calculated mean 2.92 and percentage of 58.4%.

4.4 Findings related to each domain in the questionnaire

4.4.1 The Instructions

This domain included eight norms. The results showed that teachers evaluated all these items. Table (4.2) showed that the highest calculated mean of responses was 4.20 and percentage 84%, and the lowest calculated mean was 2.98 and percentage 59.6%. Teachers percentage of the response on this domain was 74.2% and the calculated mean 3.72.

Table 4.2 Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of construction and publication of good achievement tests under the domain " The Instructions".

The Instructions	Mean	Standard Deviation	Percentage %
1. The instructions contain marks for each question.	4.20	0.97	84
2. Each question is provided by its own instructions.	4.04	1.14	80.8
3. The instructions contain the number of the questions.	3.92	1.05	78.2
4. The instructions contain allotted time for the test.	3.82	1.04	76.2
5. The instructions are simple, clear, and definite.	3.70	0.97	74
6. The test contains suitable space between the instructions and the questions.	3.64	0.98	72.8
7. The questions present the number of words and paragraphs needed for the answer.	3.44	1.01	68.8
8. The test contains general instructions.	2.98	1.26	59.6
Total	3.72	0.63	74.2

4.4.2 Content Validity

The domain of content validity comprised ten norms. Table (4.3) shows the means, standard deviation, and percentages for teacher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain "The Content Validity". The results showed that the calculated means on the items of this domain were between 3.72 and 1.00 and percentage range between 74.4%- 20%. Teachers' percentage of the response on this domain was 57.4% and the calculated mean 2.87.

Table 4.3 Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of construction and publication of good achievement tests under the domain " The Content Validity"

Content Validity	Mean	Standard Deviation	Percentage %
1. Questions are presented in different formats (essay questions, matching, true or false, multiple choices, cloze).	3.72	1.21	74.4
2. Content of the questions reflects the textbook objectives.	3.56	1.15	71.2
3. Reading skill is adequately assessed.	3.50	1.21	70
4. Content of questions assesses different cognitive levels. (Bloom's Taxonomy).	3.32	1.06	66.4
5. Writing skill is adequately assessed.	3.20	1.30	64
6. Questions are free of ambiguities.	3.20	1.04	64
7. Literature is adequately assessed.	3.18	1.11	63.6
8. Questions are proceeding from easy to more difficult items.	3.06	1.19	61.2
9. Speaking skill is adequately assessed.	1.00	0.00	20
10. Listening skill is adequately assessed.	1.00	0.00	20
Total	2.87	0.61	57.4

4.4.3 Face Validity

This domain contained five evaluation norms. According to the results, in table (4.4) the calculated means for the teachers' evaluation of the items of this domain were between 3.76 and 3.34 and percentage 75.2%- 66.8%. Teachers' percentage of the response on this domain was 71% and the calculated mean was 3.55.

Table 4.4 Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of construction and publication of good achievement tests under the domain " The Face Validity"

Face Validity	Mean	Standard Deviation	Percentage %
1. Copies of the test are clear.	3.76	0.97	75.2
2. There is a suitable space between each question and the following one.	3.60	1.07	72
3. Questions are sequenced from the beginning till the end of the test.	3.58	1.01	71.6
4. Charts, tables, or figures are printed clearly and labeled correctly.	3.44	1.23	68.8
5. The test is free from spelling, printing, and language mistakes.	3.34	1.24	66.8
Total	3.55	0.80	71

4.4.4 Essay Questions

This domain comprised two evaluation norms. Table (4.5) shows that the highest calculated mean of the teacher's evaluation on each item of this domain was 3.48 at percentage of 69.6% and the lowest mean was 3.36 at the percentage of 67.2%. Teachers percentage of the response on this domain was 68.4% and the calculated mean was 3.42.

Table 4.5 Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of construction and publication of good achievement tests under the domain" The Essay Questions"

Essay Questions.	Mean	Standard Deviation	Percentage %
1. Questions are phrased so that the task is clearly defined for the student.	3.48	0.97	69.6
2. The test has optional questions	3.36	1.10	67.2
Total	3.42	0.75	68.4

4.4.5 Short Answer Questions

The fifth domain comprised four norms of publication and constructing good ESL achievement tests. The calculated means in table (4.6) of the teacher's evaluation on each item of this domain were between 3.52 and 3.04 and the calculated percentages were between 70.4%-60.8%. Teachers percentage of the response on this domain was 65.4% and the calculated mean 3.26.

Table 4.6 Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of construction and publication of good achievement tests under the domain" The Short Answer Questions"

Short Answer Questions.	Mean	Standard Deviation	Percentage %
1. Consists of a single word or a short phrase.	3.52	1.05	70.4
2. Blanks are arranged to make answers easy.	3.28	0.93	65.4
3. Questions don't use more than two blanks within an item.	3.22	0.91	64.4
4. Questions are phrased so that there is only one possible answer.	3.04	1.06	60.8
Total	3.26	0.70	65.4

4.4.6 Cloze Questions

This domain included three norms of publication and constructing good ESL achievement tests. The results in the table (4.7) below show that the means value were between 3.16 and 2.66 and percentages between 63.2 & 53.2%. Teachers percentage of the response on this domain was 58.4% and calculated mean was 2.92.

Table 4.7 Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of construction and publication of good achievement tests under the domain" The Cloze Questions"

The Cloze Questions.	Mean	Standard Deviation	Percentage %
1. Test items have a single correct answer.	3.16	1.05	63.2
2. Blanks are either in the middle or at the end of statements rather than at the beginning.	2.94	0.96	58.8
3. Sgnificant words are omitted from the statement.	2.66	0.98	53.2
Total	2.92	0.66	58.4

4.4.7 True False Questions

True and false questions are the seventh domain in the questionnaire which included eight evaluation norms of constructing and publication of good ESL achievement tests. The results in the table (4.9) below showed that the most prevalent item's mean was 3.56 at percentage of 71.2% and the lowest prevalent mean was 2.46 at percentage of 49.2%.

Teachers' percentage of the response on this domain was 61% and the calculated mean was 3.06.

Table 4.8 Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of construction and publication of good achievement tests under the domain" The True False Questions"

True False Questions.	Mean	Standard Deviation	Percentage %
1. Statements are concise and clear.	3.56	0.97	71.2
2. Statements include a single major idea in each one.	3.24	0.94	64.8
3. Statements are arranged so that there is no discernible pattern of answers (such as T, F, T, F, T, F and T, T, F, F, T, T, F, F) for true and false statements.	3.24	1.32	64.8
4. Statements are free from double negatives.	3.20	1.03	64
5. Statements are constructed in a language that is at a lower level of difficulty than the text.	3.14	0.86	62.8
6. Statements are free from the words that give verbal clues to the correct answer, such as (always, may be, none, never, all, usually, generally, typically, sometime).	2.80	0.99	56
7. True statements are about the same length as false statements.	2.72	1.12	54.4
8. True, false identification are placed before the statements.	2.46	1.21	49.2
Total	3.06	0.57	61

4.4.8 Multiple Choice Questions

This is the last domain which comprised ten evaluation norms based on the construction and publication of the good ESL achievement tests. The calculated means in table (4.9) were between 3.50 and 2.80. The calculated percentages were between 70%-56%.Teacher's percentage of the response on this domain was 64% and calculated mean was 3.20.

Table 4.9 Means, standard deviations, and percentages of teacher's evaluation for Tawjihi tests based on the norms of construction and publication of good achievement tests under the domain " The Multiple Choice Questions"

Multiple Choice Questions.	Mean	Standard Deviation	Percentage %
1. The stem is written in simple and understandable language.	3.50	0.93	70
2. Questions avoid using all-of-the-above and none-of-the-above distractors.	3.12	1.00	62.4
3. Statements are arranged so that there is no pattern of answers (such as A,B,C ,A,B,C and C,B,C,C,BC) .	3.32	1.35	66.4
4. Test items use from three to four distractors.	3.24	1.12	64.8
5. Statements are free from double negatives.	3.20	1.03	64
6. There is only one correct or best distractor.	3.18	1.10	63.6
7. All distractors are approximately homogeneous in content, form, and grammatical structure.	3.12	0.94	62.4
8. Questions avoid using all-of-the-above and none-of-the-above distractors.	3.12	1.00	62.4
9. The language of the stem and response distractors is as simple as possible to avoid skill overlap.	3.08	0.89	61.6
10. Distractors are free from the words that give verbal clues to the correct answer, such as (always, may be, none, never, all, usually, generally, typically, sometime).	2.80	0.99	56
Total	3.20	0.64	64

4.5 Findings related to the sequence of domains in the questionnaire

The table below (4.10) presented the sequence of domains in the questionnaire as they appeared in the evaluated ESL Tawjihi tests. The findings show that instructions was the most prevalent domain in the test, the calculated value mean was 3.72 at a percentage of 74.2%, while content validity in the tests was the lowest domain, the calculated mean was 3.22 at a percentage of 57.4%. Also, the essay questions were the most frequently used in the tests at the calculated mean 3.42 and percentage of 68.4%.

Table 4.10 Means, standard deviations, and percentages for teacher's evaluation for Tawjihi tests based on the norms of construction and publication of good achievement tests for each domain in the questionnaire.

Domains	Mean	Standard Deviation	Percentage%
1. Instructions	3.72	0.63	74.2
2. Face Validity	3.55	0.80	71
3. Essay Questions	3.42	0.75	68.4
4. Short/ Answer Questions	3.27	0.70	65.4
5. Multiple- Choice Questions	3.20	0.64	64
6. True/ False Questions	3.06	0.57	61
7. Cloze Questions	2.92	0.66	58.4
8. Content Validity	2.87	0.61	57.4
Total	3.23	0.51	64.4

4.6 Findings related to the effects of the independent variables (gender, qualification, and experience) on the evaluation of English language teachers of Tawjihi English tests based on the norms of the construction and publication of good achievement tests on each domain in the questionnaire.

