# Using Arabic (L1) in Testing Reading Comprehension in English (L2) as a Foreign Language

Abdullah H. A. Al-Qudairy

PhD

The University of Edinburgh

2010

# ABSTRACT

The purpose of this study was to investigate the effect of using Arabic (L1) as a language of questions and answers in testing reading comprehension in English (L2), and to explore student and teacher opinions about this.

Both quantitative and qualitative methods were employed. To collect the quantitative data, one hundred and forty-four students were given a reading comprehension test. Both multiple-choice and short-answer questions were used. The subjects were second-year English department undergraduate Saudi students and final-year secondary school Saudi students. Other factors including gender and five reading sub-skills were considered. Twelve students and four English-language teachers participated in semi-structured interviews, the source of the qualitative data.

The findings of this study indicate that, for the population, test types and test levels investigated, there is no clear case for having reading comprehension questions and answers in L1. The use of Arabic in the English reading comprehension tests did not improve the performance of students. Interview responses were mixed, but with no consensus in favour of Arabic. Limitations of this study are discussed, and recommendations for further research in testing reading comprehension in English as a foreign language are presented.

## TABLE OF CONTENTS

CHAPTER Page

# LIST OF TABLES

LIST OF FIGURES

# CHAPTER ONE: INTRODUCTION

In this first chapter, I will briefly introduce the background of my study, and discuss the research problem, purpose and significance. Then, I will overview the remaining chapters in this thesis.

## 1.1 Background of the Study

The present research is related to my work as a teaching assistant at the College of Language and Translation, Al-Imam University in Riyadh, where I have been working since 1995. At the Department of English Language and Literature, within the college, all prospective students are required to pass an English-language proficiency test in order to join the department. Those who fail the test are given a second chance to take the test, but after attending a four-month intensive English-language programme which is administered by the college itself.

Besides teaching EFL in this intensive programme for several years, I was a member of the examination committee that was responsible for developing and administrating these proficiency tests. During these years, I had the opportunity to speak to a large number of students who joined the intensive English-language programme or those who passed the test from the first time. As expected, there were complaints about the test from the students, but I noticed that there was a common complaint about the language of the questions, i.e. its structure and vocabulary. For example, in the reading section of the entrance test, they claimed that they understood the passage but they faced difficulties in understanding some of the questions because of their language, which prevented them from getting higher scores to pass the test. They argued that if questions were translated for them into

Arabic (L1), they would have been able to answer the questions correctly. Moreover, some of them claimed that they understood the questions and knew the correct answer but they could not write down their answers because they were either afraid of grammatical and spelling mistakes or they did not know the English vocabulary needed to write their answers. In other words, if they were given the chance to write their answers in Arabic (L1), they would have been able to write their answers without any problem.

## 1.2 Statement of the Problem

These repeated complaints about the language of the test questions raised validity questions about these tests especially with receptive skills such as reading. Does the use of English (L2) in testing reading comprehension of English as a foreign language increase or decrease the validity of such tests? Would it be more valid to use Arabic (L1) in testing English-L2 reading comprehension? In other words, would it help to make reading comprehension questions in students' native language, Arabic, and also let them write down their answers in Arabic?

The validity issue was the essence of my concern. It would be to the benefit of both the students and the educational institutions to have valid tests which may reflect the real level of the test-taker. This might give qualified students the chance to perform well in such tests, and also could give these institutions the ability to choose the students they want according to their standards and needs.

Although much attention has been given to factors that might affect testing reading comprehension, there has been very little interest in the effect of using L1 in testing L2 reading comprehension. “The language chosen in L2 reading comprehension tests has been the focus of attention of only a few studies” (Martinez

& Godev, 1994, p. 2). To the researcher’s knowledge, there are no studies on the effect of the use of Arabic (L1) in testing reading comprehension in English as a foreign language in Saudi Arabia. This absence of attention to this important factor has kept the picture unclear. Alderson (2000) asks “when test-takers share a first language, might it be better to ask questions in that language” (p. 86), and then he comments that all these speculations about this issue “have yet to be confirmed by research” (p. 87).

According to Hughes (2003), the use of L1 in testing L2 reading comprehension becomes more appropriate in monolingual situations. He said that “where candidates share a single native language, this can be used both for items and for responses” (p. 153). Shohamy (1984) argues that “presenting the questions in L1 may be considered more ethical, since the decision maker obtains information on the test taker’s ability to understand the L2 text, without a carry-over from the language of the questions” (p. 158).

Off course, it might be claimed that giving the students the option of reading the questions and answering them in their native language will enable them to write exactly what they believe is the right answer without being afraid of grammatical and spelling mistakes, and without fear of not knowing the right English word or expression needed for the answer. However, this might be true in cases where they fully understand the text, but what about real-life testing situations where students usually understand part or most of the text but not all of it? In this case, is the use of English (L2) as the language of the test an obstacle preventing students from writing what they believe is the correct answer, or there are other factors that might prevent them from performing well in the test? Is the use of L2 in testing L2 reading comprehension a real reason or just an excuse for not being able to pass the test? This cannot be answered without carrying out a study that takes into consideration the independent variables that might have an effect on students performance.

To address the validity issue in using Arabic (L1) in testing reading comprehension in English (L2) as a foreign language, the researcher used both quantitative and qualitative methods. Ninety-four case-II independent sample t-test studies were carried out to trace any significant differences among the study variables that might have an effect on students performance.

For the purpose of this study, the researcher developed two reading comprehension tests for both university and secondary school students. Both tests were piloted with similar participants to the main study. The two tests were then revised according to the results of the item analysis of the pilot study which took place before the administration of the main study. More details are found in chapter five. Moreover, in order to achieve a broader understanding of the problem, the researcher included three independent variables besides the main independent variable, namely the language of the test. The other three independent variables were: testing method, gender and five reading sub-skills.

To collect the quantitative data, a total of two hundred and twenty-four male and female students participated in the study. Eighty students participated in the pilot study, and the other one hundred and forty-four students were part of the main study. Students were given a reading comprehension test. Both multiple-choice and short-answer questions were used. The subjects were second-year English department undergraduate Saudi students and final-year secondary school Saudi students. Both parts of the study (i.e. university and secondary school levels) were similar in the structure and the number of participants. The differences between them were the length and difficulty of the passages and the questions. Each group was divided into two sub-groups which were given the same version of the test. The only difference was the language of the questions and answers. The first sub-group had the questions in English and wrote their answers in English; the second sub-group had the questions in Arabic and was asked to write their answers in Arabic. I tried to ensure that students had the same level of language ability. For example, all of the secondary school students scored over 70 out of 100 in the English language exam the previous year. In addition, all of the university students passed the Department of English and Literature entrance test, and passed all the first-year English Department courses. Chapter four contains full details of the quantitative research design.

For the qualitative data, the researcher conducted sixteen one-to-one semi-structured interviews with both university and secondary school students and their English-language teachers to know their opinions about the use of Arabic (L1) in the questions and answers of the test, and also to explore their experience in learning English (L2) and its assessment. Chapter seven contains more details about the qualitative design of the present study.

## 1.3 Purpose of the Study

The present study is designed to explore three topics: the use of Arabic (L1) in testing reading comprehension in English (L2) as a foreign language; how the study independent variables (testing method, proficiency level, gender and five reading sub-skills) might affect the performance of students in both the Arabic and English versions of the test; and the views and impression of university and secondary school students and their English-language teachers about the use of Arabic (L1) in the questions and answers of the test.

The following research questions were designed to guide the present study:

1- Does using Arabic (L1) in testing English (L2) reading comprehension affect the levels of performance of upper-intermediate and post-beginner students in multiple-choice and short answer questions?

2- When Arabic (L1) is used as the language of the questions and answers of an English (L2) reading comprehension test, how would gender, testing method, proficiency level, and reading comprehension sub-skills affect the level of performance of test-takers?

3- What do university and secondary school students and their English-language teachers think of using Arabic (L1) in the questions and answers of the reading comprehension test?

## 1.4 Significance of the Study

According to Alderson (1984), “a reading ability is often all that is *needed* by learners of English as a Foreign Language (EFL), as well as of other foreign languages” (p. 1). Therefore, students’ inability to read L2 materials might negatively affect their academic progress and success. It is important to have valid reading tests in order to be able to assess these students correctly, which might help in diagnosing their ability more accurately, and consequently teach them what they really need to improve their reading ability to become successful learners. It is hoped that the results of the present study may help the Saudi university and secondary school students by giving them the chance to show their real EFL proficiency level, which will help them to succeed in their academic and every-day life.

It is hoped that the results of the present study may provide insights for the educational decision makers in Saudi Arabia to improve the current state of testing reading comprehension in Saudi schools and universities. Language testing is not well researched in Saudi Arabia, and this study is a little contribution to this important field.

Moreover, I hope that the results of this research may be an encouraging step towards a better understanding of some of the unclear aspects of using L1 in testing L2 reading comprehension. I hope that it will urge linguists and testing specialists to conduct more research on variables related to the use of L1 in testing L2 reading comprehension. Finally, I hope that the study may provide useful information for other EFL/ESL studies which have contexts similar to the one in Saudi Arabia.

## 1.5 EFL in Saudi Arabia

It is important to describe the EFL context in Saudi Arabia in order to provide a frame of reference for this thesis, especially for the readers who are unfamiliar with the Saudi educational context.

English has a special status in Saudi Arabia. It is now the only foreign language taught at all levels of education starting from primary level and ending up with graduate studies in almost all majors. Although, in several graduate programs, the language of instruction and assessment is Arabic, students are required to pass an English language proficiency test such as TOEFL or IELTS as an entrance requirement to these departments in order to make sure that students will be able to have access to and benefit from a wider range of references in their study and research.

### 1.5.1 School Types

There are two main types of schools within the Saudi school system - public schools and private schools. Public schools are owned and run by the Ministry of Education. The second type, private schools, are owned and run by Saudi individuals but under the supervision of the Ministry of Education.

