

A study of Tests Written by Teachers in Al- Ehsa' Directorate of Education for the Third Intermediate Students

Dr. Emad Salih Al-Kulaib , Dr. Mwaffag Mohammed Shatanawi

Abstract

This study aimed at investigating whether the tests constructed by the teachers of the third intermediate in Al- Ehsa' Directorate of Education cover the language components enclosed with the teacher's book in terms of structure, function, vocabulary, attitudes and values. The study attempted to answer the following questions:

- 1. To what extent do EFL teachers cover the language components in writing their test items?**
- 2. Do the EFL teachers' tests vary with respect to sex and experience variables?**

The population of the study consisted of all the teachers of the third intermediate in Al- Ehsa' Directorate of Education during the first semester of the academic year 2013/2014. The teachers had the same social background but differ in experience. The sample of the study consisted of forty tests written by forty teachers who are B.A holders (n=20 males and 20 females) in forty schools in Ehsa' Directorate of Education. They all teach the third intermediate grade. To achieve the objectives of this study, the researcher developed a model (Appendix A) to measure the four language components included in the tests (function, structure, vocabulary and attitudes and values) available in the teacher's book. The researcher used means and percentages to answer the first question and used ANOVA test to answer the second question.

The findings of the study were as follows:

- 1. Teachers cover only 16% of the language functions, 34% of the language structures, 16% of the values and attitudes and 37% of the vocabulary items. These results compared with those of the supervisors', are below the average of acceptability.**
- 2. Teachers constructed more comprehensive test items on structures and vocabulary than the functions, attitudes and values.**

3. *There are statistically significant differences among teachers due to experience in favor of long experience.*
4. *There are statistically significant differences among teachers due to sex in favor of male teachers.*
5. *There are statistically significant differences among female teachers due to experience in favor of long experience.*
6. *There are no statistically significant differences among male teachers due to experience.*

The study recommended the following:

1. *EFL teachers are recommended to vary their tests items to meet a balanced weight of language components.*
2. *EFL teachers are recommended to make a comprehensive content analysis before constructing their tests items.*
3. *Supervisors are advised to share their teachers in writing tests items.*
4. *The Ministry of Education is advised to build sample tests and provide teachers with them.*
5. *The Ministry of Education is recommended to start a training programme the aim of which is to train teachers how to build a comprehensive, authentic and valid test.*
6. *Further studies should be conducted on testing speaking, and listening in schools all over the country.*

Introduction

The majority of EFL teachers face an outstanding difficulty when they attempt to construct a test. One of their major problems involves the mastery of building a comprehensive test that takes into account the necessary language components. They often write tests that neglect most of the important test items. (generalization without evidence)

Testing plays a significant role in language teaching and learning. This significance stems from the fact that it measures students' achievement and the effectiveness of teachers' methodology. Testing also reveals the weakness of students in some areas of language, which may

come from the bad construction of the test items. Accordingly, variety of language components cornering the material assigned should be taken into consideration when writing test items. Building a test with different language components is necessary to give the appropriate value for each language item. Otherwise, some language items may be neglected or given a minor weight where it should be emphasized and given a heavier one.

Tests in their both sides, written and verbal, either at school or general examination levels; represent one of the main means or tools of evaluation because their results provide us with a quantitative digital indicator about the progress achieved by the learners. They provide them with the necessary directives through the learning process starting from inspecting the facts to analyzing and evaluating what they have extracted (Hamdan, 1981).

Tests are part of the cognitive and epistemic techniques that work on activating and recalling information within the mind of the learner then to make use of it in an effective way. Furthermore, recalling the memory in its turn controls the method of information storing and thereby learning is achieved in a perfect way (Andre, 1979; Darwazeh, 1987).

Also, questions prepared in a perfect way are considered as an effective means to develop required tendencies, to create tends and to provide the student with various methods to deal with the study material (Jaber, Al-Sheikh and Zahir, 1986). Further, they lead the learners towards deep thinking and effective response in a way that leads to higher levels of achievement (Samson et al., 1981). Additionally, they reveal the degree of the teacher's mastery, familiarity and knowledge of the material which is assigned to them to teach and reveal their capability of nature thinking (Jabes et al., 1989 & Carlson, 1988).

Tests are also considered one of the means and instruments of analogy. They are the main instruments on which teachers depend when estimating the marks and the levels of students, then to shifting them to higher classes. Besides, tests' results are adopted in defining the secondary education streams in Saudi Arabia. Further, they are the touchstone and criterion by which admission of various higher education students is decided as there is no other technique adopted to achieve that goal in a precise and a lustful way.

The importance of tests as an instrument of evaluation comes from the fact that it is achieved by the denotations of the goals. So, in order to make use of them as an instrument of evaluation, it is better to make an analysis of the teachers' questions on their various types: daily, monthly and quarterly (Morce et al., 1990, P.325).