This study aimed at the evaluation of the Tawjihi ESL tests according to the norms of construction and publication of good achievement tests by Tawjihi English language teachers. It was also an attempt to study the effect of each one of the independent variables (gender, qualification, and experience). The results are presented in the following tables of the level of each domain of the study instrument. The presentation of these results is arranged according to the order of the domains in the study instrument.

4.6.1 Findings related to the effects of the independent variable (gender) on the evaluation of English language teachers of Tawjihi English tests based on the norms of the construction and publication of good achievement tests on each domain in the questionnaire.

The results of applying the statistical analysis (mean, standard deviation, and percentage) to the data related to the effects of teacher's gender are presented in table (4.11). It showed the effect of gender in the following areas: instructions, content validity, face validity, essay questions, short answer questions, cloze questions, true false questions, and multiple choice questions.

Table 4.11 Findings related to the effects of the independent variables (gender) on the evaluation of English language teachers of Tawjihi English tests based on the norms of the construction and publication of good achievement tests on each domain in the questionnaire

Gender		Domains								Total degree
		Instruct	Content Validity	Face Validity	Essay Quest	Short Quest	Cloze Quest	T\F Quest	Multi-Quest	
Male (24)	M	3.68	2.81	3.44	3.38	3.18	2.90	3.11	3.30	3.22
	S.D	0.61	0.51	0.69	0.63	0.65	0.68	0.57	0.55	0.46
	Per %	73.6	56.2	68.8	67.6	63.6	58	62.2	66	64.4
Female (26)	M	3.75	2.93	3.64	3.46	3.35	2.94	3.00	3.11	3.24
	S.D	0.66	0.69	0.91	0.86	0.74	0.67	0.59	0.70	0.56
	Per %	75	58.6	72.8	96.2	67	58.8	60	62.2	64.8
Total (50)	M	3.72	2.87	3.54	3.42	3.27	2.92	3.06	3.20	3.22
	S.D	0.63	0.60	0.81	0.75	0.70	0.66	0.57	0.64	0.51
	Per%	74.4	57.4	70.8	68.4	65.4	58.4	61.2	64	64.4

Table (4.11) indicates that there is no difference in the ratings of the English teachers due to gender, the total mean degree for female teachers in evaluating the eighth domains was 3.24 at a percentage 64.8%, while male teacher's total mean degree for evaluating the eighth domains was 3.22 at percentage 64.4%. Both male and female teachers agreed that instructions were the most frequent domain in the tests, the calculated mean was 3.72 at percentage 74.4% and the content validity domain was the lowest calculated mean it was 2.87 at percentage 57.4%.

4.6.2 Findings related to the effects of the independent variable (experience) on the evaluation of English language teachers of Tawjihi English tests based the norms of the construction and publication of good achievement tests on each domain in the questionnaire.

Table 4.12 Findings related to the effects of the independent variables (experience) on the evaluation of English language teachers of Tawjihi English tests based the norms of the construction and publication of good achievement tests on each domain in the questionnaire

Experience		Domains								Total
		Instruct	Content Validity	Face Validity	Essay Quest	Short Quest	Cloze Quest	T/F Quest	Multi-Quest	
less than 5 Y (8)	M	4.04	3.34	3.90	4.00	3.56	3.13	3.33	3.29	3.53
	S.D	0.68	0.25	0.58	0.38	0.68	0.82	0.46	0.59	0.30
	Per %	80.8	66.8	78	80	71.2	62.6	66.6	65.8	70.6
5-10 years (24)	M	3.59	2.76	3.32	3.35	3.26	2.89	3.33	3.30	3.18
	S.D	0.58	0.56	0.78	0.54	0.61	0.60	0.46	0.65	0.49
	Per%	71.8	55.2	66.4	67	65.2	57.8	66.6	66	63.6
More than 10 years (18)	M	3.74	2.82	3.69	3.25	3.13	2.88	3.06	3.02	3.16
	S.D	0.66	0.70	0.88	0.99	0.81	0.69	0.61	0.62	0.57
	Per %	74.8	56.4	73.8	65	62.6	57.6	61.2	60.4	63.2
Total (50)	M	3.72	2.88	3.54	3.42	3.27	2.92	2.93	3.20	3.22
	S.D	0.63	0.61	0.81	0.75	0.70	0.66	0.55	0.64	0.51
	Per%	74.4	57.6	70.8	68.4	65.4	58.4	58.6	64	64.4

Based on the results shown in table (4.12), that there was a significant difference in the ratings of the English teachers due to experience, for teachers who have an experience less than five years in teaching Tawjihi students, the total mean degree for evaluating the eighth domain was 3.53 at percentage 70.6%, while teachers who have experience in teaching Tawjihi students more than five years, the total mean degree for evaluating the eighth domains was 3.16 at percentage 63%. The table also showed that all teachers with different years of experience agreed that the instructions were the most prevalent domain in the Tawjihi tests, the calculated mean was 3.72 at percentage 74.4% and they agreed that content validity domain was the least evident domain in the Tawjihi tests.

4.6.3 Findings related to the effects of the independent variable (qualification) on the evaluation of English language teachers of Tawjihi English tests based the norms of the construction and publication of good achievement tests on each domain in the questionnaire.

Table 4.13 Findings related to the effects of the independent variables (qualification) on the evaluation of English language teachers of Tawjihi English tests based on the norms of the construction and publication of good achievement tests on each domain in the questionnaire

Qualification		Domains								Total Degree
		Instruct	Content Validity	Face Validity	Essay Quest	Short Quest	Cloze Quest	T\F Quest	Multi-Quest	
B.A (46)	M	3.72	2.67	3.50	3.41	3.24	2.92	3.09	3.21	3.22
	SD	0.63	0.56	0.75	0.77	0.70	0.68	0.58	0.64	0.50
	Per%	74.4	53.4	70	68.2	64.8	58.4	61.8	64.2	64.4
M.A (4)	M	3.69	2.95	4.10	3.50	3.56	2.91	2.72	3.13	3.25
	SD	0.79	1.12	1.32	0.58	0.63	0.41	0.36	0.70	0.70
	Per%	73.8	59	82	70	71.2	58.2	54.4	62.6	65
Total (50)	M	3.71	2.87	3.54	3.42	3.27	2.92	3.06	3.20	3.23
	SD	0.63	0.60	0.81	0.75	0.69	0.66	0.57	0.64	0.51
	Per%	74.2	57.4	70.8	68.4	65.4	58.4	61.2	64	64.4

As shown in the table above (4.13), there is no difference in the ratings of the English teachers due to qualification, for teachers who are holding MA qualification, the total mean degree for evaluating the eighth domains was 3.25, while teachers who are holding the BA qualification, the total mean degree for evaluating the eighth domains was 3.22, and there wasn't teachers holding diploma qualification in the population of the study. Teachers who are holding BA qualification evaluated instructions as the highest domain in the tests with 3.72 calculated mean, where as content validity was the least prevalent domain in the tests with calculated mean 2.67. The above table showed that teachers who are holding MA qualification evaluated face validity as the highest domain in the questionnaire 4.10 mean and true false questions were the lowest mean 2.72.

4.2 The Researcher's Evaluation Tawjihi English Tests Based on Norms of the Construction and Publication of Good Achievement Tests.

The findings of the researcher's evaluation are presented in the following order: first, findings related to the evaluation of Tawjihi English tests based the norms of the construction and publication of good achievement tests for the whole questionnaire and answering the questions of the study (one, two, three), then on each domain.

4.2.1 Findings related to the first question of the study

To what extent do the Tawjihi English tests fulfill the norms of the construction and publication of good achievement tests in Palestine?

Table 4.14 Means, standard deviations and percentages of researcher's evaluation (The whole questionnaire)

The list of Required Norms of Evaluation of the ESL Tawjihi Tests Based on Norms of the Construction and Publication of Good Achievement Tests.	Mean	Standard Deviation	Percentage %
1. Distractors are free from the words that give verbal clues to the correct answer, such as (always, may be, none, never, all, usually, generally, typically, sometime).	5.00	0.00	100
2. There is only one correct or best distractor.	5.00	0.00	100
3. Questions are free from ambiguities.	5.00	0.00	100
4. The instructions contain allotted time for the test.	5.00	0.00	100
5. Test items have a single correct answer.	4.83	0.41	96.6
6. Questions are sequenced from the beginning till the end of the test.	4.83	0.41	96.6
7. The instructions contain marks for each question.	4.83	0.41	96.6
8. True statements are about the same length as false statements.	4.67	0.82	93.2
9. Questions don't use more than two blanks within an item.	4.67	0.82	93.2
10. The language of the stem and response distractors is as simple as possible to avoid skill overlap.	4.67	0.82	93.2
11. Statements are concise and clear.	4.67	0.82	93.2
12. Questions are phrased so there is only one possible answer.	4.50	0.55	90
13. The stem is written in simple and understandable language.	4.33	1.03	86.6
14. Blanks are either in the middle or at the end of statements rather than at the beginning.	4.33	1.03	86.6
15. Questions are phrased so that the task is clearly defined for the student.	4.33	1.03	86.6
16. Statements are constructed in a language that is at a lower level of difficulty than the text.	4.17	1.33	83.2
17. Statements include a single major idea in each one.	4.17	1.33	83.2
18. Charts, tables, or figures are printed clearly and labeled correctly.	4.17	1.33	83.2
19. The test is free from spelling, printing, and language mistakes.	4.17	1.33	83.2
20. Statements are free from double negatives.	4.00	1.26	80
21. Significant words are omitted from the statement.	4.00	1.26	80
22. Copies of the test are clear.	4.00	1.26	80
23. Content of the questions reflects the textbook objectives.	4.00	1.26	80
24. Each question is provided by its own instructions.	4.00	1.26	80
25. Blanks are arranged to make answers easy.	3.83	1.47	76.6
26. The instructions are simple, clear, and definite.	3.83	1.47	76.6
27. The test has optional questions.	3.67	1.52	73.2
28. Distractors are free from double negatives.	3.67	1.52	73.2
29. Questions avoid making the correct answer markedly longer or shorter than the other distractors.	3.67	1.52	73.2
30. Questions avoid using all-of-the-above and none-of-the-above distractors.	3.67	1.52	73.2
31. Statements are arranged so that there is no discernible pattern of answers (such as T, F, T, F, T, F and T, T, F, F, T, T, F, F) for true and false statements.	3.67	1.52	73.2
32. Questions are presented in different formats (essay questions, matching, true or false, multiple choices, completion).	3.33	1.52	66.6