In public schools, EFL starts from the final grade in primary school at the age of twelve. In private schools, however, EFL generally starts from the first primary grade at the age of six, and in some cases as early as preschool level. Private schools follow the same curriculum used in the public schools except for English and some elective skill-courses. They are free to select their English textbooks from preschool levels and up to the fifth grade of primary school. From the final grade in primary education, private schools have to use the same EFL books and curriculum used by the public schools. These books are developed by the ministry itself and specially designed for Saudi students where certain cultural conditions are applied (Abdan, 1991). Alabdelwahab (2002) expressed his concerns about an existing problem in teaching English as a foreign language in Saudi Arabia.

> Unlike public schools that do not teach EFL until the seventh grade, private schools often begin English instruction in grade one. Because the Ministry selects the seventh-grade EFL textbooks, seventh-grade students in private schools, who may have studied English for six years, use the same textbooks as seventh-grade students in public schools who are just beginning to study English. This can be frustrating for private-school students who have reached a stage where they can read English but are required to use the Ministry-approved textbook that teaches the alphabet, for example (p. 18 – 19).

### 1.5.2 EFL Curriculum

The exact date of the introduction of English as a foreign language in the Saudi educational system is not known, however it might be dated back to the establishment of the General Directorate of Education in 1924. In the same year, the first public primary school was opened, and English was taught as a core course (Al-Abed Al-Haq and Smadi, 1996).

The English-language Section of the General Directorate for Curriculum Development at the Ministry of Education supervises and controls the teaching of English in Saudi Arabia. It is responsible for developing the instructional materials and setting syllabus guidelines. EFL textbooks are assigned and distributed by the Ministry of Education throughout the country to all students in primary, intermediate, and secondary schools (Al-Seghayer, 2005). In other words, for each subject, all students of the same grade in any school will have the same textbook, and teachers usually teach and assess the same language skills and sub-sills in quite a similar way (Alabdelwahab, 2002).

The curriculum of English language in Saudi Arabia went through a number of major changes during the last eighty years (Al-Seghayer, 2005). The changes and developments can be summarized in table 1.

Table: 1

| Year | Textbook title | Developer |
|---|---|---|
| 1927 - 1959 | No definite curriculum | NA |
| 1961 - 1980 | Living English for the Arab World | It was adapted by the Ministry of Education from neighbouring countries' curriculum |
| 1980 - 1995 | Saudi Arabia Schools English | Ministry of Education + Macmillan Press |
| 1995 - now | English for Saudi Arabia | The Curriculum Department at the Ministry of Education + EFL specialists from King Fahad University in Saudi Arabia |

The last series of books “English for Saudi Arabia” underwent several modifications during the last fifteen years based on English-language teachers’ suggestions and supervisors’ reports.

### 1.5.3 EFL Teaching

The most popular methods of teaching English in Saudi Arabia, according to Al-Seghayer (2005), are the audiolingual method and the grammar translation method. They are preferred by English language teachers even though some facilities are not available like language laboratories which are essential to the audiolingual method. Al-Seghayer adds that English language teachers often use Arabic in teaching and sometimes depend on translation. This system, “although built on sound pedagogical objectives, fails to produce learners who can carry on a basic conversation or comprehend a simple oral or written message” (Al-Seghayer, 2005, p.129). According to Alabdelwahab (2002), many Saudi students do not know how to figure out the meaning of unfamiliar words because they do not try to explore the various connotations of the word. This might be due to teachers’ method of teaching English by concentrating on memorizing words and their direct meanings rather than using contextual clues to guess the right meaning. Al-Seghayer (2005) argues that developing communicative competence is not a priority in teaching English language in class. One main reason for that is the common belief among teachers that reading and writing skills are the most important ones. Therefore, more time is spent and more attention is paid to teaching the language content rather than its communicative aspects.

### 1.5.4 EFL Teacher Preparation

The preparation programs for English-language teachers in Saudi Arabia can be described as “non-systematic and inadequate” (Al-Hazmi, 2003, p. 341). The author adds that most of the EFL teachers who are graduates of colleges of art and colleges of education in Saudi universities take only one course of EFL teaching methodology which does not fulfil their diverse teaching needs.

> It is ironic that the MoE (Ministry of Education), which has done so much to improve and update English language curricula since 1991, has lagged behind in doing the same for EFL teacher education programs. The gap between the content of teacher education programs and the needs of the classroom widens. After graduating from university, many teachers lack essential English skills, especially the ability to speak the language …. a 1-year TEFL diploma should be a minimum requirement for newly appointed preservice teachers (Al-Hazmi, 2003, p. 342-343).

According to Al-Hazmi (2003), in a recent move, the Ministry of Education organised intensive training programs for EFL teachers in collaboration with the British Council and the American Embassy in Saudi Arabia. More than six hundred Saudi EFL teachers and supervisors are expected to join these training programs. Another recent positive move by the ministry in the last few years is sending hundreds of EFL supervisors and teachers, mainly to United Kingdom, in order to acquaint them with the current developments and changes in the theory and practice of TEFL.

## 1.6 Organization of the Thesis

The thesis is organised into eight chapters. Chapter one has presented the problem that is to be researched and its questions, the methodology to be used, the significance of the study, and a background of EFL in Saudi Arabia.

Chapter two is a literature review of reading comprehension models and variables. It traces the historical development of reading comprehension models. In addition, it gives a brief overview of the reader, text, and testing variables that may have an effect on the reading process and its assessment.

Chapter three is a literature review of some of the main issues and considerations in language testing. It covers issues such as: reliability and how it could be increased, the sources of variance in examinees’ performance, validity and how to validate a test, and Messick’s six aspects of validity. Furthermore, a brief history of validity is introduced, and a general discussion of test specifications is presented.

Chapter four describes the research design. It states the test specifications of the two reading comprehension tests including: content, testing methods, format and timing, and scoring procedure. It states the research questions, and describes the piloting process and the item and statistical analysis used in this study.

Chapter five describes in detail all the steps of data collection. It includes a comprehensive item analysis of the pilot study, and a detailed description of the test development.

Chapter six presents the findings related to the quantitative research questions. It explores the results of the ninety-four case-II independent sample t-test studies which cover all the independent and dependent variables in the main study.

Chapter seven contains a brief introduction of the interview, its history, types, structure, techniques, and schedule. It describes the sixteen semi-structured interviews with both university and secondary school students and their English-language teachers. Furthermore, an analysis of the interviews with selected quotations was presented.

Chapter eight summarizes the main findings of the present study, discusses its successes and limitations, and suggests some recommendations for further research.

# CHAPTER TWO: READING MODELS AND VARIABLES

## 2.1 Introduction

There is no agreed definition of reading comprehension. Smith (1985) affirms that "there is no point in looking for a single definition of reading…. We should not expect that a single definition for reading will be found, let alone one that throws light on its mystery" (p. 100). In the same vein, Grabe (1991) believes that "simple definitions typically misrepresent complex cognitive process such as reading" (p. 378). Alderson (2000) writes that "an overview of the study of the nature of reading is impossible" (p. 1), and Shohamy (1984) argues that reading comprehension "in a second language is even more of a puzzle because it involves unknown aspects from first and second language" (p. 148).

Although it seems that testing reading comprehension is an easy task to accomplish, nevertheless the complexity and ambiguity of the nature of reading comprehension make it a challenging one. Alderson (2000) argues, however, that one should develop reading comprehension tests even though one's understanding of its nature is "faulty, partial and possibly never perfectible" (p. 2).

## 2.2 Models of Reading

It is important to trace the historical development of reading comprehension models, which will help in understanding the reading process in a better and deeper way. However, it is also important to recognize that "the notion that there can be a single model for reading across tasks, genre, and purpose is doubtful" (Hudson, 2007, p. 31).

### 2.2.1 Bottom-up Theory

According to Pearson and Stephens (1994), reading was seen in the mid-1960s as a perceptual process: first readers decode letters in a printed text into sounds, then they listen to these sounds and understand the words. In fact, reading was seen as a similar skill to listening: the eye works only as a decoder of the written text, and comprehension happens by listening. This view of reading made it a prerequisite for teachers to teach phonetics to their students if they wanted them to understand what they read. A well-known example of such theories is Gough's bottom-up theory (1972), where reading is seen as a linear process in which letters are identified one by one and then converted to sounds; each letter is held in the memory until the next letter is identified; when words are recognized, they, too, are held in the memory until the meaning of the sentence and eventually the paragraph is understood; the reader is seen as a passive decoder. Comprehension happens when phonemic processing is rapid and efficient, however, there is no need for or effect of prior knowledge. Readers are expected to read all words in the text in order to achieve comprehension. "I see no reason, then, to reject the assumption that we do read letter by letter. In fact, the weight of the evidence persuades me that we do serially from left to right" (Gough, 1972, p. 335).

Another example of bottom-up theory is LaBerge and Samuels' model (1974) in which comprehension is based upon word identification, and the notion of "automaticity" in word identification is introduced. In this model, it is important to master the skill of decoding (through practice) to the degree that it happens automatically. Usually, beginning readers pay more attention when faced with a new or difficult word in the text, which might affect their comprehension.

LaBerge and Samuels claim that it is only possible to pay attention to one thing at a time, however it is possible to process several things at the same time if they are done automatically with no attention needed. Therefore, when text decoding is efficient and automatic, attention will be directed towards comprehension.

According to Samuels and Kamil (1988), such linear models have a serious deficiency because they “pass information along in one direction only and … do not permit the information contained in a higher stage to influence the processing of a lower stage” (p. 27). Moreover, according to Stanovich, a serious deficiency of the bottom up theory is the lack of feedback “ in that no mechanism is provided to allow for processing stages which occur later in the system to influence processing which occurs earlier in the system” (as cited in Samuels and Kamil, 1988, p. 31). Furthermore, such models do not explain why the reading process is not affected by context or prior knowledge. In addition to these shortcomings, the bottom up theory “lacks flexibility. The reader has no choice of operations or strategies to deploy in different reading tasks” (Mitchell, 1982, p. 133-134).