Tests are as old as history. Socrates had used them in teaching his students, and educators up to our present day have followed him within a pattern of spontaneity and simplicity. But tests in the second half of the present century began to take a significant form and an organized shape especially after that loud outcry of Taylor in this regard as many educators have taken the responsibility to frame and define the test, then to put for rules, limits and specifications which control it. Many definitions of the test have come in the educational literature of which is Cornback's definition (1970, P.26). He defined it as an organized procedure of an observation of an individual's behavior. Also he described it by devices with numerical measurements, categorical organization, stages or estimations. Gorow (1970, P. 28) defined it as a group of questions to be answered with the purpose of verifying the range of achieving the goal or goals that had been formed. Good (1973, P. 594) in the Dictionary of Education has mentioned three definitions of tests:

- 1. A test is a group of questions or tasks to which students are requested to respond. It aims at producing numerical representations of one of the pupil's characteristics then to plan for measuring it.*
- 2. A test is an organizational procedure in which a comparison between the behavior of two or more individuals is made.*
- 3. A test is a procedure or a touchstone which is used to specify the truthfulness or truthlessness of the presented hypothesis.*

Moartwzed, (1977, P. 1) defined the test as a certain type of measurement that includes a group of items and a group of directives which explain to students how to respond to the items. Aiken (1977, P. 332) has asserted in his definition of the test that it is an instrument used to evaluate the behavior or performance of an individual. Sax (1980, P. 13) referred to a test as a task or a group of tasks used to obtain

organized remarks which are supposed to represent educational and psychological characteristics or qualities. Sa'adeh (1984, P. 526) confirmed in his definition of the test that it is an organizational action in which the students' behavior is noticed and the range of their achievement of the drafted goals is confirmed by drafting a group of items or questions to be answered along with describing these answers through numerical measurements. Nashawati (1984, P. 601) considered a test as a certain type of measuring instruments and techniques which involves questions, statements and teaching tasks which are chosen and worded in a certain methodical way so that they provide, upon answering them by the students, a digital value of his cognitive characteristics such as achievement, cleverness, invention, or non cognitive ones such as social background, tendencies and values.

From the above definitions, it can be noticed that educators have agreed among themselves on the definition of the test as they had confirmed that a test represents that organized procedure which is concerned with the evaluation of the behavior of the learners through the group of stimulants (questions), related to a certain subject, which are to be answered in order to verify the range of the student's achievement of the defined goals of the determined educational material. One of the well-known facts is that the good question stems from the good goal and occurs with it as it is the instrument from which decision is sought interrelation to the extent of achieving the goal within its levels. The good question does not only convey the required goal, but it also presents it in a clear way. In other words, the linguistic arrangement of the question affects its clarity and is reflected on the way by which the purpose of the question is conveyed. The extent of the question's conveyance of its purpose and its good wording up performs the type of the answers received upon asking it. This requires the teachers remember each one of their students when they start to think about drafting the questions because it helps them to choose the language that suits the capacity and the mental levels of their students so that the questions are understandable, clear and specific.

Peter (1991) discussed authenticity in foreign language testing. "He stated that in foreign language testing, as in all testing; validity is the primary criterion for test quality. However plausible the concept to validity, in practice it is not always easy to arrive at congruence between

the test situation and the real life situation the learner is expected to master. Some language educators make authenticity a major criterion of test quality. However, complete congruence of test and real life situation is impossible, and there are other considerations than authenticity in testing. A language test, as a social event, is essentially different from any other social events in which the learner will need to use language. The solution is to find a reasonable balance between authenticity and abstraction in tests. Pragmatics, with its analysis of speech acts and their characteristics, can be helpful in finding the right degree of abstraction for testing. Examples of such test items include a series of sentences of which portions are illegible and the learner must supply appropriate words, or a paired or group activity in which students must elicit information from each other to complete a common task such as survey or map completion". Page number needed

Davises (1986 Page number needed) said that "the good test is an obedient servant since it follows and apes the teaching". Progress in language testing looks at movement in the field since 1980, based on the themes and content of national and international conferences; trends in test content, method, and analysis; and work on the nature of proficiency and of language learning. It is proposed that movement evidenced by conferences is largely side way and back word; that while improvements have been made in test content, method and analysis, there is little evidence that these improvements represent real advancements; and the research on the nature of proficiency and of language learning is still in its early stage. Four main reasons are given for the lack of progress: the relative youth of the discipline, dearth of replication, team work, and agenda in research; inadequacy of funding; and lack of a coherent frame-work or model. Areas in which attention will be important in the next decade or outlined, including: research on language learning; the wash back effect of testing; validity of test content; knowledge of the structure of language proficiency; computer based language testing and the impact of technology on testing; learner-central testing; the role judgment in language testing; and traditional concerns about test validity and reliability".

Generally speaking, the proper relationship between teaching and testing is surely that of partnership. It is true that there may be occasions when teaching is good and appropriate and testing is not; we are then likely to suffer from harmful backwash. This would scam to be the

situation that leads Davies to confine testing to the role of servant of teaching. But equally there may be occasions when teaching is poor or inappropriate and when testing is able to exert beneficial influence, we cannot expect testing only to follow teaching. What should demand of it, however is that it should be supportive of good teaching and, where necessary, exert a corrective influence on bad teaching, if testing always had beneficial backwash on teaching, it would have a much better reputation amongst teachers.