33. Statements are free from specific determiners such as (always, may be, none, never, all, usually, generally, typically, sometimes).	3.00	1.41	60
34. Statements are arranged so that there is no pattern of answers (such as A,B,C ,A,B,C and C,B,C,C,BC) .	2.67	1.37	53.2
35. Test items use from three to four distractors.	2.67	1.37	53.2
36. Reading skill is adequately assessed.	2.67	1.37	53.2
37. All distractors are approximately homogeneous in content, form, and grammatical structure.	2.50	1.22	50
38. There is a suitable space between each question and the following one.	2.50	1.22	50
39. Writing skill is adequately assessed.	2.50	1.22	50
40. The test contains suitable space between the instructions and the questions.	2.33	1.21	46.6
41. Consists of a single word or short phrase.	2.33	1.21	46.6
42. Questions are proceeding from easy to more difficult items.	2.33	1.21	46.6
43. The questions present the number of words and paragraphs needed for the answer.	2.33	1.21	46.6
44. Literature is adequately assessed	2.00	0.89	40
45. Content of questions assesses different cognitive levels. (Bloom's Taxonomy).	2.00	0.89	40
46. The instructions contain the number of the questions.	1.16	0.41	23.2
47. True, false identification are placed before the statements.	1.00	0.00	20
48. Speaking skill is adequately assessed.	1.00	0.00	20
49. Listening skill is adequately assessed.	1.00	0.00	20
50. The test contains general instructions.	1.00	0.00	20
Total	3.47	0.06	69.4

Table (4.14) shows the results of researcher's evaluation of Tawjihi achievement tests at the level of each item in the questionnaire. It includes the means, the standard deviation, and the percentages for each item in the questionnaire. According to table (15), the calculated mean of researcher's evaluation for the English Tawjihi achievement tests ranged between 5.00 at percentage of 100% and 1.00 at percentage of 20%, no calculated mean was less than the criterion adopted by the researcher. The calculated means and percentages of responses show that:

1. Distractors are free of the words that give verbal clues to the correct answer, such as (always, may, none, never, all, usually, generally, typically, sometime).
2. There is only one correct or best distractor.
3. Questions are free of ambiguities.
4. The instructions contain allotted time for the test.

The above items were prevalent items in the all the tests that the researcher analyzed at a percentage of 100% and calculated mean 5.00. Also, the calculated means and percentages of responses show that the following items were the least prevalent items in the evaluated tests at percentage of 20% and calculated mean was 1.00.

1. True, false identification are placed before the statements.
2. Speaking skill is adequately assessed.
3. Listening skill is adequately assessed.
4. The test contains general instructions.

4.2.2 Findings Related to the Second Question of the study

To what extent are the Tawjihi English tests presented in different questions formats?

The table above (4.14) showed that Tawjihi English tests presented the content of curriculum in different questions formats at a percentage of 66.6% and calculated mean 3.33. The findings show that essay questions were the most prevalent domain in the test; the calculated value mean was 4.00 at a percentage of 80%, while the cloze questions was the least frequently used at the calculated mean was 3.47 at a percentage of 69.4%

4.2.3 Findings related to Each Domain in the questionnaire

4.2.3.1 The Instructions

This domain included eight norms. Table (4.15) showed that the highest calculated means of responses was 5.00 and percentage 100%, and the lowest calculated means was 1.00 and percentage 20%. The total evaluation percentage of this domain was 72.6% and calculated mean 3.62.

Table 4.15 Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain "The Instructions".

The Instructions.	Mean	Standard Deviation	Percentage %
1. The instructions contain allotted time for the test.	5.00	0.00	100
2. The instructions contain marks for each question.	4.83	0.41	96.6
3. Each question is provided by its own instructions.	4.00	1.26	80
4. The instructions are simple, clear, and definite.	3.83	1.47	76.6
5. The instructions contain the number of the questions.	1.16	0.41	23.2
6. The questions present the number of words and paragraphs needed for the answer.	2.33	1.21	46.6
7. The instructions contain the number of the questions.	1.16	0.41	23.2
8. The test contains general instructions.	1.00	0.00	20
Total	3.62	1.02	72.6

4.2.3.2 The content validity

The domain of content validity comprised ten norms. Table (4.16) shows the means, standard deviation, and percentages for researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain "The Content Validity". The results showed that the calculated means on the items of this domain were between 5.00 and 1.00 and percentage range between 100%- 20%. The total evaluation of this domain was 51.6% and calculated mean 2.58.

Table 4.16 Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Content Validity"

The Content Validity.	Mean	Standard Deviation	Percentage %
1. Questions are free from ambiguities.	5.00	0.00	100
2. Content of the questions reflects the textbook objectives.	4.00	1.26	80
3. Questions are presented in different formats (essay questions, matching, true or false, multiple choices, completion).	3.33	1.52	66.6
4. Reading skill is adequately assessed.	2.67	1.37	53.2
5. Writing skill is adequately assessed.	2.50	1.22	50
6. Questions are proceeding from easy to more difficult items.	2.33	1.21	46.6
7. Literature is adequately assessed.	2.00	0.89	40
8. Content of questions assesses different cognitive levels. (Bloom's Taxonomy).	2.00	0.89	40
9. Speaking skill is adequately assessed.	1.00	0.00	20
10. Listening skill is adequately assessed.	1.00	0.00	20
Total	2.58	1.18	51.6

4.2.3.3 The Face Validity

This domain contained five evaluation norms. According to the results, in table (4.17) the calculated means for the researcher's evaluation on the items of this domain were between 4.83 and 2.50 and percentage 96.6%- 50%.The total percentage of evaluation this domain was 78.6% and calculated mean was 3.93.

Table 4.17 Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Face Validity"

The Face Validity.	Mean	Standard Deviation	Percentage %
1. Questions are sequenced from the beginning till the end of the test.	4.83	0.41	96.6
2. Charts, tables, or figures are printed clearly and labeled correctly.	4.17	1.33	83.2
3. The test is free from spelling, printing, and language mistakes.	4.17	1.33	83.2
4. Copies of the test are clear.	4.00	1.26	80
5. There is suitable space between each question and the following one.	2.50	1.22	50
Total	3.93	1.06	78.6

4.2.3.4 The Essay Questions

This domain comprised two evaluation norms. Table (4.18) shows that the highest calculated mean of evaluation the each item of this domain was 4.33 at a percentage of 86.6% and the lowest mean was 3.67 at a percentage of 73.2%.The total percentage of evaluation for this domain was 80% and the calculated mean 4.00.

Table 4.18 Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Essay Questions"

The Essay Questions.	Mean	Standard Deviation	Percentage %
1. Questions are phrased so that the task is clearly defined for the student.	4.33	1.03	86.6
2. The test has optional questions.	3.67	1.52	73.2
Total	4.00	1.26	80

4.2.3.5 The Short Answer Questions

The fifth domain comprised four norms of publication and constructing good ESL achievement tests. The calculated means in table (4.19) of the teacher's evaluation on each item of this domain were between 4.67 and 2.33 and the calculated percentages were between 93.2%-46.6%.The total percentage of evaluation was 76.6% and the calculated mean 3.83.

Table 4.19 Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Short Answer Questions"

The Short Answer Questions.	Mean	Standard Deviation	Percentage %
1. Questions don't use more than two blanks within an item.	4.67	0.82	93.2
2. Questions are phrased so there is only one possible answer.	4.50	0.55	90
3. Blanks are arranged to make answers easy.	3.83	1.47	76.6
4. Consists of a single word or short phrase.	2.33	1.21	46.6
Total	3.83	0.41	76.6

4.2.3.6 The Cloze Questions

This domain included three norms of the publication and constructing good ESL achievement tests. The results in the table (4.20) below show that the means value were between 4.83 and 4.00 and percentages between 96.6%- 80%.The total percentage of this domain was 69.4% and the calculated mean was 3.47.

Table 4.20 Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Cloze Questions"

The Cloze Questions.	Mean	Standard Deviation	Percentage %
1. Test items have a single correct answer.	4.83	0.41	96.6
2. Blanks are either in the middle or at the end of statements rather than at the beginning.	4.33	1.03	86.6
3. Significant words are omitted from the statement.	4.00	1.26	80
Total	3.47	0.98	69.4

4.2.3.7 The True False Questions

True and false questions are the seventh domain in the questionnaire which included eight evaluation norms of constructing and publication of good ESL achievement tests. The results in the table (4.21) below presented that the most prevalent item's mean was 4.67 at percentage of 93.2% and the lowest prevalent mean was 1.00 at percentage of 20%. The total evaluation of this domain was 73.4% and the calculated mean was 3.67.