Despite all the criticism it received, the bottom-up theory was an important step in understanding the process of reading comprehension. It inspired other researchers to do more work, which enriched the field of reading in general. The following theory, top-down, may offer a better explanation of the reading process and how comprehension is achieved.

### 2.2.2 Top-down Theory

As we have seen, bottom-up models start with the printed text and then move up to derive meaning. Top-down models, as we will see, start with higher levels of cognitive processes by making predictions and assumptions and then trying to confirm or reject them by working down to the printed text. Top-down models benefited a lot from the work of Smith (1971) and Goodman (1969, 1982), who emphasized the role of the reader as an essential participant in the reading comprehension process (Alderson, 2000).

The model that best represents the top down theory is Goodman’s psycholinguistic model of reading (1970). He believes that reading is meaning oriented: readers bring their knowledge and experience when they read. They sample, select from the text and make assumptions and predictions based on cues from letters, words, and syntax, and while they read they accept or reject their predictions by using their knowledge and experience that they had from the beginning. Goodman argues that:

> Reading is a selective process. It involves partial use of available minimal language cues selected from perceptual input on the basis of the reader’s expectation. As this partial information is processed, tentative decisions are made to be confirmed, rejected, or refined as reading processes (Goodman, 1976, p. 498).

Goodman does not see readers as passive decoders of letters and words but as active constructors of comprehension (Alderson, 2000). Thus, comprehension in reading is a continuous and active process from the very beginning until the end. As Goodman (1970) stated, “Reading is a psycholinguistic guessing game. It involves an interaction between thought and language”(p. 108). Comprehension does not come from accurate perception and identification of all letters and words as in the bottom up theory, but from the ability to choose the cues essential to making the right guesses. Readers use only some of the text cues to predict, which explains the errors children make when they read aloud and replace some words with different ones that do not change the grammar of the sentence (Goodman, 1970).

Goodman's model has received some criticism. For example, according to Mitchell (1982) the model does not give enough details about the reading process. Also, most of Goodman's work was with children, who have different reading strategies from adults. Mitchell believes that the model is inadequate and does not describe fluent reading. According to Samuels and Kamil, (1988), one of the shortcomings of the top down model is that "for many texts, the reader has little knowledge of the topic and cannot generate predictions" (p. 32). Therefore, if beginners are presented with a new subject in which their knowledge is limited, then how can they  compensate from bottom-up ability, since it is itself is weak and incomplete (Alderson, 2000). Another shortcoming, according to Samuels & Kamil, (1988), is that the amount of time needed to make predictions, even for the skilled reader, is more than the time needed to decode and identify words. In other words, it is easier and more efficient for a skilled reader to identify all the words than to make predictions and then confirm or reject them. Consequently, the top-down model fails to describe skilled reading behaviour.

Bottom-up and top-down theories are almost the opposite of each other, and seem mutually exclusive, but neither of them can fully explain the reading process. Nevertheless, both have some valid insights. In an attempt to build on these insights, a new theory was developed to explain the process of reading comprehension in a better and deeper way. This new theory was called 'interactive'.

### 2.2.3 Interactive Theory

Both bottom-up and top-down models are essential in understanding reading comprehension. However, the interaction between the two models is complicated and unclear. The interaction differs according to the text, reader, and purpose (Alderson, 2000). Interactive models attempt to combine valid insights of both bottom-up and top-down. However, they differ in their degree of focus on process or product respectively.

The term 'interactive' is used to represent different views. According to Grabe (1991), some writers use it to refer to an interaction between the reader and the text, while others refer to the interaction among different reading skills.

Stanovich was one of the researchers who succeeded in combining both bottom up and top down theories in a new model. In his interactive compensatory model, Stanovich (1980) argued that:

> interactive models of reading appear to provide a more accurate conceptualization of reading performances than do strictly top down or bottom up models. When combined with an assumption of compensatory processing (that a deficit in any particular process will result in a greater reliance on other knowledge sources, regardless of their level in the processing hierarchy), interactive models provide a better account of the existing data on the use of orthographic structure and sentence context by good and poor readers (as cited in Samuels & Kamil, (1984) p. 212).

According to Stanovich's model, reading involves a number of processes. Readers who are weak in one process will rely on other processes to compensate for the weaker one. For example, a poor reader who is slow and inaccurate at word identification (bottom up) but knows a lot about the text subject will overcome his or her weakness by relying on his or her knowledge (top down). Thus, the Stanovich model is interactive in the sense that any process, regardless of its position, may communicate (interact) with any other process in order to achieve comprehension. “However, the evidence that such compensation does in fact occur is controversial” (Alderson, 2000, p. 19).

Perfetti (1985) proposed the “verbal efficiency theory”, which consists of local text processes and text modelling processes. The local text processes are restricted in their interaction.

> Its interactions are restricted to occur only within the specific data structure of lexical formation (i.e. letters, phonemes, and words). It allows no influences from outside lexical data structures, no importation of knowledge, expectations, and beliefs. Skilled word recognition is context-free (Perfetti, 1991, p. 34).

The second part of the verbal efficiency theory, text modelling processes, is more interactive. The reader can make use of his or her background knowledge of the world in order to understand the text.

The verbal efficiency theory works as a theoretical framework for understanding the nature of individual differences in reading comprehension ability. Perfetti’s verbal efficiency theory, in its general form, claims that “individual differences in reading comprehension are produced by individual differences in the efficient operation of local processes. … This is not to say that comprehension differences are not also produced by schema-related processes” (Perfetti, 1985, p. 100).

According to Perfetti (1985), the essence of verbal efficiency theory is that local processes play the main role in reading comprehension. The theory involves specific cognitive processes such as working memory and attention. Working memory helps in comprehending sentences: it stores the information from previous sentences, or even part of a sentence, and makes it available to understand the following word, sentence, or paragraph. Attention also helps in comprehension: without it, memory will not work properly, which affects reading comprehension indirectly. Less attention is needed when reading an easy or familiar subject. The amount of attention depends, also, on the reader's processing skill, which can be improved by learning and practice. Verbal efficiency theory argues that reading comprehension will not be optimal until word decoding is efficient. Therefore, besides teaching high levels of reading comprehension skills such as making predictions and assumptions, teachers should also teach ESL students how to decode quickly and accurately.

Like Stanovich compensatory model, Perfetti (1985) argues that "weaknesses in decoding may be compensated for, to some extent, by other comprehension processes" (p. 236) because "reading comprehension is the result of processes that operate on many different levels" (p. 233).

Rumelhart (1977) proposed an interactive reading model, which starts with a 'visual information store' as a first stage in the reading process. The important features are then selected by a cognitive 'feature extraction device'. After that, a 'pattern synthesizer' comes up with the most probable hypothesis of the text based on the selected important features combined with linguistics and world knowledge. These multiple knowledge sources continue interacting with each other until a final acceptable hypothesis is reached. The reading process in the Rumelhart model is both perceptual and cognitive (Alderson 2000, Hudson 2007).

Another interactive reading model was proposed by Just and Carpenter (1980). They focused on eye-fixation, which is considered a major difference between reading and listening, because the reader controls the amount and speed of input when reading a text. They also claim that the reader reads almost every single word in the text, which is different from Goodman’s view that the reader samples the text to make the right guesses. They say:

> Almost every content word is fixated at least once. There is a common misconception that readers do not fixate every word, but only some small proportion of the text, perhaps one out of every two or three words. However, the data … show that during ordinary reading, almost all content words are fixated (Just and Carpenter, 1980, p. 329 - 330).

They add, however, that the fixation time on words varies considerably. A shorter time is needed for function words, while unfamiliar words usually require a longer time. “Readers make longer pauses at points where processing loads are greater. Greater loads occur while readers are accessing infrequent words, integrating information from important clauses, and making inferences at the end of the sentence” (p. 329). The fixation continues as long as the reader processes the word. The more difficult is the word, the longer the fixation time. The interaction happens among all levels simultaneously until the correct meaning is reached.

Person and Tierney (1984) proposed a reading/writing model. This interactive model is based on the interaction between the reader and the author. “The thoughtful reader … is the reader who reads as if she were a writer composing a text for another reader who lives within her” (Person and Tierney, 1984, p. 144). Therefore, when writing, authors assume that readers will compose meaning from the text, and at the same time, readers read the text with the expectation that it has enough clues and information to convey the author’s intended message. Comprehension happens when the reader succeeds in decoding what the author wants to convey. Therefore, the reader is seen as a composer and not as a mere reciter.

The reader plays four different roles, namely planner, composer, editor, and monitor. In the first stage, the reader positions him/herself towards the text and its content and style. Then, as a composer, the reader looks for key words and ideas as well as the organisation of the text. After that, the reader, as an editor, works as an examiner in order to make sure that he/she has reached the correct propositions. Finally, the reader acts as a monitor over all the previous three roles, and coordinates the work among them.

> The model focuses on the thoughtful reader with the four interactive roles of planner, composer, editor, and monitor. As a planner, the reader creates goals, mobilizes existing knowledge, and decides how to align him/herself with the text…. As a composer, the reader searches for coherence, often needing to fill in gaps with inferences about the relations within the text…. In the role of editor, the reader stands back and examines his or her developing interpretations…. Simultaneously with the above three roles, the reader acts as an executive or monitor. This monitor role directs the three previously mentioned roles, deciding which particular role should dominate at any particular moment in the reading process (Hudson, 2007, p. 49-50).

Interactive models of reading comprehension have received some criticism. According to Urquhart and Weir (1998), such models are valid only for careful reading, and are not always applicable when English language teachers deal with different kinds of readers. Therefore, more elaboration is needed to explain and describe other types of reading. Eskey & Grabe (1988) argue that models of reading have little to offer because they are "models of the "ideal," completely fluent reader with completely developed knowledge systems and skills; whereas the second language reader is, almost by definition, a developing reader with gaps and limitations in both of these categories" (p. 227). Moreover, in their view, these models do not explain clearly how readers use both bottom-up and top-down processes to achieve comprehension; especially when we talk about specific readers.