Statement of the problem:

The majority of EFL teachers face an outstanding difficulty when they attempt to construct a test. One of their major problems involves the mastery of building a comprehensive test that takes into account the necessary language components. They often write tests that neglect most of the important test items.

Significance of the study:

The Significance of this study is that it attempted to highlight the necessity of building a test which covers the language components appropriately according to the planned material. This study is also significant as it will reveal the quality of tests written by teachers and the difficulties teachers face during writing their test items. This study is also expected to make teachers give balance weight to each language component. As well, this study is significant because school tests in all their types, which are built and prepared by teachers, play a major role in the educational evaluation process as they are essentially designed to measure the product of class learning.

This study is also significant as the results will participate in developing training programs and workshops which aim at improving the performance of English language teachers in writing and constructing their written tests. The results of this study will participate in improving students' abilities as well as highlighting the weakness in building the written test questions.

Purpose of the study:

The Purpose of the study stems from the need of overcoming the difficulties and problems teachers face in building comprehensive and satisfactory tests. This study aimed at investigating whether the tests constructed by the teachers of the third intermediate in Al-Ehsa' Directorate of Education cover the language components enclosed with

the teacher's book in terms of structure, function, vocabulary, attitudes and values.

Questions of the study:

This study attempted to answer the following questions:

- 1. To what extent do EFL teachers cover the language components in writing their test items?*
- 2. Do EFL basic stage teachers' tests vary with respect to sex and experience variables?*

Limitations of the study:

The study was limited to the four language components (structure, function, values and vocabulary). It is also limited to the forty tests of the third intermediate students in Al-Ehsa' Directorate of Education.

Definition of Terms:

The following terms will have the associated meaning whenever they appear in the study:

- 1. Language components: The four language components mentioned in the teachers' book (structure, function, attitudes, values and vocabulary).*
- 2. Structures: The structures mentioned in the Teacher's book of the third intermediate.*
- 3. Vocabulary: The vocabulary mentioned in the Teacher's book of the third intermediate.*
- 4. Attitudes and Values: The attitudes and values mentioned in the Teacher's book of the third intermediate.*
- 5. Functions: The functions mentioned in the Teacher's book of the third intermediate.*
- 6. A Comprehensive test item: A test item that covers more than 50% of the language components mentioned in the teacher's book of the third intermediate.*
- 7. An in comprehensive test item: A test item that covers less than 50% of the language components mentioned in the teacher's book of the third intermediate.*

Review of Related Literature

This chapter deals with the review of related literature. To the best of the researcher's knowledge, there have not been any previous studies on the English Language on the topic of the relation between teachers' questions and the components of the content of the teachers' guide curriculum in Saudi Arabia. Therefore, it is expected that this study will have a significant role in attracting teachers', as well as researchers', attention to the levels of written testing questions constructed by teachers to cover the content of the teachers' guide. In addition, it is hoped that this study will improve the quality of teaching the English language in Saudi Arabia as well as adding something new to the English Educational Library.

The researchers reviewed some of the studies, which, were conducted on topics related to this study. Staus (1970) mentioned that Hedges had analyzed questions of Science teachers for the secondary stage in Virginia, USA, where the total number of question items had amounted to 1400 ones. Further, it had been found that questions which measured the remembering level had occupied the first place with a high rate 78%).

Tinsley and Davis (1971) studied in Texas the relation between the classroom questions of teachers of the Social materials and the level of their questions in school tests prepared by the teachers of the eighth and the eleventh grades. The study sample consisted of 67 teachers who had been randomly chosen then randomly distributed into four groups. The first group contained 15 teachers, the second 17 teachers, the third 18 teachers, and finally, the fourth 17 teachers. The first and the third groups were requested to prepare not less than fifteen classroom questions for the eighth and the eleventh grades. Meanwhile, the same number of questions for school tests was requested from the second and the fourth groups. The two researchers sought the assistance of two specialists in analyzing the questions of all the groups according to Guilford's classification. The analysis results indicated that the questions were either classroom or written (school tests) at the evaluation then the remembering levels and showed the importance of an advance preparation of the two types of questions which had led to an increase in the level of the high cognitive questions. Not clear rephrase.

Zaki (1973) conducted a study on analyzing the Science textbooks of the first and third grades of the preparatory stage in Egypt according to

Bloom's classification of the educational goals of the cognitive domain. The result of the analysis revealed that the questions in the remembering level were 73% of the questions of the text book of the first preparatory class, while the rate of the understanding level were 26% for the same class. As for the third preparatory class 87% measured the remembering level and the rest (13%) measured the understanding level.

Billah (1974) aimed at defining the mental processes contained in the questions of high school teachers who teach the three sciences (Biology, Chemistry and Physics). The researcher adopted Bloom's classifications of the cognitive domain's goals as a criterion in analyzing the questions.