Table 4.21 Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The True False Questions"

The True False Questions.	Mean	Standard Deviation	Percentage %
1. True statements are about the same length as false statements.	4.67	0.82	93.2
2. Statements are concise and clear.	4.67	0.82	93.2
3. Statements are constructed in a language that is at a lower level of difficulty than the text.	4.17	1.33	83.2
4. Statements include a single major idea in each one.	4.17	1.33	83.2
5. Statements are free from double negatives.	4.00	1.26	80
6. Statements are arranged so that there is no discernible pattern of answers (such as T, F, T, F, T, F and T, T, F, F, T, T, F, F) for true and false statements.	3.67	1.52	73.2
7. Statements are free from specific determiners such as (always, may be, none, never, all, usually, generally, typically, sometimes).	3.00	1.41	60
8. True, false identification are placed before the statements.	1.00	0.00	20
Total	3.67	0.82	73.4

4.2.3.8 The Multiple Choice Questions

This is the last domain which comprised ten evaluation norms based on the construction and publication of the good ESL achievement tests. The calculated means in table (4.22) were between 5.00 and 2.50. The calculated percentages were between 50%-100%.The percentages for this domain was 75.6% and the calculated mean was 3.78.

Table 4.22 Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests under the domain" The Multiple Choice Questions"

The Multiple Choice Questions.	Mean	Standard Deviation	Percentage %
1. Distractors are free from the words that give verbal clues to the correct answer, such as (always, may be, none, never, all, usually, generally, typically, sometime).	5.00	0.00	100
2. There is only one correct or best distractor.	5.00	0.00	100
3. The language of the stem and response distractors is as simple as possible to avoid skill overlap.	4.67	0.82	93.2

4. The stem is written in simple and understandable language.	4.33	1.03	86.6
5. Distractors are free from double negatives.	3.67	1.52	73.2
6. Questions avoid making the correct answer markedly longer or shorter than the other distractors.	3.67	1.52	73.2
7. Questions avoid using all-of-the-above and none-of-the-above distractors.	3.67	1.52	73.2
8. Statements are arranged so that there is no pattern of answers (such as A,B,C ,A,B,C and C,B,C,C,BC) .	2.67	1.37	53.2
9. Test items use from three to four distractors.	2.67	1.37	53.2
10. All distractors are approximately homogeneous in content, form, and grammatical structure.	2.50	1.22	50
Total	3.78	0.97	75.6

4.2.5 Findings related to the Sequence of domains in the questionnaire

The table below (4.23) represented the sequence of domains in the questionnaire as they appeared in the evaluated ESL Tawjihi tests. The findings show that essay questions were the most prevalent domain in the test; the calculated value mean was 4.00 at a percentage of 80%, while content validity in the tests was the lowest domain, and the calculated mean was 2.58 at a percentage of 51.6%. Also, the essay questions were the most frequently used format in the tests at the calculated mean 4.00 and percentage of 80% and the cloze questions was the least frequently used format at the calculated mean was 3.47 at a percentage of 69.4%

Table 4.23 Means, standard deviations, and percentages of researcher's evaluation for Tawjihi tests based on the norms of constructions and publication of good achievement tests for each domain in the questionnaire.

Domains	Mean	Standard Deviation	Percentage %
1.Essay Questions	4.00	1.26	80
2.Face Validity	3.93	1.06	78.6
3.Short/ Answer Questions	3.83	0.41	76.6
4.Multiple-Choice Questions	3.78	0.97	75.6
5.True/ False Questions	3.67	0.82	73.4
6.Instructions	3.62	1.02	72.6
7.Cloze Questions	3.47	0.98	69.4
8.Content Validity	2.58	1.18	51.6
Total	3.47	0.06	69.4

The results of the teachers' evaluation showed that 4.20 is the highest value at 84% and the lowest value is 1.00 at 20%. Also, the total evaluation for the whole questionnaire is 69.4. also the results of the researcher's evaluation showed that 5.00 at percent of 100% is the highest value and 1.00 at 20% is the lowest value. The explanation of these results will be in the next chapter.

Chapter 5

Discussion of the Findings and Recommendations

In this chapter, the researcher discusses the results of evaluation of the Twajih English Tests by both the English teachers and the researcher herself according to the norms of construction and publication of good achievement tests. Also, the effects of the independent variables (gender, experience and qualification) are discussed.

5.1 Discussion of the findings of data analysis related to the evaluation of the norms of evaluating the Twajih English Tests based on norms of the construction and publication of good achievement tests for the whole questionnaire.

5.1.1 Discussion of the findings of data analysis related to the First Question of the study.

To what extent do the Tawjih English tests fulfill the norms of the construction and publication of good achievement tests in Palestine?

The results of the study showed that English language teachers evaluated all the fifty items stated in the questionnaire. The highest percentage value was 84% which shows that the instruction items in the questionnaire were very prevalent items in the tests and this is a high percentage. This finding is similar to the finding of Marshall (1967) who states that the tests he analyzed have clear and explicit instructions. Also, Thompson (2002) stated in his study that tests should include simple, clear, intuitive instructions and procedures.

This finding is different from the researcher's finding that distractors are free of the words that give verbal clues to the correct answer, such as (always, may be, none, never, all, usually, generally, typically, sometime) were the most prevalent items in the tests at percentage 100%. The researcher's finding is different from the results in Victor (1972), Talmir (1991) and Al-Janazrah (1999) studies who stated that teachers have problems in designing multiple choice questions, for example the distractors weren't written appropriately and accurately.

Listening and speaking are very important skills that should be included in constructing ESL achievement test were absent in the Tawjih tests. Both teachers and the researcher are agreed on this finding that these skills were 20% and it is very low value. These findings agreed with the results in Lerkkanen (2004), Abbott (2006) and Cohen (2006) studies which showed that tests evaluate the reading skill while listening, writing, and speaking were avoided.

Also, findings of this study were similar to Nicosia (2005), who stated in her study that teachers in constructing their tests concentrate only on the writing skill rather than reading, listening and speaking. Different from the researcher and teachers findings; Nair Vanugopal (1991), Al-Ekbaty (1998), and Mills (1998) found that

tests focus on testing students in listening and speaking skills because they reduced tests anxiety.

Listening and speaking aren't tested in the Tawjihi tests because of the lack of facilities, money and suitable equipment.

The percentage of evaluation that was given to the whole questionnaire by teachers was 64.4% and the researcher evaluation was 69.4%. So teachers and the researcher agreed that Tawjihi English tests match the norms of the construction and publication of good achievement tests in a medium average and good percentage. These results were similar to Marso & Pigge (1993) who stated that teachers have good testing tools due to frequent use of self-constructed tests in their classroom. It also agrees with the findings of Johnstone(2003), Donlan(2005), and Levacic (2006) that tests play a main role in students performance when they are clearly constructed. Moreover, this finding agrees with Popham (2005) who said that standardized tests constructed by experienced teachers and experts were better than teacher-made tests.

This is so because these tests were constructed by professional and experienced teachers who are holding educational qualification in testing. Also these tests were revised by the Directorate General of Assessment, Evaluation & Examinations.

Table 5.1 Discussion of the findings of data analysis related to the First Question of the study. (Teachers and the researcher)

The list of Required Norms of Evaluation the ESL Tawjihi Tests Based on Norms of The Construction and Publication of Good Achievement Tests.	Teachers' Mean	Researcher's Mean
1. The instructions contain marks for each question.	4.20	4.83
2. Distractors are free of the words that give verbal clues to the correct answer, such as (always, may be, none, never, all, usually, generally, typically, sometime)..	2.80	5.00
Total	3.23	3.47

5.1.2 Discussion of the findings of data analysis related to the Second Question of the study.

To what extent are the Tawjihi English tests presented in different questions formats?

The results of the teachers' evaluation showed that Tawjihi English tests present the content of curriculum in different question formats at a percentage of 74.4%. Also the essay questions were the most frequent format used in the tests at the percentage of 68.4% and the cloze questions were the least frequent format at the percentage of 58.4%.

This finding is similar to the researcher's results. Although the percentages are different, the content of curriculum which was presented in different question formats was 66.6%, the essay questions were 80% and the cloze questions were 69.4%. Both teachers and the researcher agreed that tests presented the content of the curriculum in an acceptable value. These findings were similar to Marshall &

Hales (1971) and William and others (1984), Al-Omar (1989), Abu-Taleb (1991), The Administration for Education Research in Jordan (1995), Al-Ekbaty, (1998) and Al-Janazrah (1999) findings which revealed that essay tests were the most frequent format used to test students especially the Arabic and English language teachers.

Nevertheless, this finding disagrees with the findings of Marso & Pigge (1988) who found out that teachers used multiple choice, matching and short questions in their achievement tests mostly and they rarely used the essay questions

Also, the result that cloze questions were the least frequently used agreed with the findings in Abu-Taleb (1991) and Al-Janazrah (1999) that cloze questions were rarely used, but this finding disagreed with Al-Ekbaty (1998) and Al-Omar (1989) results that cloze tests are the most frequent format especially in Math and Science.

Teachers use essay questions mostly because they are less time-consuming to construct. Essay tests also give teachers an opportunity to comment on student's progress, the quality of their thinking, the depth of their understanding, and the difficulties they may be having.

Matching questions were never used in Tawjihi tests. This may be due to the fact that teachers found that matching questions, it is extremely difficult to develop a set of premises for a matching exercise that will measure high levels of the cognitive domain. Also, it is difficult to find enough important and homogeneous ideas to form a matching set. Moreover, the construction of a homogeneous set of matching items often places an overemphasis on a rather small portion of the content area to be tested.

5.1.3 Discussion of the findings of data analysis related to each domain in the questionnaire.

5.1.3.1 The Instructions

The results of both teachers and the researcher's evaluation showed that the instructions contain marks for each question. Also, each question is provided by its own instructions and the instructions contain the number of the questions which scored the highest percentages. In addition, the test which contains general instructions has the lowest prevalent item in the instructions domain. Moreover, teachers and the researcher gave the same high value.

Similar findings were revealed by Marshall (1967), that tests have clear and adequate instructions. Also, they agreed with Al-Janazrah (1999) results that instructions contain marks for each question because it is easy to correct the student's responses. Moreover he found that most of tests lack test general instructions because teachers used to give them orally to students during the test.