## 2.3 Variables Affecting Reading Comprehension

Alderson (2000) argues that reading comprehension is a complex cognitive process that might be affected by different variables. Based on their source, these variables can be divided into two main groups. The first group of variables are related to the reader, while the second main group of variables are related to the text. For the purpose of this review, I propose a third group of variables that is related to testing.

The following is a brief overview of these three groups of variables that may have an effect on the reading process. I will start with the reader variables and their direct and indirect relation to reading comprehension.

### 2.3.1 Reader Variables

The reader plays an important role in the process of reading. He or she is no longer seen as a passive decoder of the written symbols, but as an active meaning-constructor in the reading process. The reader especially in the top-down approaches contributes more than do “the visual symbols on the page” (Grabe, 1991, p. 377). However, readers differ in their level of contribution to the reading process because they “vary by shared background knowledge, language skills, strategies … and ‘other personal characteristics’” (Alderson, 2000, p. 128). The following is the first reader variable.

#### 2.3.1.1 Schemata and background knowledge

Carrell (1987) distinguishes between two types of schemata. According to her,

> one type of schema, or background knowledge, a reader brings to a text is a *content schema*, which is knowledge relative to the content domain of the

> text. Another type is a *formal schema*, or knowledge relative to the formal, rhetorical organizational structures of different types of texts (p. 461).

Alderson (2000) argues that content schemata can be divided into general knowledge and subject-matter knowledge. The general knowledge is not necessarily relevant to any particular text, while the subject-matter knowledge is directly related to the topic of the text. Formal schemata and subject-matter knowledge can interact: readers who are weak in one of them will rely on the other to compensate for their weaknesses. Furthermore, Alderson argues that “tests based on texts which are too specialised might test subject matter knowledge rather than reading ability” (p. 102-103)

Several studies have found that the background knowledge of the text has a positive effect on levels of performance in reading comprehension tests. Test takers are generally found to perform better whenever they are introduced to subject areas with which they are familiar. (Alderson, 2000; Bachman, 1990; Grabe, 1991; Kitao & Kitao, 1999; Tierney & Cunningham, 1984). However, Clapham (1998) found in a study of 842 non-native speakers of English that background knowledge did not increase levels of reading comprehension.

According to Kitao & Kitao (1999), being either very familiar or completely unfamiliar with the text can adversely affect performance in a reading comprehension test. When test takers are very familiar with the text subject matter, then they may answer questions without comprehending the text. At the other extreme, fluent readers may not perform well if they have no background knowledge at all of the text.

#### 2.3.1.2 Cultural schema

Another important factor is cultural knowledge. Cultural differences may affect readers’ comprehension in a negative or positive way. For example, when an English reader reads about the funeral or marriage ceremonies in India or parts of Africa, his/ her level of understanding may be affected by his/ her cultural schema. Johnson (1981) conducted a study on 65 American and Iranian students to investigate the effect of the cultural origin of folklore stories on reading comprehension. She found that “the cultural origin of the story had more effect on the comprehension of the ESL students than the level of syntactic and semantic complexity” (p. 169). Rosowsky (2000) investigated the influence of culture on reading comprehension of bilingual students. He found that “there is a strong link between reading comprehension performance and cultural bias … An awareness of cultural schemata on the part of teachers is therefore necessary if we do not want to confuse poor reading comprehension with cultural confusion” (p. 50). Furthermore, in a study by Pritchard (1990) on 60 11th-grade students, he examined the influence of cultural schemata on students' processing strategies and reading comprehension. He found that students’ level of comprehension was affected by the cultural schemata, and that “students recalled significantly more idea units and produced more elaborations, as well as fewer distortions, for the culturally familiar than for the unfamiliar passage (p. 273).

#### 2.3.1.3 Formal schemata

Chang (2006) conducted a study on forty American college third-year students who study Chinese as a Foreign Language. She investigated the effects of linguistic difficulty and topic familiarity on the reading strategies and mental representations. The results showed that the topic familiarity was “found to have a facilitative effect on the mental representations of the reading passages whereas no effects due to linguistic difficulty was found” (p. 172).

Kaivanpanah & Alavi (2008) examined the role of grammatical knowledge in inferring the meaning of new vocabulary items. There was enough evidence to suggest that grammar knowledge has a clear influence on inferring the meaning of new words. Khaldieh (2001) examined the role of the knowledge of vocabulary in reading comprehension; forty-six American learners of Arabic as a foreign language participated in the study, and he found that vocabulary knowledge had a major positive effect on the level of reading comprehension.

#### 2.3.1.4 Gender

Phakiti (2003) investigated gender differences in the use of cognitive strategy in a reading comprehension test of English as a foreign language. Three hundred eighty four Thai university students participated in the study. Results showed that there are no gender differences among male and female students in their reading comprehension performance and their use of cognitive strategies. In addition, Powers (1995) compared the performance of three hundred thirty five male and female students. The objective of this unusual test-like task was to “uncover any gender differences in approaches to and performance on a task requiring examinees to answer reading comprehension questions without reading the passages on which the questions were based” (p. 1). The findings showed that there were almost no gender differences between male and female students.

Brantmeier (2003) examined the effects of readers' gender and passage content on the second language (L2) reading comprehension of seventy-eight university-level students. “Findings reveal significant interactions between readers' gender and passage content with comprehension …. (They) provide evidence that subject matter familiarity has a facilitating effect on second language (L2) reading comprehension by gender” (p. 1).

In a study by Bugel & Buunk (1996) on 2980 secondary school students who studied English as a foreign language for a minimum of three years. They examined the impact of text topic on gender differences in foreign language reading comprehension. Results showed that gender differences in foreign language reading comprehension tests are affected by the topic of the text. Male students scored significantly higher on topics about volcanoes, cars, laser thermometers, and football players, while female students scored significantly higher scores on text topics about midwives, a sad story, and a housewife’s dilemma.

Although it can be argued, based on research results, that there are usually no gender differences in performance levels in reading comprehension, research has shown that text topics can make a gender difference. This can be supported by other research which shows that text familiarity has a positive effect on reading comprehension ( Alderson, 2000; Bachman, 1990; Chang, 2006; Grabe, 1991; Johnson, 1981; Kitao & Kitao, 1999; Pritchard, 1990; Rosowsky, 2000; Tierney & Cunningham, 1984).

### 2.3.2 Text Variables

The second group of variables affecting reading comprehension besides the reader relate to the text. Text variables might facilitate the reading process or might make it a difficult one. They include variables such as text layout, the language of the text, text structure, and text type.

### 2.3.2.1 Text layout

Several studies have shown that text layout has a positive effect on levels of reading comprehension. Lorch, Jr., Lorch and Klusewitz (1995) investigated the effect of typographical cues like capitalization, boldface, underlining, italics and colour variation on text memory. Eighty college students participated in the study. Their results showed that when the reader encounters "signaled information, the reader attends more carefully to the content…; processing of unsignaled content is unaffected…. The increased attention to the signaled content results in better memory for the signaled content, whereas memory for the unsignaled content is unaffected" (p. 63).

Moreover, Lonsdale, Dyson and Reynolds (2006) examined the effects of text layout as a whole on the speed and accuracy of a reading task in an examination-type situation. Thirty undergraduate and postgraduate students participated in the study. With the layout conforming to legibility guidelines, the results of the study showed that shorter time was needed to finish the task, and a higher number of correct answers was achieved. Moreover, students pointed out that it was easier for them to find answers in this layout. Lonsdale et al. (2006) emphasised the importance of using legibility guidelines in tasks requiring the reader to scan the text for specific information under time pressure similar to examination-type situation.

Hsu and Yang (2007) compared two texts in Chinese on the moon phase with different print and image integration. The main differences between the two layouts were “ the structure of print (e.g., ambiguity of words, technicality), structure of image (e.g., representational structure, modality, salience), and interaction of print and image (e.g., starting point and reading path implied in texts)” (p. 656). One hundred thirty-two junior secondary school students took part in this study. Results showed that students who read the new-layout text scored significantly higher, and their comprehension was better than students who read the traditional text. In addition, students who read the traditional text faced greater difficulty in making sense of the images than the students who read the new-layout text. These results might help in solving the problem of explaining difficult-to-learn science concepts.

#### 2.3.2.2 Text structure

Knowledge of text structure is essential for reading comprehension. Identifying the organizational structures of texts promotes readers’ speed and accuracy in reading comprehension. Carrell (1985) conducted a study on twenty-five intermediate-level ESL students to find out whether explicit training of text structure could promote ESL reading comprehension. She found that teaching text structure considerably increased the amount of recalled information by intermediate-level ESL students. Furthermore, in a similar study by Armbruster, Anderson, and Ostertag (1987), they investigated the effect of text structure training on reading comprehension. A total of eighty-two fifth-grade students participated in the study. The results suggest that explicit training in text structure helped students in their reading processes.

Moreover, Geva (1983) carried out a study to investigate the effect of flowcharting on the reading process. Forty-eight first-year college students participated in this study. The results suggested that using flowcharting to identify text structure led less skilled readers to more careful reading of expository texts.

### 2.3.2.3 Language of the text

Yano, Long, and Ross (1994) conducted a study on 483 university EFL Japanese students. They investigated the effects of simplified and elaborated texts on foreign language reading comprehension. Even though text simplification and elaboration can make text comprehension easier for non-native speakers of English, nevertheless there were no differences between readers of simplified and elaborated texts based on the results of reading comprehension tests of passage content. Yano et al. (1994) emphasised that “elaboration appears to serve the twin functions of most foreign and second language reading lessons: (a) improving comprehension and (b) providing learners with the rich linguistic form they need for further language learning” (p. 214). In another similar study, Oh (2001) studied the effect of simplification and elaboration of the text on reading comprehension. One hundred eighty Korean students participated in the study. Results showed that text simplification facilitated reading comprehension for only high proficiency students, while text elaboration helped both low and high proficiency students. Moreover, when comprehension levels of both simplified and elaborated texts were compared, there was no difference for both low and high proficiency students. Oh (2001) recommended that “input should be modified in the direction of elaboration rather than by artificial simplification, because elaboration retains more nativelike qualities than and is at least equally successful as- if not more successful than- simplification in improving comprehension” (p. 69).