The population of the study consisted of the secondary schools in Beirut, of which the researcher randomly chose 25 schools that is (30%) of the population of the study, and from each school he randomly chose two classes resembling the students of the seventh and the tenth grades successively. This was followed by a recording of a complete unit. Then a teacher who taught this unit was requested to draft a test which needed an hour to be answered. After that, thirty three tests were collected from 18 schools. Finally, the questions of these tests were analyzed by a committee consisting of three arbitrators who have experience and knowledge of classifying educational goals.

The results of the analysis showed that the level of the cognitive questions was very high and suitable for the three materials as this level took 77.6%, 73.1%, and 63.9% for Biology, Chemistry and Physics respectively.

Furthermore, the results showed that the cognitive level at the class level was high as the questions' level were 79.2% in the seventh grade but retreated to be 65.72% in the tenth grade. Additionally, this study indicated the absence of the other three levels (analysis, synthesis and evaluation) in the teachers' questions.

Kneip & Grossman (1979) summarized some studies which reported that students' achievement was significantly and positively affected when teachers use mostly high level questions. For example, Rayan (1973) compared the effects of high to low level questions on the Social Studies' achievement of fifth and sixth grade students. Results indicated that questions which demand high cognitive levels, beyond the recall level,

were superior to the low level recall and recognition questions in producing not only the high level understanding but also producing high level of achievement.

The results of the analysis showed that the novel of the questions of the first test did not exceed the remembering level, and this was the case of the second text. But they showed that the rate of cognitive questions (remembering) were high in the third test although it had contained some questions which measure some other levels.

Azar (1980) conducted a study dealing with the Science textbooks in Iran for the secondary stage aiming at the textbook's readability, trying to share the student, the included questions, and measuring the subordinate questions. In analyzing the included and the subordinate questions, the researcher applied Bloom's classification with its well known levels.

Redfield (1981) conducted a study to examine the effect of teachers' question on student achievement. In this study, 20 studies on teachers' use of higher application questions and lower recall and recognition cognitive questions were reviewed. Higher cognitive questions require the student to manipulate information to create and supply a response; lower cognitive recall and recognition questions call for verbatim recall or recognition of factual information. Results of the studies reviewed showed that teachers' use of higher cognitive questions had a positive effect on student general achievement on the retention level of learning.

To summarize, some researchers found that high level questions (i.e., use a generalization) have a greater effect on students' achievement than low level ones (e.g., Al- Nayef, 1989; Kneip & Crossman, 1979; Redfield, 1981; Reckards & Vesta , 1974; Royer & Konold, 1984; Watts & Andreson, 1971). Others found that low level questions had a greater effect on students' achievement than high level ones (e.g., Felker & Dapra, 1975; Perkins et al., 1990; Samson et al., 1987). Few studies did not find significant differences among the different levels of questions: remember an instance, remember a generalization, or use a generalization on later learning.

Panailla had conducted a study which was mentioned by Morgensten & Renner (1984), in which he had analyzed 41 criterion test in Biology

*for the tenth grade aiming at recognizing its content of the cognitive levels according to Bloom's classification. The total number of the items of these tests was 2689 ones. Then an analyzing process was conducted by a special committee consisting of 12 arbitrators. Finally, the results of this study indicated that the levels of the questions of five of these tests were 100% cognitive levels was 90%. *إلى مراجعة* Also, the results have shown that the tests which their questions have supposed the application level are those built by teachers who have experience in teaching Biology.*

In a study conducted by Morgenstern & Renner (1984) to compare questions of five scientific subjects: Biology, Chemistry, Earth Sciences, General Sciences, and Physics in order to verify the extent of the measurement of the criterion test. The body of the study consisted of 30 tests, while the gross total number of the items amounted to 1077 of which 60% were randomly chosen. So the total number of the sample of the study came to 648 items. The two researchers took the 'Ten Mental Capabilities' which have been chosen by the Committee of Educational Policies in the USA as a criterion for classification.

The results of analysis have shown that the questions of Chemistry, Earth Sciences, and Physics did not outreach the 'remembering' level, and it has been noticed that the level of remembering question of all of the five subjects was high; while the General Sciences' test contained seven mental capabilities only.

Abu Helew (1984) conducted an analytic study of the content of the Social Education textbooks assigned for the students of the fourth, fifth and sixth elementary stages in the Jordanian public schools in order to explore the strength and stage aspects in the content as well as the general characteristics of these concerned Social Education books. For this purpose, the researcher has used the cognitive scope levels in analyzing the questions.

Royer and Konold (1984) examined Hunkin's (1969) study in which he investigated the effect of two levels of questions, knowledge (low – level) and evaluative (high – level) on students' achievement in two groups. One group studied social materials provided with knowledge questions, the other group studied the same passage but was provided with evaluative questions. After four weeks, all students took an

examination consisting of questions of all six levels of Bloom's Taxonomy. Results showed that the two groups did not differ on items. الكلام لتالي يخالف هذه النتيجة That is, students receiving higher level evaluation during the study phase performed significantly better on high level questions in the posttest. The results revealed that the questions had been worded in order to measure the cognitive, understanding and application levels. Although, the concentration was focused mainly on the cognitive level; therefore, the questions which measure the higher levels were few in the three text-books.