This finding disagrees with the findings of Madsen (1982), Kirby & Oescher (1987), Al-Omar (1989), Al-Janazrah (1999) and Thompson (2002) which showed that most of tests lack presence of test instructions, the provided instructions were unclear, inaccurate and inadequate so they were a source of test anxiety and have an influence on the test performance.

5.1.3.2 The Content Validity

The results of the study showed that teachers and the researcher agreed that Questions are presented in different formats (essay questions, matching, true or false, multiple choices, cloze) and the content of the questions that reflect the textbook objectives were the most prevalent items in Tawjihi tests and the items "speaking and listening were adequately assessed" weren't used at all in Tawjihi tests. In addition to that, both teachers and the researcher agreed that the content of the test isn't sufficient prevalent in Tawjihi tests.

The findings related to that Tawjihi test is a level test that allows students to follow up their higher education, so tests measure the success in teaching the curriculum and so "teaching to the test" which agreed with the findings of The National Research Council, (2002), Fleming and Chambers (1983), Johnstone (2003), Kirby & Oescher (1987), Al-Aga (1994), Al-Ekbaty (1998), Al-Janazrah (1999), Thompson (2002) and Haladyna (2002) who stated that tests don't cover the curriculum content, they present this content in different formats of questions and in different cognitive levels.

Moreover, listening and speaking which are main parts of the content validity were avoided to be used in the Tawjihi tests agreed with the results in Lerkkanen (2004), Nicosia (2005), Abbott (2006), and Cohen (2006).

The decreasing of the content validity in Tawjihi tests may be due to the result in failure to conform to the table of specifications and thus cause a bias in content sampling. Moreover, the content of the test wasn't presented in different formats especially in the objective questions.

5.1.3.3 The Face Validity

According to the results of this domain, teachers and the researcher agreed that all the items of this domain were covered in Tawjihi tests, they found out that:

1. The questions are sequenced from the beginning till the end of the test.
2. The test is free of spelling, printing, and language mistakes.
3. Charts, tables, or figures are printed clearly and labeled correctly.
4. Copies of the test are clear.
5. There is suitable space between each question and the following one.

The findings of this domain disagree with the findings of Scannell & Marshall (1966), Marshall (1967), Al-Omar (1989), Administration for Education Research in Jordan (1995), Al-Omarie (1997), Al-Janazrah (1999), Thompson (2002) and Dolan (2005) that tests was unclear handwriting tests, crowded questions, and the

readability of tests were poor. Also, the tests were full of spelling, punctuation, and language mistakes or errors.

Disagreement between the findings of this study and studies related to the same topic can be attributed to the fact that Tawjihi tests were printed tests and constructed by experienced teachers. In addition to that, these tests are revised by the Directorate General of Assessment, Evaluation & Examinations in order to be accessible and understandable to wide variety of students.

5.1.3.4 The Essay Questions

The teacher's evaluation on each item of this domain was 68.4% and it is a medium value, This finding is different from the researcher's results that the total evaluation of this domain was 80% and this is high value. Both of them agreed that:

1. Essay tests are phrased so that the task is clearly defined for the student.
2. Essay questions have optional questions.
3. Essay questions were mostly used.

This result is similar to Marshall & Hales (1971), William and others(cited in Al-Janazrah) (1984), Al-Omar (1989), Ekbaty (1998) Abu-Taleb (1991) and Al-Janazrah (1999) who found out that tests mostly used essay tests which fell into three categories: simple recall questions, short answer questions and discussion questions.

At the same time these findings were different from Marso & Pigge (1988) and Al-Janazrah (1999) who stated that tests rarely used essay questions and they measure the lower cognitive levels. Also, the researchers in the Administration for Education Research in Jordan (1995) found that essay tests were poorly constructed.

Teachers preferred to use essay questions because they are less time-consuming to construct. They have a good effect on students' learning and students do not have to memorize facts, but try to get a broad understanding of complex ideas, to see relationships, etc.

Table 5.2 Discussion of the findings of data analysis related to the essay questions. (Teachers and the researcher)

The Essay Questions.	Teachers' Mean	Researcher's Mean
1. Questions are phrased so that the task is clearly defined for the student.	3.48	4.33
2. The test has optional questions.	3.36	3.67
Total	3.42	4.00

5.13.5 The Short Answer Questions

The results of this domain showed that teachers and the researcher agreed that short answer questions were a main format that is used in Tawjihi tests although they gave it a different percentage. Teachers gave it 65.4% and the researcher 76.6%. Furthermore, they agreed that short questions:

1. Consist of a single word or a short phrase.
2. Blanks are arranged to make answers easy.
3. Questions don't use more than two blanks within an item.
4. Questions are phrased so there is only one possible answer.

These findings were supported by other studies; Marso & Pigge, (1988) and Abu-Taleb (1991). Such studies revealed that achievement tests used short questions. The possible explanations of the previous results could be that teachers prefer to use this format because the scoring of responses is easier and is likely to be completed with more consistency than for extended answers. Moreover, they can be completed in enough time to include several items of this type. Also, their production format allows a range of variation that probably provides a more accurate reflection of student differences in learning

5.1.3.6 The Cloze Questions

The findings of this domain showed that teachers agreed that cloze questions were high in the tests at a percentage of 80%.and this result is different from the researcher's finding that cloze questions are prevalent format in the Tawjihi tests. But both of them found that cloze questions:

1. Test items have a single correct answer.
2. Blanks are either in the middle or at the end of statements rather than at the beginning.
3. Sgnificant words are omitted from the statement.

The teachers findings were consistent with theoretical speculations of Abu-Taleb (1991) and Al-Ekbaty (1998) that cloze tests were the most frequent formats that were used by teachers. Also, Al-Omar (1989) stated that science and math teachers used cloze questions mostly.

Cloze tests were used in high percent because they are easy to construct, it simplifies the item development task and reduces the amount of time needed for item construction. It minimizes the chance of guessing the correct answer, and when the item is constructed to yield only one correct response, it is simple to make a scoring key. It measures the recall of information rather than recognition.

However, these findings were inconsistent with the findings of Al-Janazrah (1999) who found that cloze questions weren't used at all and this result agreed with the researcher's findings.

Table 5.3 Discussion of the findings of data analysis related to the cloze questions. (Teachers and the researcher)

The Cloze Questions.	Teachers' Mean	Researcher's Mean
1. Test items have a single correct answer.	3.16	4.83
2. Blanks are either in the middle or at the end of statements rather than at the beginning.	2.94	4.33
3. Sgnificant words are omitted from the statement.	2.66	4.00
Total	2.92	3.47

5.1.3.7 The True False Questions

The results of teachers and the researcher's evaluation revealed that true false questions were frequently used in Tawjihi tests and they agreed that this format has the following points of strength:

1. Statements are concise and clear.
2. Statements include a single major idea in each one.
3. Statements are arranged so that there is no discernible pattern of answers (such as T, F, T, F, T, F and T, T, F, F, T, T, F, F) for true and false statements.
4. Statements are free of double negatives.
5. Statements are constructed in a language that is at a lower level of difficulty than the text.
6. Statements are free of the words that give verbal clues to the correct answer, such as (always, may, none, never, all, usually, generally, typically, sometime).
7. True statements are about the same length as false statements.

Also, they agreed that true false identification is never placed before the statements. The preceding findings disagree with Al-Janazrah (1999) and Al-Omar (1989) findings that the true false questions contain the words that give verbal clues to the correct answer, such as (always, may be, none, never, all, usually, generally, typically, sometime).

The reason behind such result is that true false questions are particularly useful for factual information that is central to understanding and they can be constructed quickly, with less attention to the creation of distractors that have an equivalent plausibility as correct answers.

5.1.3.8. The Multiple Choice Questions

The results of this domain showed that teachers found out that the multiple choice questions match the norms of constructing good multiple choice questions at a percent of 64% and this is a medium value, while the researcher found that multiple choice questions match these norms at a percent of 50% and this is a low value. Although they have this disagreement, they agreed that multiple choice questions match the following norms:

1. The stem is written in simple and understandable language.
2. Questions avoid using all-of-the-above and none-of-the-above distractors
3. Statements are arranged so that there is no pattern of answers (such as A, B, C, A, B, C and C, B, C, C, B, C).
4. Test items use from three to four distractors.
5. Statements are free of double negatives.
6. There is only one correct or best distractor.
7. All distractors are approximately homogeneous in content, form, and grammatical structure.
8. Questions avoid using all-of-the-above and none-of-the-above distractors.
9. The language of the stem and response distractors is as simple as possible to avoid skill overlap.

10. Distractors are free of the words that give verbal clues to the correct answer, such as (always, may, none, never, all, usually, generally, typically, sometime).

Marso & Pigge (1988) agreed with the teachers' findings that teachers used multiple choice widely. This may be due to that they are flexible and adaptable to all types and levels of knowledge, capable of generating many items, easy to score, and they have the potential of generating reliable results. In contrast, the researcher's outcomes supported by other studies Victor & others (1972), Newman (1981), Al-Ekbaty (1998), Talmir (1991) and Al Al-Janazrah (1999) who revealed that teachers have problems in constructing multiple choice questions especially using poor distractors, inaccurate and inappropriate information, specific determiners such as (always, may, none, never, all,...), negative or positive statements, keying words like all of the above or none of the above or (A+C) that led to guess the correct answer. This may be due to that teachers aren't skillful enough to acquire the facility in item writing and they may fail to find three to four related and suitable distractors.