McNamara, Kintsch, Songer, and Kintsch (1996) examined the effect of coherence on the comprehension of science texts. They found that coherent text helped readers with little background about the subject matter, while readers who know a lot about the domain of the text relied less on its coherence.

### 2.3.3 Testing Variables

Just as reader and text variables affect reading comprehension, testing variables represent an essential part in the reading process. In testing, it is very important to minimize the effect of these factors in order to maintain test validity and reliability.

#### 2.3.3.1 Testing method

Shohamy (1984) argues that "traits which are evaluated indirectly put a heavy burden on the testing method and therefore may create greater variations in the scores obtained as a result of these methods" (p. 149). Bachman (1993) believes that the testing method has an obvious effect on the performance of the test taker. He said "indeed, one of the major findings of language testing research over the past decade is that performance on language tests is affected not only by the ability we are trying to measure but also by the method we use to measure it." (p. 185). Moreover, Alderson & Urquhart (1988) argue that there is a clear effect of the testing method on student's performance in a reading comprehension test. In other words, the test score might be the result of the testing method rather than the skill being measured. If this is true, then it is important to know which testing method has the least or the most effect on the test taker? In addition, how can these effects be reduced? Do testing methods affect performance in all the four skills (reading, writing, listening, and speaking) equally or differently? Moreover, is there a best way to test each skill? Alderson (2000) believes that “no single test method can fulfil all the varied purposes for which we might test” (p. 203). Each method has its own advantages and disadvantages. To reduce the effect of testing methods, test developers should pay special attention to test specifications, and are encouraged to use more than one method in testing each skill whenever possible to minimize the effect of the testing method.

#### 2.3.3.2 Language of the test

Although there has been much work on factors that might affect testing reading comprehension, there has been very little interest in the effect of using L1 in testing L2 reading comprehension. “The language chosen in L2 reading comprehension tests has been the focus of attention of only a few studies” (Martinez & Godev, 1994, p. 2). According to Bernhardt (2005), one reason for the popular use of L2 in testing L2 reading comprehension is that researchers do not know the language of the subjects in their studies; therefore they (researchers) use their own. “If readers are assessed in comprehension tasks in their stronger language (almost always L1 until the highest proficiency/fluency levels), their comprehension seems to be much more significant than when it is measured within the context of their impoverished second language skills. … The field will not progress until researcher deficiencies no longer interfere with the ability to provide solid and trustworthy data” (Bernhardt 2005, p. 141).

Shohamy (1984) found “that the testing methods – MC, OE, and language of questions (L1, L2) – can make a difference in the assessment of the trait and can affect the scores that students obtain on reading comprehension tests” (p. 157). She pointed out that the use of the native language lowered the test difficulty and the anxiety of the test takers only for low-level students; and that the L1 wording might give clues to the test takers. Shohamy conducted her study on 655 twelfth-grade students in Israel whose native language was Hebrew. She investigated the effect of the testing method on testing reading comprehension. In particular, she examined multiple-choice (MC) and open-ended (OE) questions, and added the language factor to them. Questions were written in English (L2) and then translated into Hebrew (L1). The OE questions were the same as MC questions but without the distractors. In both English and Hebrew questions, answers were always in English (L2).

Hamdan & Diab (1997) found similar results showing that the use of the native language in testing reading comprehension in English (L2) helped only the low-level students. They conducted their study on sixty secondary school students in Jordan whose native language was Arabic. In their study, Hamdan & Diab developed an English short-answer reading comprehension test, and then translated the questions into Arabic (L1). Based on their results in a reading comprehension cloze test, students were divided into two sub-groups, equal in size and level of proficiency. The first sub-group took the English version of the test, and their answers were in English. The second sub-group took the Arabic version of the test, and their answers were in Arabic.

Martinez & Godev (1994) examined the use of the target language (Spanish) in writing the questions and answers of reading comprehension tests. They found that errors in the students' answers "stemmed from the combined factors of not understanding the wording of the questions and from not being able to word the answers in Spanish" (p. 14). Moreover, based on their results, they argue that the use of L2 in assessing L2 reading comprehension will reduce test validity. In their study, forty-six undergraduate students at George Mason University, USA, participated in the study. Each student was given one reading passage in Spanish and two sets of short-answer questions. The first set of questions were written in Spanish, and students were asked to answer each question twice, in Spanish and then in English. After collecting the first set of questions, students were given the second set, which contained the same questions as in the first set, but were written in English, and students were asked for the second time to answer them in English and then in Spanish. So, all students answered the same question twice in English (L1) and twice in Spanish (L2). However, the repetition of answers may raise questions about the validity of the test and consequently the scores obtained from it.

## 2.4 L2 Reading Sub-skills

Identifying reading sub-skills is a major concern in second language reading research (Hudson, 2007). "It is commonplace in theories of reading to seek to identify skills which underly or contribute to the reading process" (Alderson, 1990, p.425). According to Alderson and Lukmani (1989), many lists of reading sub-skills have been assembled as a result of the work of researchers who have long tried to identify reading sub-skills by "giving subjects a series of passages and asking them questions intended to test different levels of understanding of the passages. The answers to these questions are then subjected to factor analysis, in order to see whether the different questions measure different ‘subskills’” (p.255). These lists of sub-skills have been drawn up by both first and second language researchers based on the assumption that a number of different sub-skills are included in the reading comprehension process. (Gomez, Noah, Schedl, Wright, and Yolkut, 2007).

Since the late 1970s, L2 students were taught a variety of reading sub-skills or strategies in order to improve their comprehension ability (Barnett, 1988). Grabe (1991) stated that L2 students “needed to be taught strategies to read more efficiently (e.g., guess from context, define expectations, make inferences about the text, skim ahead to fill in the context, etc.)” (p.377). According to Barnett (1988), reading strategies are the comprehension processes which readers use to comprehend what they are reading. These processes may include skimming, scanning, guessing word meanings from context, etc. She argues that using efficient reading strategies will help L2 students to understand more than those who do not. In addition, these taxonomies of reading “have been widely used not only for reading syllabus design..., but also for the development of test specifications and items in EFL/ESL reading assessment” (Lee, 2004, p.79).

The distinction between reading ‘skills’ and reading ‘strategies’ is not always clear, and they are sometimes used interchangeably. “It is not always easy to distinguish skills from strategies” (Hudson, 2007, p.77). According to Rupp, Ferne and Choi (2006), ‘skills’ are unconscious automatic abilities that facilitate comprehension such as text decoding and the use of background knowledge, while ‘strategies’ are conscious techniques which are used intentionally for successful reading such as skimming and scanning.

### 2.4.1 Separability and Hierarchy of Sub-skills

The notion of dividing reading ability into different sub-skills is common in ESL teaching and assessment (Lumley, 1993). Alderson (2000) points out that “the notion of skills and subskills in reading is enormously pervasive and influential, despite the lack of clear empirical justification” (p.10). Song (2008) stated that there are at least three different views with regard to the separability of reading sub-skills.

> “The first view is that there is little clear evidence that distinct separate subskills exist, so reading is a unitary, integrated skill.…The second view is that reading is divisible into subskills, even though there is no consensus on how many subskills might be empirically identifiable…. The third position is that reading is essentially divided into two processes: decoding (word recognition) and comprehension” (p. 438).

The hierarchy of sub-skills is also a common notion in reading literature, and there have been many attempts in second language reading research to develop sub-skill hierarchies (Hudson, 2007). Alderson (1990) argues that “there are serious reasons for doubting whether a skill can be said to be “higher” or “lower” than another skill in any hierarchy that implies relative difficulty or some differential stage of acquisition (at least for ESL readers)” (p.436). According to Hudson (2007), in general, the results of both the first and second language reading research do not support the existence of strictly hierarchically ordered sub-skills.

According to Hudson (2007), the work of Munby (1978) reflects the research into the separability and hierarchy of L2 reading comprehension sub-skills. Munby’s list “has had enormous influence in the area of the teaching and testing of English as a Foreign Language” (Alderson and Lukmani, 1989, p.256). Munby distinguishes the following reading microskills (as cited in Alderson, 2000, pp.10-11):

> recognising the script of a language
> deducing the meaning and use of unfamiliar lexical items
> understanding explicitly stated information

understanding information when not explicitly stated
understanding conceptual meaning
understanding the communicative value of sentences
understanding relations within the sentence
understanding relations between parts of text through lexical cohesion devices
understanding cohesion between parts of a text through grammatical cohesion devices
interpreting text by going outside it
recognising indicators in discourse
identifying the main point or important information in discourse
distinguishing the main idea from supporting details
extracting salient to summarise (the text, an idea)
extracting relevant points from a text selectively
using basic reference skills
skimming
scanning to locate specifically required information
transcoding information to diagrammatic display

Munby developed his list of reading sub-skills to help teachers and curriculum designers by providing them with a comprehensive list of sub-skills from which they can choose what suits students' levels and needs (Hudson, 2007). Munby's list has not escaped criticism for various reasons including "the level of conjecture it relies upon and its lack of an empirical base, as well as its impracticality. Its impact, however, in the area of needs analysis, an important aspect of syllabus design, has been considerable" (Lumley, 1993, p.212).

Alderson and Lukmani (1989) investigated the questions regarding the existence of the separability of reading sub-skills and the notion of hierarchy of those sub-skills according to their level of cognitive ability: lower, middle or higher order sub-skills. Nine experienced teachers at the University of Lancaster were given the task of examining a reading comprehension test used at the University of Bombay. They were asked to determine what each test item was testing, and then had to classify each item according to whether it was measuring lower, middle or higher order sub-skills. After that, the teachers were given a list of reading sub-skills and were asked to indicate which of the sub-skills were tested by each item in the test. The results showed that "for 27 out of 41 items there was very little agreement on the levels being tested. In addition, for approximately the same number of items judges disagreed considerably over the skills being tested" (Alderson and Lukmani, 1989, p.263). In contrast with these findings, Lumley (1993) reported a study by Brutten, Perkins and Upshur (1991) who found "a high level of agreement between four raters about the skills tested by individual test items, using the Iowa test of basic skills taxonomy of reading skills" (Lumley, 1993, p.216). This finding goes with the present study where all the five English-language teachers in the panel have reached a consensus on the sub-skills tested by each test item (more details on this can be found in chapter five). In the same vein, Lumley (1993) examined the notion of sub-skills and their difficulty in reading comprehension tests. Based on the perception of five experienced ESL teachers, Lumley has found that it is possible for a group of teachers to reach a high level of agreement on the sub-skills tested by each item in the test.