Thissen, Wainner, and Wang (1994) agreed with Bridgeman and Rock's results when they did not find a significant difference between multiple-choice and essay questions on students' achievement. They used 2000 students who took Computer Science and Chemistry tests of the College Board's Advanced Placement Program, and divided them into two groups: one group received multiple-choice questions, and the other one received essay question. Results showed that essay sections have the same effect on students' achievement as the multiple-choice on solving problem (application level) test.

Hiyagineh (1998) aimed at finding out the levels of written questions which are made by the teachers of Arabic Language at the secondary educational stage according to Bloom's Taxonomy and their relationships with a number of personal variables: sex, academic and professional qualification and experience of the teacher. The study tried to provide answers to the following four questions:

- 1. What are the levels of written questions made by the teachers of Arabic in the secondary stage according to Bloom's Taxonomy in the objectives of the cognitive domain in the schools of the general directorate of Education in Irbid?*
- 2. What are the types of written questions prepared by Arabic Language teachers in the secondary stage according to Bloom's Taxonomy?*
- 3. Is there a relationship between the levels of written questions in the school of (high - low) made by the teachers of Arabic in the secondary stage, due to sex, experience and teachers academic and professional qualification?*

The study sample consisted of 140 teachers; (46 males and 58 females). Question papers produced by teachers were collected and put into certain tables. The overall items were 9769. A certain guide developed by the researcher by the help of Bloom's Taxonomy was used to classify the written questions according to their cognitive levels in the secondary stage. The study revealed that:

- 1. All teachers used six levels of cognitive questions in Bloom's Taxonomy at different proportions. The percentage of knowledge questions used was 44.2% comprehension questions, 32.4% application questions 16.10%, analysis questions 4.60%, synthetic questions 2%, and evaluation questions 7% of the total number of the written questions.*
- 2. All the teachers used all patterns (types) with one level of Bloom Levels for cognitive questions with focus on law patterns or types.*
- 3. There were statistically significant differences ($\alpha=.05$) among the level of the written question (low - high) due to either sex, academic or experience.*

To conclude, most of the previous studies have shown that low level written questions (remembering and understanding) have occupied the first rank in the teacher's constructing tests. The previous studies have also shown that the rate of high level written questions (Application, Analysis, Synthesis and Evaluation) have occupied the second rank of the teacher's concern.

*The researcher concluded that some researchers (e.g., Arrasmith, Sheehan & Applebaum, 1984; and Roderick & Aderson, 1968) found that essay type questions have more effect on students learning than do multiple-choice ones especially on high levels of learning. كلام مخالف
Others, (e.g., Bridgman, 1992; Bridgman & Rock, 1993; Frase, 1968; Thissen, Wainner & Wang, 1994; Williams, 1963).
جملة غير مفيدة*

Through previous studies, it has been noticed that there is an urgent need to study the levels of a written testing questions prepared by teachers of English language and the relations of these questions to a number of personal variables of the teacher such as sex, academic

qualification and experience due to scarcity of such studies that tackle written testing questions in Saudi Arabia.

Methodology and procedures

Population

The population of this study consisted of all the teachers of third intermediate grade in Al-Ehsa' Directorate of Education during the academic year 2013/2014 in the first semester. The teachers had the same social background but differ in experience.

Sample

The sample of this study consisted of forty teachers (20 males and 20 females) who wrote forty tests and who are B.A holders in forty schools in Al-Ehsa' Directorate of Education. They all taught the third intermediate grade classes.

Table 1 shows the distribution of the subjects of the study in terms of experience and sex.

Table (1) : Distribution of the subjects of the study in Terms of experience and sex.

Sex	Experience		Total
	1-10	Above 10	
Male	10	10	20
Female	10	10	20
Total	20	20	40

Validity and Reliability of the instrument

The instrument of the study was given to a jury of eight TEFL teachers, four TEFL supervisors and two university professors. Their comments and recommendations were highly appreciated and taken into consideration. The validity was also achieved as the instrument has been used by the authors of Petra series for the basic stage since 1985.

Data Collection Procedure

The researcher collected forty final tests of the third intermediate grade classes of forty different teachers in the schools of Al-Ehsa' Directorate of Education.

The researcher developed a model (Appendix A1-4) to measure the four language components included in the tests (function, structure,

vocabulary, attitudes and values) available in the Content Analysis Table of the teacher's book.

To assess the acceptable percentage of the tests written by the subjects of the study, four M.A holders of TEFL who work as supervisors in Al-Ehsa' Directorates of Education and have an experience of five years in the field of test construction and evaluation were asked to construct a test based on the content analyses table in the teachers' book of the third intermediate grade. The supervisors' tests were analyzed by a jury of judges that consisted of four professors who work as supervisors of English with an experience of ten years and above.