Table 5.4 Discussion of the findings of data analysis related to the multiple choice questions. (Teachers and the researcher)

The Multiple Choice Questions.	Teacher's Mean	Researcher's Mean
1. Distractors are free of the words that give verbal clues to the correct answer, such as (always, may, none, never, all, usually, generally, typically, sometime).	3.50	5.00
2. There is only one correct or best distractor.	3.12	5.00
3. The language of the stem and response distractors is as simple as possible to avoid skill overlap.	3.32	4.67
4. The stem is written in simple and understandable language.	3.24	4.33
5. Distractors are free of double negatives.	3.20	3.67
6. Questions avoid making the correct answer markedly longer or shorter than the other distractors.	3.18	3.67
7. Questions avoid using all-of-the-above and none-of-the-above distractors.	3.12	3.67
8. Statements are arranged so that there is no pattern of answers (such as A,B,C ,A,B,C and C,B,C,C,BC) .	3.12	2.67
9. Test items use from three to four distractors.	3.08	2.67
10. All distractors are approximately homogeneous in content, form, and grammatical structure.	2.80	2.50
Total	3.20	3.78

5.1.4 Discussion of the findings of data analysis related to the sequence of domains in the questionnaire

The findings that appeared in this domain showed that teachers and the researcher disagree in the sequence of the questionnaire domains; the teachers found out that instructions were the most prevalent domain at percent of 74.2% and this result was supported by Marshall (1967), however the researcher found out that essay questions were the most prevalent domain at a percent of 80% and many studies agreed this finding in Marshall & Hales (1971), William and others (1984), Al-

Omar (1989), Ekbaty (1998) Abu-Taleb (1991) and Al-Janazrah (1999) studies. Although they have this disagreement, they agreed that the content validity was the least prevalent domain in the Tawjihi tests and this result is consistent with the results of The National Research Council (2002), Fleming and Chambers (1983), Johnstone(2003),Kirby & Oescher (1987), Al-Aga (1994), Al-Ekbaty (1998), Al-Janazrah (1999), Thompson (2002) and Haladyna (2002) studies.

Table 5.5 Discussion of the findings of data analysis related to the sequence of domains in the questionnaire. (Teachers and the researcher)

Domains	Teacher's Mean	Researchers' Mean
1.Essay Questions	3.72	4.00
2.Face Validity	3.55	3.93
3.Short/ Answer Questions	3.42	3.83
4.Multiple-Choice Questions	3.27	3.78
5.True/ False Questions	3.20	3.67
6.Instructions	3.06	3.62
7.Cloze Questions	2.92	3.47
8.Content Validity	2.87	2.58
Total	3.23	3.47

5.1.5 Discussion of the findings of data analysis related to the effects of the independent variables (gender, qualification, and experience) on the evaluation of English language teachers of Tawjihi English tests based the norms of the construction and publication of good achievement tests on each domain in the questionnaire.

This study aimed at the evaluation of the Tawjihi ESL tests according to the norms of construction and publication of good achievement tests by Tawjihi English language teachers. It was also an attempt to study the effect of each one of the independent variables (gender, qualification, and experience).

5.1.5.1 Discussion of the findings of data analysis related to the effects of the independent variable (gender) on the evaluation of English language teachers of Tawjihi English tests based the norms of the construction and publication of good achievement tests on each domain in the questionnaire.

The results indicated that there is no difference in the ratings of the English teachers due to gender. Both of them agreed that Tawjihi tests match the norms of the construction and publication of good achievement tests at a percent of 64%. Also they agreed that instruction was the most frequent domain in the tests, and the content validity domain was the least frequent domain in the Twajih tests. The researcher didn't find any study related to the same topic but she found that there is no significant difference due to the gender of teachers in constructing the achievement tests which supported in Garadat (1988) and Daniel & king (1998) studies. However, Al-Janazrah (1999) stated that Male teachers holding educational

qualifications were better in the dimensions of writing test constructions and the test publication dimension; whereas female teachers without educational qualifications were better in writing test items in general and items with different types in particular.

5.1.5.2 Discussion of the findings of data analysis related to the effects of the independent variable (experience) on the evaluation of English language teachers of Tawjihi English tests based the norms of the construction and publication of good achievement tests on each domain in the questionnaire.

The results showed that there is a difference in the ratings of the English teachers due to experience as the following:

1. Teachers who have an experience less than five years in teaching Tawjihi students found that Tawjihi tests reflected these norms at a percent of 70.6% and this value is high.
2. Teachers who have an experience in teaching Tawjihi students more than five years, found that Tawjihi tests reflected these norms at a percent of 63% and this is a medium value.

The researcher found no relevant studies related to this topic, but she found that Marshall (1967), Garadat (1988), Nair-Venugopal (1991), Boothroyd (1992), Hynie, (1992), the supervisors committee study in Jordan(1995), Al-Omarie (1997), Daniel & king (1998), Al-Janazrah (1999) and Ediger, (2001) all agreed that teachers have no experience in constructing their tests appropriately and they aren't acquainted with the norms of constructing good tests. So teachers have to attend training sessions in order to increase their experience in constructing their tests. These results disagreed with Newman (1981), Dereshiwsky (1993) and the Ministry of Education in Jordan (1995) who held a number of workshops and found out that teachers have a good knowledge in constructing their tests.

5.1.5.3 Discussion of the findings of data analysis related to the effects of the independent variable (qualification) on the evaluation of English language teachers of Tawjihi English tests based the norms of the construction and publication of good achievement tests on each domain in the questionnaire.

The findings of this domain showed that there is no difference in the ratings of the English teachers due to qualification. Newman (1981), Garadat (1988) and Hynie, (1992) agreed that there is no role due to the qualification in constructing achievement tests.

To conclude this chapter, the results of the statistical analysis for both teachers and the researcher have shown that Tawjihi English tests are presented the content of curriculum in different questions formats at a medium level. In addition to that, both teachers and researcher agreed that the content of the test wasn't sufficient evidence in Tawjihi tests. Also, the results revealed that instructions and the face validity were presented at high level in Tawjihi tests. Moreover, the essay questions were the most frequently format used in the tests. The short answer questions, multiple choice questions and cloze questions were a main format used in Tawjihi

tests, although the cloze questions were the least frequently format used. The findings of the study showed that matching questions were never used in Tawjihi tests. In addition to that, speaking and listening skills weren't evident at all in Tawjihi tests. Furthermore, the results indicated that there is no significant difference in the ratings of the English teachers due to gender and qualification they agreed that Tawjihi tests match the norms of the norms of the construction and publication of good achievement tests.

5.2 Recommendations of the Study

Based on the findings of the study, the researcher recommended the following:

1. Attention should be paid to assessment and evaluation courses especially in the Tawjihi English tests construction and publication.
2. The Directorate General of Assessment, Evaluation & Examinations should write clear instructions which could be understood by large number of students.
3. The Directorate General of Assessment, Evaluation & Examinations should use different formats of questions especially the matching questions.
4. The Directorate General of Assessment, Evaluation & Examinations should concentrate more on higher cognitive levels.
5. The Directorate General of Assessment, Evaluation & Examinations should include the four skills of the English language appropriately especially listening and speaking.
6. The Directorate General of Assessment, Evaluation & Examinations have to organize training programs for pre and in-service teachers in assessment and evaluation.
7. Researchers in other fields have to carry out similar researches on different Tawjihi tests.
8. Researchers in other fields have to carry out researches evaluating the teachers-made tests.

List of References

- Abbott, M. (2006):** ESL Reading Strategies: Differences in Arabic and Mandarin Speaker Test Performance, *Language Learning*, p633-670. ERIC (<http://www.eric.ed.gov>). 17/11/2006.
- Alderson, C. (2000):** *Assessing Reading*, Cambridge University Press, Britain.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999):** Washington, DC : American Psychological Association. 17/11/2006.
- American Association of Colleges for Teacher Education. (2000):** Counterpoint article: *USA Today* January 31, We need time to do the job right, www.aacte.org/governmental_relations/time_do_job_right.htm. 7/11/2006.
- Bachman, L.(1990):** *Fundamental considerations in Language Testing*, Oxford University Press, Oxford.
- Baker, E. (2002):** Visions of test results dance in their heads. ERIC (<http://www.eric.ed.gov>) 17/11/2006.
- Beaton, A. (1992):** Considerations for National Examinations. ERIC (<http://www.eric.ed.gov>). 1/12/2006.
- Black, P. and Wiliam, D. (1998):** Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-148. ERIC (<http://www.eric.ed.gov>). 17/11/2006.
- Blanche, P. (1988):** Self-Assessment of Foreign Language Skills (19) 1 pp. 75-93. ERIC (<http://www.eric.ed.gov>). 17/11/2006.
- Bloom, B.S. (1956):** *Taxonomy of Educational Objectives*, New York, USA.
- Boothroyd, R. (1992):** What Do Teachers Know about Measurement and How Did They Find Out? ERIC (<http://www.eric.ed.gov>). 1/12/2006.
- Bordonaro, K. (2006):** *Journal of Academic Librarianship*, p518-526. ERIC (<http://www.eric.ed.gov>). 1/12/2006.
- Childs, R. (1998):** *Constructing Classroom Achievement Tests*. ERIC (<http://www.eric.ed.gov>). 1/12/2006.
- Cohen, M. (2006):** Reading Disabilities among Hebrew-Speaking Children in Upper Elementary Grades: The Role of Phonological and Nonphonological Language Skills Reading and Writing. ERIC (<http://www.eric.ed.gov>). 13/12/2006.

Coulson, J. & Silberman,H. (1960): Effect of three Variables in Teaching Machine. Journal of Education Psychology.

Cross, T.L. (1990): Testing in the College Classroom. Paper presented at the Annual Meeting of American Educational Research Association, Boston, April 1990.

Daniel,L and King,D. (1998): Knowledge and use of Testing and Measurement Literacy of Elementary and Secondary Teachers. The Journal of Educational Research, 91, 330-344.

Dolan, R. P. (2005): Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. Journal of Technology, Learning, and Assessment, 3(7). (<http://www.bc.edu/research/intasc/jtla/journal/v3n7.shtml>). 26/11/2006.

Dereshiwsky, M.(1993): When "Do It Yourself" Does It Best: The Power of Teacher-Made Surveys and Tests.ERIC (<http://www.eric.ed.gov>). 1/12/2006.