The following five sub-skills were used in the present research. From my experience as an English-language teacher, these sub-skills are among the most commonly used sub-skills in teaching and assessing reading comprehension in Saudi schools.

### 2.4.2 Skimming and Scanning

According to Brown (1994), skimming and scanning are "the two most valuable reading strategies for learners" (p.293). Skimming means reading quickly through a text to get a general picture of it (Ionescu, 2008), while the second sub-skill, scanning, refers to the ability to search quickly to extract a specific piece of information (e.g. names, dates, ..) without reading the whole text. It enables the reader to predict the purpose of the text and its main topic or message (Brown, 1994). Scanning needs less attention to the text than skimming does (Nuttall, 2005). Duggan and Payne (2009) carried out a study on the process and effectiveness of text skimming under time pressure. Their findings have shown that skimming is an effective sub-skill that can be used to grasp the main points of a text. The study has shown that when readers are under time pressure, "they are indeed able to gain a greater understanding of a text by skimming rather than by reading linearly through an imposed half of the text" (p. 240). Although both skimming and scanning are valuable reading sub-skills, "they do not remove the need for careful reading, but they enable the reader to select the texts, or the portions of a text, that are worth spending time on" (Ionescu, 2008, p.149).

### 2.4.3 Making Inferences

Norvig (1989) defines an inference as "any assertion which the reader comes to believe to be true as a result of reading the text, but which was not previously believed by the reader, and was not stated explicitly in the text" (p.569). Inference making is considered to be an essential part of skilled reading (Oakhill, Barnes and Bryant, 2001). Cain and Oakhill (1999) believe that, for skilled readers, inference making is an important part of the process of integrated and coherent representation of the text. According to Zwaan and Brown (1996), several researchers have argued that the ability to make inferences is reflected in the differences between skilled and less skilled comprehenders.

The relation between inference making and comprehension skill is well established (Cain and Oakhill, 2006; Cain and Towse, 2008; Haenggi, Gernsbacher and Bolliger, 1994; Oakhill, 1993; Oakhill, Barnes and Bryant, 2001). Casteel (1993) states that “the ability to generate inferences is an important skill that determines, to a large extent, the degree to which a passage will be understood” (p.346). According to Cain and Towse (2008), “Poor reading comprehenders’ language processing difficulties extend to many of the skills essential for adequate text comprehension, such as inference generation” (p.1539). Oakhill, Barnes and Bryant (2001) examined the relation between comprehension skill and inference-making ability and found a strong relation between them even when knowledge was equally available to all participants. The less skilled readers made considerably fewer inferences than the skilled readers did.

### 2.4.4 Pronoun Resolution

According to Kennison (2003), pronouns are among the most commonly used words in English, and the processes used in pronoun resolution are among the most commonly used comprehension processes. In the process of pronoun resolution, the reader usually uses semantic, syntactic, and discourse information to choose the suitable coreferent for a pronoun (Matthews and Chodorow, 1988). Similarly, Pretorius (2005) argues that “there are several syntactic, semantic, textual, and pragmatic variables that affect anaphoric resolution. Three factors in particular have been shown to influence anaphoric resolution: ease of antecedent identifiability, topic continuity/discourse focus, and distance between anaphor and antecedent” (Pretorius, 2005, p.524).

Understanding pronoun referent is considered to be an essential component of reading comprehension (Al-Jarf, 2001; Badecker and Straub, 2002; Crawley and Stevenson, 1990; Demel, 1990; Huang, 2005; Pretorius, 2005; Wolf, Gibson, and

Desmet, 2004). In a study by Berkemeyer (1994) on anaphoric resolution and text comprehension, she found a significant positive relationship between overall text comprehension and coreferential tie comprehension for readers of German. In another study by Demel (1990), she investigated the relationship between overall reading comprehension and the comprehension of coreferential ties for second language readers of English. She found that there is a significant relationship between the comprehension of coreferential ties and overall comprehension for L2 readers. Demel (1994) investigated the relationship between overall comprehension and coreferential tie comprehension for second language readers of Spanish literature. She found that there are significant correlations between coreferent identification and overall comprehension for both beginning and advanced levels. Furthermore, Huang (2005) investigated the relationship between referential understanding and academic reading comprehension among EFL college students. The results showed a significant relationship between them.

According to Sweet and Snow (2003), skilled readers need to figure out the coreferent whenever there is a pronoun in the text. To comprehend a text, it is more important to understand the relationships between propositions in that text than just understanding each single sentence in it (Huang, 2005). Therefore, Rose (2010) states that "it is no surprise then, that pronouns get early treatment in most English language teaching curricula" (p.1).

### 2.4.5 Guessing the Meaning of Unknown Words

L2 learners are usually exposed to materials that are meant for native speakers in which they may encounter many unfamiliar words. According to Zhang and Annual (2008), "if a text contains too many difficult words for the students….

comprehension will diminish even if the text is highly cohesive." (p.61). For L2 readers, vocabulary knowledge is closely tied to reading comprehension (Cain, Oakhill, and Lemmon, 2004; Pearson, Hiebert, and Kamil, 2007; Zhang and Annual, 2008). When a reader is faced with an unknown word, he or she usually looks it up in a dictionary, guesses its meaning or ignores it. According to Kaivanpanah and Alavi (2008), it is most common for readers to guess the meaning of the unknown word. The ability to guess the meaning is affected by the following four textual variables: (1) the characteristics of the word itself (2) the level of text difficulty (3) the presence of contextual clues (4) topic familiarity (Kaivanpanah and Alavi, 2008).

Guessing the meaning of unfamiliar words is considered to be an essential sub-skill for reading. Aebersold and Field (1997) consider it "the most useful vocabulary skill that readers can have"(p.142). Students who can use the context to guess the meaning of unknown words "have a powerful aid to comprehension and will ultimately read more quickly" (Nuttall, 2005, p.72). Readers' accuracy in guessing the meaning is closely related to their proficiency level. The higher the level the more accurate the guesses are (Kaivanpanah and Alavi, 2008). Cooper (1984) adds that "practised readers are distinguished from unpractised readers by their ability to use the whole context to decode the meaning of unfamiliar words" (p.128). Therefore, many L2 reading textbooks include tasks that require learners to use the context to guess the meaning of unknown words.

# CHAPTER THREE: LANGUAGE TESTING

## 3.1 Introduction

The previous chapter discussed reading models and the main variables that might have an effect on reading comprehension and assessment. These variables differ in the way and degree they affect the test taker. Test developers and users usually pay special attention to these variables in order to reduce their effect on the level of performance and therefore increase test validity and reliability. In this chapter, I will discuss some issues and considerations in language testing especially validity and reliability and test specifications.

Numbers play an important role in research. In language testing, for example, numbers (i.e., test scores) give a clear representation of the issue or problem with which we are dealing. Nevertheless, test scores are not always a dependable source of information. Certain fundamental considerations must be met before we may make inferences based on test scores. The two main considerations are reliability and validity.

## 3.2 Reliability

Reliability refers to the degree of consistency of test results (Richards, J.C., Platt, J., & Platt, H., 1992). It deals with different kinds of questions such as: Would we obtain approximately the same results if the test was given in a different time or place? Would we obtain the same results if we used different samples of the same test? Would different scorers rate the performance of all test takers in the same way? Maintaining and promoting reliability is not always an easy task. “How we conceptualise and operationalise reliability is problematic, especially in the light of what is known about variation in language performance” (Alderson and Banerjee, 2002, p. 101). Reliability is a condition for validity. A valid test must be reliable but a reliable test is not necessarily valid (Alderson, Clapham, and Wall, 1995; Bachman, 1990). It is possible for a test to be reliable but not valid, for example “a written multiple-choice test of pronunciation which is highly reliable but which fails to identify students whose actual pronunciation is good or bad” (Alderson et al., 1995, p. 187).

The reliability of test scores is usually expressed in the reliability coefficient ($r_{tt}$) and the standard error of measurement (SEM), which is derived from the reliability coefficient. The reliability coefficient can go as high as +1 for a perfectly reliable test and as low as 0 for a test that is useless (Hughes 2003). For example, a reliability coefficient of .85 indicates that 85 percent of the variation in the observed scores can be attributed to the true scores variation and the remaining 15 percent is due to error.

The reliability coefficient can be estimated in a number of ways; one of them is internal consistency, which is estimated by conducting the test once then computing the consistency of the response within the test. The odd-numbered and even-numbered items on the test are scored separately. These two halves are calculated as if they were two different versions of the same test. The internal consistency can be estimated by different methods. In the present study, the reliability coefficient is estimated by using both Spearman-Brown prophecy formula and the Guttman split-half estimate which does not require calculating the correlation coefficient between the two halves because it is based on the assumptions that the test halves are independent and not equal. It provides a direct estimate of the reliability of the whole test.

The standard error of measurement:

All scores are only estimates, and they are not true scores. Therefore, the standard error of measurement is used to estimate the true score. “Our tests are not perfectly reliable, and errors of measurement - error scores - cause observed scores to vary from true scores…. The more reliable the test is, the closer the obtained scores will cluster around the true score mean, resulting in a smaller standard deviation of errors” (Bachman, 1990, p. 198-199).

The standard error of measurement (SEM) is derived from the reliability coefficient and it shows how many points must be added to or subtracted from an individual’s test score. For example, if a test taker obtained 84 on a test and the SEM for this test is 4, then we can conclude that the student would score between 80 and 88.

### 3.2.1 Making Tests More Reliable

Hughes (2003) mentioned a number of ways to maximize test reliability. In the following section, some of these ways and methods are discussed.