The supervisors' tests cover 58% of the functions, 64% of the structures, 56% of the attitudes and values and 73% of the vocabulary in the teachers' book

Findings of the Study

As for the first question which deals with the extent to which the EFL teachers cover the language components in writing their testing items, table 2 presents the means and standard deviations of the number and percentages of language functions that are included in the tests written by the subjects of the study.

Table 2 shows that language function 3 which is extracting information from a timetable, took the highest percentage with an average of 28%. Language function 8, which is giving instructions and explaining how things work, was the second with an average of 27%. Language function 2, 16 and 35 took the third place with an average of 25%. Table 2 also shows that the average mean of the percentages of the fifty language functions is 16%.

Table 3 presents the number and percentages of structures included in the tests.

The results presented in table 3 show the percentages of each structure included in the subjects' tests. The table shows the average mean of the percentages of the structures is 34%.

Table 4 presents the means and standard deviation of language values and attitudes included in the subjects testing items.

Table 4 shows the means and standard deviation of the number and percentages of language values and attitudes. It is noticed that the average means of the percentages of the values and attitudes is 16%.

Table five presents the means and standard deviations of vocabulary included in the subjects' testing items.

Table 2 : Means and Standard Deviations of the Numbers and Percentages of the Language Functions Included in Tests

Language functions numbers	Means of percentages	Standard deviation	Language functions	Mans of percentages	Standard deviation
3	28	0.45	34	17	0.38
8	27	0.45	38	17	0.38
2	25	0.44	23	15	0.36
16	25	0.44	29	15	0.36
35	25	0.44	37	15	0.36
7	22	0.42	47	15	0.36
15	22	0.42	22	13	0.43
25	22	0.42	24	13	0.34
40	22	0.42	30	13	0.34
43	20	0.40	42	13	0.34
6	20	0.40	44	13	0.34
9	20	0.40	45	13	0.34
14	20	0.40	21	12	0.32
27	20	0.40	41	12	0.32
36	20	0.40	50	12	0.32
46	20	0.40	10	10	0.30
1	18	0.39	20	10	0.30
11	18	0.39	39	10	0.30
28	18	0.39	4	08	0.28
31	17	0.39	19	08	0.28
5	17	0.38	32	07	0.25
13	17	0.38	48	07	0.25
18	17	0.38	49	07	0.18
33	17	0.38	56	03	0.04
Q ALL				16%	

Table 3 : Means and Standard Deviations of the Numbers and Percentages of Structures Included in the Subjects' Tests

Structures	Means of percentage	SD	Structures	Means of percentage	SD
3	50	0.50	2	33	0.48
2	50	0.50	27	33	0.48
2	43	0.50	30	33	0.48
4	43	0.50	1	32	0.47
23	42	0.50	17	32	0.47
10	40	0.49	25	32	0.47
28	40	0.49	29	32	0.47
21	38	0.49	16	30	0.46
39	38	0.49	31	30	0.46
40	38	0.49	6	28	0.45
13	37	0.49	7	28	0.45
18	37	0.49	26	38	0.45
24	37	0.49	37	28	0.44
8	35	0.48	12	27	0.44
35	35	0.48	33	27	0.43
36	35	0.48	11	25	0.42
5	33	0.48	34	25	0.40
9	33	0.48	15	23	
14	33	0.48	32	22	
19	33	0.48	38	20	
Q ALL				34%	

Table 4 : Means and standard deviations of Number Percentages of Language Values and Attitudes in the Subjects' Tests

Values and attitudes	Means of percentages	SD	Values and attitudes	Means of percentages	SD
27	0.28	0.45	10	15	36
24	27	45	11	15	36
22	23	43	1	13	34
28	23	43	2	13	34
18	22	42	17	13	34
21	22	42	3	12	32
23	22	42	13	12	32
25	22	42	6	10	30
29	18	42	12	10	30
4	18	39	16	8	28
15	18	39	9	5	22
19	17	39	14	3	18
5	17	38			
8	17	38			
20	17	38			
62	15	38			
7	15	36			
QALL				16%	

Table 5 : Means and standard Deviations of the Numbers and Percentages of Vocabulary Included in the Subjects' Tests

Vocabulary	Means of percentages	SD	Vocabulary	Means of percentages	SD
10	56	0.50	5	32	47
1	52	0.50	6	32	47
8	48	0.50	7	32	47
4	45	0.50	14	32	47
13	44	0.50	16	32	47
12	42	0.50	18	32	47
2	39	49	21	31	46
3	39	49	22	31	46
9	39	49	20	29	46
11	39	49	19	27	45
17	35	48	24	27	45
23	35	48	15	26	44
QALL				37%	

The Table above presents the means and standard deviation of the numbers and percentages of vocabulary included in the test written by the subjects of the study. It also shows that the average mean of the percentages of the vocabulary items is 37%.

Table 6 presents a summary of the overall means of the teachers tests (functions, structures, vocabulary and attitudes and Values) compared to the overall means of the tests constructed by the supervisors.