Ediger, M. (2001): Teacher Involvement To Evaluate Science Achievement, ERIC (<http://www.eric.ed.gov>). 1/12/2006.

Erkaya, O. (2005): Benefits of Using Short Stories in the EFL Context Online Submission, Asian EFL Journal. ERIC (<http://www.eric.ed.gov>). 13/12/2006.

Fleming, M.&Chambers,B.(1983): Teacher-made Tests: Windows on the Classroom,W. E. Hathaway, Testing in the schools, San Francisco: Jossey-Bass. ERIC (<http://www.eric.ed.gov>). 17/11/2006.

Fradd, S. and S. Hudelson. (1995): Alternative Assessment: A Process that Promotes Collaboration and Reflection. TESOL Journal, 5, 1, p. 5.

Friel, M.(1989): Reading Technical Texts: A Class Test, English Teaching Forum, 27, 1, pp. 32-33. ERIC (<http://www.eric.ed.gov>). 6/11/2006.

Genesee, F and Upshur, J. (1998): Classroom-Based Evaluation Second Language Education. Cambridge University Press, Britain.

Gentry,D. (1989): Teacher-Made Test Construction. Paper Presented at the Annual Meeting of the Mid- South Educational Research Association (Little Rock, AR, Novemeber 8-10, 1989) ERIC (<http://www.eric.ed.gov>). 2/2/2007.

Gronlund, N. and Linn, R. (1990): Measurement and Evaluation in Teaching. (6th ed.) New York: Macmillan.

Haladyna, M. (2002): Large-scale assessment programs for all students: Validity, technical adequacy, and implementation, (pp. 213-231). ERIC (<http://www.eric.ed.gov>). 26/11/2006.

Heaton J.(1997): Classroom Testing. Longman, England.

Hughes, A.(1990): Testing English for University Study, Modern English Publications and The British Council, Hong Kong.

Hughes, A.(1989):Testing For Language Teachers, Cambridge University Press, Britain.

Hynie,W. (1992): Post Hoc Analysis of Test Terms Written by Technology Education Teachers, Journal of Technology Education, 4(1) 128-141.

Johnstone, C. (2003): Improving validity of large-scale tests: Universal design and student performance, Minneapolis, MN: National Center on Educational Outcomes. (<http://education.umn.edu/NCEO/OnlinePubs/Tech44/>).26/11/2006.

Johnstone, C. (2006): Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners, Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. (<http://education.umn.edu/NCEO/OnlinePubs/Tech44/>).26/11/2006.

Karabenick, S. (2000): Impact of State Testing on Students and Teaching Practices: Much Pain, No Gain? ERIC (<http://www.eric.ed.gov>). 1/12/2006.

Kane, S. (2000): American Association of Colleges for Teacher Education. [www.aacte.org/governmental_relations/time do job right.htm](http://www.aacte.org/governmental_relations/time_do_job_right.htm) May 7, 2001.

Katz, L.(1997): A Developmental Approach to Assessment of Young Children. ERIC (<http://www.eric.ed.gov>). 6/11/2006.

Kirby, P. and.Oescher, J. (1987): Testing for Critical Thinking: Improving Test Development and Evaluation Skills of Classroom Teachers. ERIC (<http://www.eric.ed.gov>). 1/12/2006.

Kirby, P. (1988): Reflective Teaching and Teacher Effectiveness: Measurement Considerations. ERIC (<http://www.eric.ed.gov>). 1/12/2006.

Kopriva, T (2001): Assessment using think aloud methods. ERIC (<http://www.eric.ed.gov>). 26/11/2006.

Kjellin, S. (2006): Children's Engagement in Different Classroom Activities European Journal of Special Needs Education, p285-300. ERIC (<http://www.eric.ed.gov>). 23/1/2007.

- Lerkkanen, M.(2004):** The Developmental Dynamics of Literacy Skills during the First Grade, Educational Psychology, p793-810. ERIC (<http://www.eric.ed.gov>). 6/11/2006
- Levacic, R.(2006):** Evaluating the Effectiveness of Specialist Schools in England School Effectiveness and School Improvement, p229-254. ERIC (<http://www.eric.ed.gov>). 22/2/2007.
- Madsen, H. (1982):** Retrospective Evaluation of Testing in ESL Content and Skills Courses. ERIC (<http://www.eric.ed.gov>). 17/10/2006.
- Marshall,J. and Hales, L. (1971):** Classroom Test Construction, Addison-Wesley Publishing Company, United States of America.
- Marso,R and Pigge,F. (1988):** An Analysis of Teacher – Made Tests: Testing Practices Cognitive Demands and item Construction errors. ERIC (<http://www.eric.ed.gov>). 1/12/2006.
- Mayerhof, E.(1992):** Communication Dynamics as Test Anxiety Therapy. English Teaching Forum, 30, 1, pp. 45-47. ERIC (<http://www.eric.ed.gov>). 2/2/2007.
- McKeachie, W. (1986):** Teaching Tips, Lexington. ERIC (<http://www.eric.ed.gov>). 6/11/2006.
- McNamara, T. (2000):** Language Testing, Oxford University Press, Oxford.
- Milanovic, M. (1999):** Studies In Language Testing 3. Cambridge University Press, Britain.
- Mills, R. (1998):** Development of Program and Individual Student Evaluation Models for Foreign Language in the Elementary School. ERIC (<http://www.eric.ed.gov>). 26/10/2006.
- Nair-Venugopal, S. (1991):** Continuous Assessment in the Oral Communication Class: Teacher Constructed Test. ERIC (<http://www.eric.ed.gov>). 1/12/2006.
- National Research Council. (2002):** Performance assessments for adult education: Exploring the measurement issues: Report of a workshop. Washington, DC: National Academy Press,USA.
- Newman, D. (1981):** Teacher Competency in Classroom Testing Practices . Dissertation Abstract,(42:3 P(1111-A). ERIC (<http://www.eric.ed.gov>).
- Nicosia, G. (2005):** Developing an Online Writing Intensive Course: Will It Work for Public Speaking, International Journal of Instructional Media, p163. ERIC (<http://www.eric.ed.gov>).

- Nitko, A. (2001):** Educational assessment of students (3rd Ed.). ERIC (<http://www.eric.ed.gov>). 17/11/2006.
- O'Neil, T. (1992):** Putting Performance Assessment to the Test, Educational Leadership, 49, 8, pp. 14-19. ERIC (<http://www.eric.ed.gov>). 22/2/2007.
- O'Neil, T.(2004):** Evaluating the Consistency of Test Content Across Two Successive Administrations of a State-Mandated Science Assessment. ERIC (<http://www.eric.ed.gov>). 22/2/2007.
- Pierce, L. and O'Malley, M. (1992):** Performance and Portfolio Assessment for Language Minority Students, Washington ,National Clearinghouse for Bilingual Education. 2/2/2007.
- Pigge, F. and Marso, R. (1993):** A Summary of Published Research: Classroom Teachers' and Educators' Attitudes toward and Support of Teacher-Made Testing. ERIC (<http://www.eric.ed.gov>). 6/11/2006.
- Popham, W. (2005):** For Assessment Eductopia, George Lucas Educational Foundation. ERIC (<http://www.eric.ed.gov>).28/11/2006.
- Popham, W. (2006):** Assessment for learning, Educational Leadership, 63(5), 82–83. ERIC (<http://www.eric.ed.gov>). 17/11/2006.
- Posner, D.(2004):** What's Wrong with Teaching to the Test? ERIC (<http://www.eric.ed.gov>). 22/2/2007.
- Rea-Dickins, P. and Rixon, S.(1997):** The Assessment of Young Learners of English as a Foreign Language, In Encyclopedia of language and education, language testing and assessment, Kluwer Academic Publishers. Netherlands.
- Rodriguez, M. (2002):** Choosing an item format. ERIC (<http://www.eric.ed.gov>). 26/11/2006.
- Shepard, L. (2000):** Evaluating Test Validity, Darling-Hammond, Review of research in education, Washington, 19, pp. 405-450. American Educational Research Association.
- Shohamy, E. (1995):** Performance Assessment in Language Testing, Annual Review of Applied Linguistics, pp. 188-211. ERIC (<http://www.eric.ed.gov>).
- South Dakota Department of Education, 2004):** South Dakota's State Assessment System ERIC (<http://www.eric.ed.gov>). 22/2/2007.
- Talmir, N. (1991):** Multiple Choice Items: How to Gain the Most out of Them, Biochemical Education, ERIC (<http://www.eric.ed.gov>). 1/12/2006.

Thompson, S. (2002): Universal design applied to large scale assessments, Minneapolis, MN: National Center on Educational Outcomes. (<http://education.umn.edu/NCEO/OnlinePubs/Tech44/>). 26/11/2006.

Tindal,G. and Marston, D. (1990): Classroom-Based Assessment. Merrill Publishing Company. United States of America.

Victor, H. (1972): Introduction to Educational Measurement. Michigan State University. Houghton Mifflin Company, USA.

Wang, C. (2000): How to grade essay examinations, Performance Improvement, 39(1), 12-15. ERIC (<http://www.eric.ed.gov>). 17/11/2006

Wiggins, G.(1998): A True test: Toward more Authentic and Equitable Assessment. Phi Delta Kappa, 70, pp. 703-713. ERIC (<http://www.eric.ed.gov>). 17/11/2006

المراجع العربية

أبو طالب، ج. (1991): أساليب و ممارسات تقييم الأداء لمبحث العلوم العامة في الصفوف الرابع و الخامس و السادس الأساسية. الجامعة الأردنية، الأردن. (رسالة ماجستير غير منشورة).

الأغا، أ. (1994): تحليل أسئلة الامتحانات النهائية لمقرر العلوم للصف الثالث الإعدادي بمدارس. قطاع غزة. جامعة الأزهر، فلسطين.