1. “Take enough samples of behaviour”: It is not possible to determine how reliable a test is by only having the examinees to answer one or two items because that may not reflect the ability being measured. Therefore, test developers usually increase the number of items in the test; the higher the number of items the higher the reliability of the test. However, it is important that the added items are not a sort of repetition of the main ones. They should be independent and add value to the test.

The empirical research proved that adding more independent items to the test would promote its reliability (Hughes, 2003). However, this positive effect should not persuade us to make tests longer than they should be. Balance is needed; otherwise, test validity might be affected.

2. “Exclude items which do not discriminate well between weaker and stronger students”: Based on item discrimination analysis, items with low discrimination values should be excluded. An item that does not discriminate between low- and high-level students will not add real value to the test and therefore might affect test reliability. Moreover, in the case of a multiple-choice test, a distractor analysis is performed, and any distractor that does not discriminate well should also be excluded. At the same time, Hughes argues that some of the easy non-discriminating items can be used at the beginning of the test in order to raise confidence and minimise the anxiety of the test taker.

3. “Do not allow candidates too much freedom”: More freedom in the test means more variety in answers, which might have a negative effect on test reliability. In some tests, for example, students are given the choice to answer three question out of five which are not necessarily of equal difficulty, and which might require the use of different skills and strategies. Another example is that student may be asked to write a composition about a general subject, but eventually scorers find themselves scoring compositions that covers completely different sub-subjects. For example, If students are asked to write about ‘safety’, students might write about different sub-subjects such as road safety, fire safety, maintenance safety, and so on. However, if they were asked to write more specific subject such as ‘the importance of using car

seat belts for children", this way will minimise the difference in students' performance, and consequently promote test reliability.

4. "Write unambiguous items": Any ambiguity in the language of the questions might affect the understanding of the examinees. Moreover, questions with multiple unpredictable correct answers will produce different answers because of their ambiguity or because of the different possible unpredictable correct answers they might have, and these answers are not necessarily included in the answer key and therefore might be considered as wrong answers. Such questions and answers will not promote test reliability. Therefore, to avoid these problems, a test draft can be reviewed by a panel of experienced teachers and test developers. In addition, the draft can also be pre-tested and piloted with a small number of students similar to the group who will take the test. Reviewing and piloting the test before its real use will help a lot in solving these problems, and eventually increase test reliability.

5. "Provide clear and explicit instructions": Comprehensible oral and written instructions play an important role in test reliability. Some test writers think that the required task is clear and therefore they do not explain carefully and explicitly what students are required to do. Moreover, oral instructions are also a possible source of confusion; therefore, it is always advised that oral instructions should be read from a written text. Furthermore, Hughes emphasise that, sometimes, unclear or indirect instructions may have a negative effect on high-level students more than low-level students. The abilities of the high-level students might make them think of several different possible answers and eventually they might not answer the question correctly. Therefore, in order to predict and solve such a problem, instructions can be

revised by a panel of experienced teachers and test constructors, and also by piloting the test with a small group of students similar to the intended candidates.

6. “Provide uniform and non-distracting conditions of administration”: Different conditions between one administration and the other might cause difference in students’ performance. Administration conditions include different factors such as place, time, equipments …etc. Thus, it is important to maintain similar conditions in order to have stable results, which means higher reliability.

All the previous ways of maintaining and promoting test reliability are related to the performance of the candidates; however, Hughes (2003) also mentions other ways of promoting reliability that are related to the scorer. The following is a brief discussion of these ways.

1. “Use items that permit scoring which is as objective as possible”: This does not mean using multiple-choice items, but includes different methods such as short answer, cloze test, true/false questions …etc. The more objective the scoring, the higher the reliability that can be reached. However, writing skills, for example, are usually assessed by subjectively scored items. Therefore, taking the other recommendations in this section into account might help to overcome this issue.

2. “Provide a detailed scoring key”: Scoring should not be left to guessing. All acceptable answers should be written in advance to insure that all correct answers are considered correct. Moreover, the assigned points for each item should also be stated from the beginning in a clear way. Using a panel of experienced teachers and piloting the answer key at an early stage can help in identifying any potential problems and help also in refining the scoring key, which eventually leads to a higher reliability for both the scorer and the test.

3. “Train scorers”: Subjective questions are widely used for different reasons, and it is not always possible to have objective questions all the time. Therefore, to deal with this issue, Hughes emphasised the importance of training for scorers, especially the less experienced ones. For example, the same answered paper can be corrected by different scorers, then a panel of all the scorers is held to compare and analyse the scoring process in order to reach an optimal level of scoring. Training is an essential factor in scorer reliability.

4. “Employ multiple, independent scoring”: Reliability can be promoted by using multiple independent scoring especially in subjective testing. For example, in writing tests, the same text may be scored by two separate independent scorers, then a third scorer can examine and compare the two scores in order to reach a more reliable judgment.

### 3.2.2 Source of Variance

Brown (2005) argues that the performance of students on tests can vary for different reasons. He divides the sources of variance into two general types:
(1) variance related to the purpose of the test, which he called ‘meaningful variance’
(2) variance caused by other extraneous sources, which he called ‘measurement error’ or ‘error variance’.

In the second source of variance, which is closely connected to reliability, Brown (2005) divided the potential sources of measurement error into five types, which are: (1) variance due to environment (2) variance due to administration procedures (3) variance due to scoring procedures (4) variance due to the test and test items (5) variance due to examinees. Table 2 summarizes these potential sources of error variance (Brown, 2005, p. 172).

Table: 2

**Variance due to environment**

- location
- space
- ventilation
- noise
- lighting
- weather

**Variance due to administration procedures**

- directions
- equipment
- timing
- mechanics of testing

**Variance due to scoring procedures**

- errors in scoring
- subjectivity
- evaluator biases
- evaluator idiosyncrasies

**Variance attributable to test and test items**

- test booklet clarity
- answer sheet format
- particular sample of items
- item types
- number of items
- item quality
- test security

**Variance attributable to examinees**

- health
- fatigue
- physical characteristics
- motivation
- emotion
- memory
- concentration
- forgetfulness
- impulsiveness
- carelessness
- testwiseness
- comprehension of directions
- guessing
- task performance speed
- chance knowledge of item content

As mentioned above, there are several factors and variables that may have a negative or positive effect on test reliability. According to their source, some of them are not controllable especially those related to the examinees, however it is relatively possible to control the other variables related, for example, to the test itself or the administration procedures and environment, and hence increase test reliability.

## 3.3 Validity

Test validity is the most important issue in language testing. “Since the 1960s, the central location of intense language assessment (and testing) research has been validation” (Kunnan, 1998, p. 1). Validity used to be defined as follows: “does a test measure what it is supposed to measure? If it does, it is valid” (Lado, 1961) as cited in Chapelle, 1999, p. 255). Although validity is considered the cornerstone of any test, a good language test must be both reliable and valid at the same time. Moreover, according to Alderson et al. (1995), “validity is not an all-or-nothing matter…. (it) is relative rather than absolute” (p.170, p. 175).

Construct validity is considered now the overarching concept of validity; however, before discussing this important key concept, I will start with the other aspects of validity: Criterion-Related Validity and Content Validity.

### 3.3.1 Criterion-Related Validity

There are two types of criterion validity: predictive validity and concurrent validity. Both types deal with the relationship between the test in question and a recognized valid test of the same skill.

3.3.1.1 Predictive validity. It “concerns the degree to which a test can predict candidates’ future performance” (Hughes, 2003, p. 29). In the case of the Graduate Record Examinations (GRE), for example, ‘future performance’ is success in a graduate program. This kind of validity is essential for all predictive tests. Predictive validity can be established by giving the test to a number of students who are, for example, starting a program of study and then compare their test performance with some later measures of success such as their grade point average (GPA). If the correlation is high between them, we can affirm that the test has a high predictive validity. However, "In predictive validity studies, it is common for test developers and researchers to be satisfied when they have achieved as low as +.3!" (Alderson et al., 1995, p. 182).

#### 3.3.1.2 Concurrent validity.

It is "established when the test and the criterion are administered at about the same time" (Hughes, 2003, p.27). For example, if the same students who took the reading comprehension test in this study took also a recognized reading comprehension test elsewhere, and the results of both tests were found to be highly correlated, we can say that our test has a high concurrent validity.

### 3.3.2 Content Validity

If the content of the test includes representative samples of the domain to be measured, then the test maintains content validity. To determine content validity, the test's domain needs to be spelled out. According to Hughes (2003), to judge the content validity of a test, "we need a specification of the skills or structures, etc. that it is meant to cover. Such specification should be made at the very early stage in test construction" (p. 26). These specifications should be based on empirical evidence and theoretical bases as we will see later in the next type, construct validity.

Lissitz and Samuelsen (2007), propose a new framework for validity in which they argue that content validity should be the central issue in test validation, and suggest that "an inquiry into the validity of a test should first concern itself with the characteristics of the test that can be studied in relative isolation from other tests, from nomothetic theory, and from the intent or purpose of the testing" (p. 437). Moreover, they argue to separate the validation of the construct validity from the validation of the test, and believe that it is a mistake to concentrate on the construct validity particularly in the educational measurement. Furthermore, they suggest that "content validity, or internal validity, should be acknowledged as the critical initial characteristic to consider when evaluating the quality of a test" (p. 446).

Lissitz and Samuelsen (2007) proposal received some criticism. Sireci (2007) argued that the framework needs revision and more clarifications. In addition, he stated that he

> cannot agree with Lissitz and Samuelsen that we should approach validity independent of the testing context and purpose of the testing…. how would the results be evaluated if not with respect to test specifications designed for a particular purpose? Furthermore, if a test were found to have good content coverage for a specific purpose (e.g., sixth-grade math in Massachusetts), would it still be content valid for another purpose (e.g., adult mathematics literacy)? (pp. 478-9).

In the same vein, Embretson (2007) argues that removing construct validity from the validation process "could have an adverse impact on the quality of educational tests" (p. 449). Moreover, Gorin (2007) suggests that "Lissitz and Samuelsen's conceptualization returns to methods shown historically to be problematic for score use and interpretation" (p. 456).