Table 6 : Overall means of the subjects and the supervisors tests

Components of the CAT	Functions	Structures	Attitudes values	vocabulary
Subjects	16%	34%	16%	37%
Supervisors	58%	64%	56%	73%

Table 6 shows that the overall means of the subjects' tests compared to those of the supervisors are quite far from the degree of acceptability which was obtained through calculating the means for the four supervisors' test scores. As for the functions, the subjects' tests cover 16% while that of supervisors cover 58%.The supervisors' tests cover 64% of the structures whereas the subjects tests cover 34%. Concerning the attitudes, values and vocabulary, the subjects cover 16% and 37% respectively, while the supervisors' tests cover 56 and 73, respectively.

Concerning the second question which deals with whether or not there are significant differences in the EFL basic teachers' tests due to sex and experience variables, Table 7 presents a summary of the performance of the subjects.

Table 7 : Subjects Performance with Experience Males and Females

SEX	Experience		Total
	short	Long	
Male	0.18	0.18	0.18
Female	0.12	0.16	0.14
Total	0.15	0.17	

The table above shows that there are statistically significant differences among teachers due to experience. The average mean of long experience teachers is 0.34 while that of short experience is 0.30.

It could also be seen from the table that the overall average mean of males' performance is 0.36 whereas that of the females is 0.28. It is also obvious from the table that there are no statistically significant differences among male subjects due to experience. Both short and long experience subjects got an average mean of 0.18. It can be also seen that there is a statistically significant difference among female teachers due to experience. Female teachers of long experience got an average mean of 0.16 where as the average mean of short experience female teachers is 0.12. Moreover, it is noticed that the performance of male teachers with short experience is better than that of short experience female teachers. Male teachers with short experience got an average mean of 0.18 while female teachers with short experience got an average mean of 0.12. In addition, the average mean of long experience male teachers' performance is 0.18 whereas that of the females is 0.16. All in all, it is clearly seen that the performance of male teachers (long and short experience) is better than that of the female subjects.

The superficial reading of Table 7 shows that the male subjects are better than the female ones. But by looking deeply to the Table, we can see that the male subjects do not benefit from experience while the female subjects grow up during experience.

Concerning the performance by sex and experience and the interaction between them, Table 8 presents summary of the performance of the subjects.

Table 8 : Subjects performance of sex experience and interaction between them.

	Sum of sq.	DF	M.Sq	F	Significance Of F
Sex	0.037	1	0.037	31.228	0.00
Experience	0.019	1	0.019	6.858	0.011
Sex*experience	0.006	1	0.006	2.059	0.157
Residual	0.155	56	0.003		
Total	0.215	59	0.004		

Significant at $\alpha=0.05$

From the table above, it can be seen that there is a statistically significant difference between male and female teachers. It can also be seen that there is a statistically significant difference among teachers due to experience. As for the interaction between sex and experience, there is no statistically significant difference at ($\alpha=0.05$).

Discussion, conclusions and recommendations

The present study aimed at investigating whether or not the tests constructed by the teachers of the third intermediate in Al-Ehsa' Directorate of Education cover the language components enclosed with the teachers' book in terms of structure, function, vocabulary and attitudes and values.

The researcher used percentages to reveal the extent to which teachers write comprehensive tests that cover the components. He also used ANOVA test to identify the differences between male and female subjects due to sex and experience.

Discussion of results:

The findings concerning the first question of the study, which deals with the extent to which teachers write comprehensive tests that cover the components , indicate that the teachers cover only 16% of the language functions included in the TG as shown in Table 2. The teachers' tests also cover 34% of the language structures included in the TG as indicated in Table 3. In addition, the tests only cover 16% of the values and attitudes in the TG, while the subjects' tests cover 37% of the vocabulary items in the TG.

These results, compared with those of the supervisors', are below the average of acceptability. The subjects' tests cover 16% of the functions, 34% of the structures, 16% of the attitudes and values and 37% of the vocabulary items whereas the supervisors' tests cover 55% of the functions, 60% of the structures, 52% of the attitudes and values and 65% of the vocabulary as can be seen in Table 6.

This result could be attributed to the following issues:

- 1. The teachers' book doesn't include any sample tests that could be referred to when constructing tests.*
- 2. Teachers do not have a theoretical background on tests construction which could be attributed to the lack of teaching a course on assessment and evaluation at university.*
- 3. Most of the in-service training programs held by the Ministry of Education neglect topics dealing with test construction.*
- 4. From the researcher's experience, teachers tend to construct tests that could be easily marked regardless of their comprehensiveness.*
- 5. Most Teachers in the field usually do not take into consideration the feedback from the supervisors on the test they construct.*
- 6. Teachers complain most of the time that the Ministry Education*
- 7. Regulations (for passing and failing) impose a limitation upon their testing procedures since all students should pass; therefore, there is no need for a serious and a systematic testing process.*
- 8. Teachers restrict themselves to the textbook so they emphasize explicit components of the TGs in their teaching and they are reluctant to go beyond the text. This means that they neglect implicit components such as attitudes and values and functions.*
- 9. Teachers in general are not aware of the integration of the components in the TGs, so they tend to concentrate on certain components (structures and vocabulary) while ignoring others (attitudes and values and functions).*

10. *Teachers still look at English as only a school subject rather than a language of real life situations in which functions and attitudes and values should be highlighted.*

Table 6 also shows that the subjects of the study constructed more comprehensive test items on structures and vocabulary than the functions, attitudes and values. The researcher believes that this result is expected as teachers in general focus on vocabulary and structures most of the time. This could be attributed to the fact that teaching functions, attitudes and values is more difficult than teaching structures and vocabulary. In addition, teachers find it easy to build a structure and a vocabulary test as most of the test items written by the teachers have the form of true/false, multiple choice, filling in gapes and matching. These test items could be marked more easily than other sets of questions.