جردات، م. (1988): مدى معرفة و ممارسة معلمي العلوم للمرحلة الإعدادية لكفايات بناء الاختبارات المدرسية. جامعة اليرموك، الأردن.

القباطي، م. (1998): أساليب و ممارسات التقييم المدرسي في مديرية الحديدة و مستوى جودة الاختبارات التحصيلية. الجامعة الأردنية، الأردن. (رسالة ماجستير غير منشورة).

العمري، ح. (1997): تقويم الاختبارات المدرسية في ضوء معايير تطوير الاختبارات (تخطيطها، و إخراجها، و تطبيقها، و تصحيحها). جامعة اليرموك، الأردن. (رسالة ماجستير غير منشورة).

العمر، س. (1989): أساليب و ممارسات التقييم لمدرسي مديرية عمان الكبرى و مستوى جودة اختباراتهم التحصيلية. الجامعة الأردنية، الأردن. (رسالة ماجستير غير منشورة).

الجنازرة، أ. (1999): تقويم الاختبارات التحصيلية لمادة الكيمياء للصف العاشر وفق معايير تصميم و إخراج الاختبار التحصيلي الجيد. جامعة القدس، فلسطين. (رسالة ماجستير غير منشورة).

وزارة التربية و التعليم. (1995): قسم القياس و التقويم. (http://www.moe.gov.sa/ishraf/MOE/Qunfozah.htm)30/11/2006

Appendix A

The Questionnaire before the Judgment.

The Required Norms of Evaluation the ESL Tawjihi Tests Based on Norms of The Construction and Publication of Good Achievement Tests.	Poor	Fair	Good	Very Good	Excellent
1. The Instructions.					
51. Test contains general instructions.					
52. Test contains suitable space between the instructions and the questions.					
53. Instructions contain allotted time for the test.					
54. Instructions contain the number of the questions.					
55. Test is allotted marks for each question.					
56. Instructions are simple, clear, and definite.					
57. Each question has instructions.					
2. The Content Validity.					
8. Content of the questions reflects the textbook objectives.					
9. Content of questions assess different cognitive levels. (Bloom's Taxonomy)					
10. Questions are presented in different formats (essay questions, matching, true or false, multiple choices, completion).					
11. Reading skill is adequately assessed.					
12. Listening skill is adequately assessed.					
13. Speaking skill is adequately assessed.					
14. Writing skill is adequately assessed.					
15. Literature is adequately assessed.					
3. The Face Validity.					
16. Test is free of spelling, printing, and language mistakes.					
17. There is a suitable space between each question and the following one.					
18. Charts, tables, or figures are printed clearly and labeled correctly.					
19. Copies of the test are clear.					
20. Questions are sequenced from the beginning till the end of the test.					
21. There is suitable space for answers.					
22. Questions are free of ambiguities.					
23. Questions are proceeding from easy to more difficult items.					
4. The Essay Questions.					
24. Questions are phrased so that the task is clearly defined for the student.					
25. Test have optional questions.					
26. Questions indicate the number of the points to be earned for correct response.					
5. The Short Answer Questions.					
27. Consists of a single word or short phrase.					
28. Blanks are arranged to make answers easy.					
29. Questions don't use more than two blanks within an item.					
30. Question is phrased so there is only one possible answer.					

6. The Cloze Questions.					
31. Blanks are at the end of a statement rather than at the beginning.					
32. Important words are omitted from the statement.					
33. Test items have a single correct answer.					
7. The True and False Questions.					
34. Statements are concise and clear.					
35. Statements are free of specific determiners such as (always, may be, none, never, all, usually, generally, typically, sometimes).					
36. Statements are arranged so that there is no pattern of answers (such as T, F, T, F, T, F and T, T, F, F, T, T, F, F) for true and false statements.					
37. Statements are free of double negatives.					
38. Statements include single major idea in each one.					
39. Statements are constructed in language at a lower level of difficulty than the text.					
40. True, false identification are placed before the statements.					
41. True statements are about the same length as false statements.					
8. The Matching Questions.					
42. Instructions are clearly stated the basis for matching two columns, A: the column of premises, B: the column of responses.					
43. Material in the premises and responses are clearly related to each other.					
44. Premises and responses are short ranging from 5-6 premises and responses.					
45. Premises and responses are clearly and easy to read.					
46. The number of responses is more than number of premises.					
47. Premises are formed in numbered column at the left and the response choices are in a lettered column of the right.					
48. All items are on a single page.					
49. The list of responses is arranged in alphabetical or numerical order, in order to save reading time.					
9. The Multiple Choice Questions.					
50. There is only one correct or best distractor.					
51. All distractors are approximately homogeneous in content, form, and grammatical structure.					
52. Questions avoid using all-of-the-above and none-of-the-above distractors.					
53. Questions avoid making the correct answer markedly longer or shorter than the other distractors.					
54. All response distractors are in the same length.					
55. The language of the stem and response distractors is as simple as possible to avoid skill contamination.					
56. Distractors are free of the words that give verbal clues to the correct answer, such as (always, maybe, none, never, all, usually, generally, typically, sometime).					
57. The stem is written in simple, and understandable language					

58. Distractors are free of double negatives.					
59. Test items use from three to four distractors.					
60. The placement of the correct answer is on a random basis.					

Teacher's General Comments:

Appendix B

The list of Referees

- Dr. Adnan Shehadeh, Polytechnic University.
- Dr. Ahmad Atawneh Hebron University
- Dr. Hana Tushiya Bethlehem University.
- Dr. Hazem Eid Bader Hebron University
- Dr. Jeanne Kattan Bethlehem University
- Dr. Mohd Farrah Hebron University
- Dr. Nimer Abu Zahrah Hebron University.
- Dr. Raghad Dweik Hebron University
- Dr. Salah Shrouf Hebron University
- Dr. Will Edmundsun Hebron University

Appendix C
The questionnaire after the Judgment

Dear Teachers,

I would like to present this questionnaire to you hoping that you will fill it objectively and seriously as you are known for being so. It is about:

Evaluation of the Tawjihi English Tests Based on Norms of the Construction and Publication of Good Achievement Tests.

Thus, the researcher hopes you would respond to all items precisely and frankly assuring you that your responses will be confidential and for academic purposes. After filling out some general information about you, you are requested to read each item carefully and write (X) in the square of the degree. The researcher greatly appreciates your help in answering the questionnaire faithfully.

With Thanks

Background Information:

Put (x) in the suitable place.

1. Gender: Male
 Female

2. Experience: less than 5 years
 5-10 years
 More than 10 years.

3. Qualifications: Diploma
 B.A
 MA and more

The Required Norms of Evaluating the ESL Tawjihi Tests Based on Norms of The Construction and Publication of Good Achievement Tests.	Poor	Fair	Good	Very Good	Excellent
1. The Instructions.					
58. The test contains general instructions.					
59. The instructions are simple, clear, and definite.					
60. The test contains suitable space between the instructions and the questions.					
61. The instructions contain allotted time for the test.					
62. The instructions contain the number of the questions.					
63. The instructions contain marks for each question.					
64. Each question is provided by its own instructions.					
65. The questions present the number of words and paragraphs needed for the answer.					
2. The Content Validity.					
9. Content of the questions reflects the textbook objectives.					
10. Content of questions assesses different cognitive levels. (Bloom's Taxonomy)					
11. Questions are free of ambiguities.					
12. Questions are proceeding from easy to more difficult items.					
13. Questions are presented in different formats (essay questions, matching, true or false, multiple choices, completion).					
14. Reading skill is adequately assessed.					
15. Listening skill is adequately assessed.					
16. Speaking skill is adequately assessed.					
17. Writing skill is adequately assessed.					
18. Literature is adequately assessed.					
3. The Face Validity.					
19. The test is free of spelling, printing, and language mistakes.					
20. There is suitable space between each question and the following one.					
21. Charts, tables, or figures are printed clearly and labeled correctly.					
22. Copies of the test are clear.					
23. Questions are sequenced from the beginning till the end of the test.					
4. The Essay Questions.					
24. Questions are phrased so that the task is clearly defined for the student.					
25. The test has optional questions.					
5. The Short Answer Questions.					
26. Consists of a single word or short phrase.					
27. Blanks are arranged to make answers easy.					
28. Questions don't use more than two blanks within an item.					

29. Questions are phrased so there is only one possible answer.					
6. The Cloze Questions.					
30. Blanks are either in the middle or at the end of statements rather than at the beginning.					
31. Significant words are omitted from the statement.					
32. Test items have a single correct answer.					
7. The True and False Questions.					
33. Statements are concise and clear.					
34. Statements are free of specific determiners such as (always, may be, none, never, all, usually, generally, typically, sometimes).					
35. Statements are arranged so that there is no discernible pattern of answers (such as T, F, T, F, T, F and T, T, F, F, T, T, F, F) for true and false statements.					
36. Statements are free of double negatives.					
37. Statements include a single major idea in each one.					
38. Statements are constructed in a language that is at a lower level of difficulty than the text.					
39. True, false identification are placed before the statements.					
40. True statements are about the same length as false statements.					
8. The Multiple Choice Questions.					
41. There is only one correct or best distractor.					
42. All distractors are approximately homogeneous in content, form, and grammatical structure.					
43. Questions avoid using all-of-the-above and none-of-the-above distractors.					
44. Questions avoid making the correct answer markedly longer or shorter than the other distractors.					
45. The language of the stem and response distractors is as simple as possible to avoid skill overlap.					
46. Distractors are free of the words that give verbal clues to the correct answer, such as (always, may be, none, never, all, usually, generally, typically, sometime).					
47. The stem is written in simple, and understandable language.					
48. Distractors are free of double negatives.					
49. Test items use from three to four distractors.					
50. Statements are arranged so that there is no pattern of answers (such as A,B,C ,A,B,C and C,B,C,C,BC) .					

Teacher's general comments:

Appendix D
Recommendation letter.

Appendix E
The list of Tawjihi Tests (2000-2006)