Concerning the second question which deals with whether or not there are statistically significant differences between male and female EFL teachers due to experience and sex, it was found that there was a statistically significant difference among teachers in favor of long experience. This result is expected as those of long experience have more practice in test construction than those of short experience. It is also supposed that teachers of long experience have attended more training courses on test construction.

Moreover, the interaction and exchange of experience of long experience teachers do play a part in developing their competence to build more comprehensive tests. Having more interaction and exchange of experience, help them to keep in touch with modern techniques and improvements in testing and teaching methodologies. This result goes with that of Maribor, (1972); Hassan, (1984); Al Hader, (1991); and Al-Hayajneh, (1998). They found that there is a positive relation between experience and the level of test items.

Regarding the sex variable, it was found that there was a statistically significant difference between male and female teachers in favor of male teachers. This result could be attributed to the fact that male teachers attend regularly training courses more than female teachers. Male teachers are also actively involved in the in-service training programs more than female teachers as female teachers have other daily life interests. Male teachers also exchange visits the aim of which is to

exchange experience more than female teachers. The fact that more male teachers participate in marking the General Certificate Examination gives them the chance to develop their competence for building more comprehensive test items.

Recommendations:

Based on the findings of the study, the researcher recommends that:

- 1. EFL teachers vary their test items to achieve a balanced weight of language components.*
- 2. EFL teachers make a comprehensive content analysis before constructing their test items.*
- 3. Supervisors share their teachers in building tests items.*
- 4. The Ministry of Education provides teachers with sample tests.*
- 5. The Ministry of Education starts a training program the aim of which is to train teachers on how to build a comprehensive, authentic and valid test.*
- 6. Further studies should be conducted on testing speaking, and listening in schools all over the country.*

References:

- Alderson, C. (1991). *Language Testing in the 1990s*. ERIC.365145.
- Anderson, R. & Biddle, W.B. (1975). On Asking People Questions about What They are Reading. In G. Bower (ED.), *Psychology of learning and motivation* . Newyork Academic Press, Vol.9, pp89-132.
- Andre,T. (1979). Does Answering High- Level Questions while Reading Facilitate Productive Learning? *Review of Educational Research* 49(2), 280-318.
- Black R. T. (1980). An Analyses of Levels of Thinking in Nigerian Science Teachers Examinations, *Journal of Research in Science Teaching*, 17, (40), pp.301-306.
- Cronbach L. J. (1970). *Essentials of Psychological Testing*. Third Edition. Happer and Row Publishers, Inc. Newyork,P.26.
- Darwazeh, A.N.(1982). Students-generated versus Teachers-generated Adjunct Questions: Their Effect on Remembrance and

- Application Level Learning. *Doctoral dissertation*, Syracuse University.
- Doye, P. (1991). Authenticity in Foreign Language Testing. *ERIC.35089*.
 - Goro, F. F. (1972). *The Learning Game: Strategies for Secondary Teachers*. Charles. E. Marrill Publishing Company, Columbus, OH, op(68).
 - Hugs, A. (1989). *Testing for Language Teachers*. Cambridge University Press. London, England.
 - Kneip, W.M. and Grossman, G.(1979). The Effect of High Level Questions in Competitive and Cooperative Environments on the Achievement of Selected Social Studies Concepts. *The Journal Education research,73(2),82-85*.
 - Lewis R. A. (1977). *Psychological Testing and Assessment* , Second Edition. Allyn and Bacon , Boston,P. 332.
 - Perkins, K. (1990). A level of Questions in ESL Reading Comprehension .(*ED 322709, 1982-1991*).
 - Redfield, D.L. (1981). A Meta –analysis of Experimental Research Teacher Questioning Behavior. *Journal of Educational Research,51(2)*, pp.237-245.
 - Sax, G.(1980). *Principles of Educational and Psychological Measurement and Evaluation*. Second Edition. Wardwarth Publishing Company, Inc. Belemont, California.P.13o.
 - Tinsley, D.C.and Davis,O.L.(1971). Questions Used by Secondary Student Teachers to Guide Discussion and Testing in Social Studies: A study in Planning . *The Journal of Teacher Education,22(1),59-65*.
 - Watts, G.H.and Anderson,R.C.(1971). Effects of Three Types of Inserted Questions on Learning from Prose. *Journal of Educational Psychology, 62(5)9*, pp. 387-394.