### 3.3.3 Construct Validity

According to Chapelle (1999), construct validity is considered an essential element in all validation procedures. It refers to the underlying theory of a test (Bachman, 1990; Chapelle, 1999; Hughes, 2003). In addition, it evaluates how well a test measures its constructs. In reading comprehension, for example, some theories state that skimming, scanning, etc are different constructs that should be tested in order to have a valid reading comprehension test (Alderson et al., 1995). "Construct validity concerns the extent to which performance on tests is consistent with predictions that we make on the basis of the theory of abilities, or constructs" (Bachman, 1990, p. 254-255). The concept of construct validity is traced back to an article by Cronbach and Meehl in 1955 "Construct Validity in Psychological Tests" (Tighezze, 2008).

According to Tighezze (2008), many types of validity were proposed in the first half of the 20th century, but the differences among them are not always clear. The following are some of  Tighezze's examples of validity types and names at that time: concurrent validity, predictive validity, correlational validity, content validity, criterion validity, discriminant validity, convergent validity, curricular validity, factorial validity, statistical validity, experimental validity , practical validity, empirical validity, logical validity, incremental validity , intrinsic validity, synthetic validity, …etc.  In 1954, the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) issued the first edition of their standards titled "Technical Recommendation for Psychological tests and Diagnostic Techniques" in which they proposed their classification, dividing validity into four types; content validity, predictive validity, concurrent validity, and construct validity. However in the second edition of the AERA, APA and NCME standards in 1966, they changed these four types into three; content validity, criterion validity, and construct validity. The new type "criterion validity" included predictive validity and concurrent validity. It seems that these four- or three-type classification of validity were widely accepted. That is, they dominated measurement research, teaching and publication from that time till now.

In the third edition of AERA, APA and NCME standards (1974), they made a little but important change in describing validity. Instead of using “types of validity” as in the first and second edition of the standards, they used “aspects of validity”. This important change from “type” to “aspect” reflects the new development in understanding the nature of validity, and reflects also the move towards the unity among validity aspects instead of the previous view of validity as independent types (Tighezze 2008).

Later in 1985, the unitary nature of validity was emphasised in the fourth edition of AERA, APA and NCME standards. They “replaced the former definition of three validities with a single unified view of validity, one which portrays construct validity as central. Content and correlational analyses were presented as methods for investigating construct validity” (Chapelle, 1999, p. 256). In the fourth edition of the standards, validity was referred to as “the appropriateness, meaningfulness, and usefulness of the specific inferences made from the scores” (as cited in Hubley, A. and Zumbo, B. (1996) p. 207). According to this definition, it is not the test itself that we need to validate, but the inferences and their appropriateness, meaningfulness, and usefulness. However, the definition did not talk about the “interpretation” process of scores which leads us to the inferences. The fifth edition of AERA, APA and NCME standards (1999) solved the previous problem and reflected the advancement in understanding the concept of validity. In their definition, validity referred to: “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (as cited in Stobart, G, 2008, p. 4). This definition emphasizes that the correct interpretation of test scores is based on evidence and theory according to test use.

Nowadays, the definition of validity includes the outcome of the analyses and uses of tests as fundamental and inescapable features (Alderson et al, 2002). Thus, test context is highly related to its validity. According to Messick (1980), "Construct validity is indeed the unifying concept that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationship" (as cited in Bachman, 1990, p. 256). Messick (1989) proposed a new definition of validity. He defined it as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (as cited in Sireci, S., 2007). In addition, Bachman (1990) defined validity as "the extent to which the inferences or decisions we make on the basis of test scores are meaningful, appropriate, and useful" (p. 25). The previous views of Messick (1980) (1989) and Bachman (1990) emphasize different aspects. First, validity is a unitary concept, and construct validity is the essential one. Second, validity is associated with test scores not the test itself. Third, the definitions lay emphasis on the consequences of actions based on test scores.

According to Alderson & Banerjee (2002), "it is currently impossible to say exactly what a score might mean. This we might term The Black Hole of language testing" (p. 100-1). Therefore, test users should be very careful in interpreting test scores. Moreover, this argument urges test developers and users to maintain the highest level of both reliability and validity in order to promote scores' accuracy and clarity.

Chapelle (1999) summarised the differences between past and current views of validity in table 3 (p. 258):

Table: 3

| Past | Current |
|---|---|
| Validity was considered a *characteristic of a test:* the extent to which a test measures what it is supposed to measure. | Validity is considered an *argument* concerning test interpretation and use: the extent to which test interpretations and uses can be justified. |
| Reliability was seen as distinct from and a necessary *condition for validity.* | Reliability can be seen as *one type of validity evidence.* |
| Validity was often established through *correlations* of a test with other tests. | Validity is argued on the basis of a number of types of *rationales and evidence,* including the consequences of testing. |
| Construct validity was seen as one of *three types of validity* (the three validities were content, criterion-related, and construct). | Validity is a *unitary concept* with construct validity as central (content and criterion-related evidence can be used as evidence about construct validity). |
| Establishing validity was considered within the purview of *testing researchers* responsible for developing large-scale, high-stakes tests. | Justifying the validity of test use is the responsibility of *all test users.* |

Messick (1995) distinguished six aspects of validity evidence. They are: content aspect, structural aspect, generalizability aspect, external aspect, and consequential aspect. The following is a brief discussion of each one.

1. The content aspect: Test content should be relevant to and representative of the ability being measured. It is based on the experts' analysis and judgment of test content and what it is expected to measure. However, in language tests, according to Bachman (1990), "we seldom have a domain definition that clearly and unambiguously identifies the set of language use tasks from which possible test tasks can be sampled" (p. 245). Moreover, a clear limitation of content relevance is that it does not take into account the actual performance of examinees which may vary

from one situation to another. “Content validity is a test characteristic. It will not vary across different groups of examinees or vary much over time. However, the validity of test score interpretations will vary from one situation to another” (Hambleton and Swaminathan, 1978, p. 38). The content aspect is more concerned with the test rather than the scores. However, in spite of the previous limitations, content aspect is an important part of the validation process, but it is not sufficient by itself (Bachman, 1990).

2. The substantive aspect: According to Messick (1995), the substantive aspect emphasizes the need for two things: “the need for tasks providing appropriate sampling of domain processes … [and] the need to move beyond traditional professional judgment of content to accrue empirical evidence that the ostensibly sampled processes are actually engaged by respondents in task performance” (p. 6). The substantive aspect distinguishes itself with the requirement of empirical evidence, and it does not rely on experts’ judgment as in the content aspect. The evidence for the substantive aspect can be obtained by different methods, for example, think aloud protocols, interviews or questionnaires performed immediately after the test in order to know how they answered each question and what strategies and processes they used to reach the correct answer.

3. The structural aspect: According to Messick (1995), “the theory of the construct domain should guide not only the selection or construction of relevant assessment tasks but also the rational development of construct based scoring criteria and rubrics” (p. 6-7). The higher the consistency between the theory of the domain and the internal structure of the assessment (scoring rubrics and assessment tasks), the better the validity of the test. According to Chaplle (1999), the structural aspect

of validity investigates the consistency of the observed dimensionality of response data and the hypothesized dimensionality of the construct in use. Unidimensional models are usually investigated by several methods such as true-score reliability. However, the problem is that “many language tests are developed on the basis of multidimensional construct definitions” (p. 261), but at the same time the work on multidimensional psychometric models is tentative (Chapelle, 1999).

4. The generalizability aspect: The degree to which score properties and interpretations are applicable and transferable to other groups, settings, and tasks is the essence of generalizability aspect. According to Messick (1995), invistagating the generalizability and the boundaries of score meaning is “meant to ensure that the score interpretation not be limited to the sample of assessed tasks but be generalizable to the construct domain more broadly” (p. 7). The various methods of estimating reliability can be used as an evidence to support the generalizability of test scores (Bachman, 1990).

5. The external aspect: According to Messick (1995), the external aspect of validity refers to “the extent to which the assessment scores’ relationships with other measures and nonassessment behaviors reflect the expected high, low, and interactive relations implicit in the theory of the construct being assessed” (p. 7). The convergent and discriminant evidence supporting the external aspect of validity can be drawn from multitrait-multimethod (MTMM) research design where tests designed to measure the same construct should correlate more highly amongst themselves than with tests measuring other different constructs.

6. The consequential aspect: The final aspect of validity deals with the value implications of score interpretations and with the social consequences of test use.

Messick (1995) argues that the consequential aspect of validity “includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term, especially those associated with bias in scoring and interpretation or with unfairness in test use” (p. 7). Without the consequential aspect of validity, several sources of invalidity will remain unknown and might have a negative effect on the educational measurement practice (Messick 1998). An important concern in regard to any negative effect of the consequential aspect of validity is that it should not be related to other sources of test invalidity which, if not present, test takers will show their ability and competence. (Messick 1995).

The more different evidence of validity can be established, the better the validity of the test (Alderson, Clapham et al, 1995). However, it is important to remember that:

> there is no one best way to validate the inferences to be made from test scores for particular purposes. Rather, there are a variety of different perspectives from which evidence for validity can be accumulated, and thus in a sense, validation is never complete (Alderson et al. 2002, p. 102).

However, Alderson et al (2002) argue that there are some questions remain without an answer, for example “how much evidence is enough? … what to do when the various sources of evidence contradict each other or do not provide clear-cut support for the validity argument” (p. 105).

## 3.4 Test Specifications

Writing test specification is the first step in constructing any test. The clearer and more accurate the test specifications are, the higher the reliability and validity of the test. Test writers use test specifications as a manual to write different samples of the same test. “The specifications are the blue print to be followed by test and item writers, and they are also essential in the establishment of the test’s construct validity” (Alderson et al. 1995, p. 9).

According to Hughes (2003) a typical test specifications will include the following specifications (see p.62 for the test specifications of the present study):

Statement of problem: By stating the problem, we are describing the purpose of the test and deciding what type of test we will use. In addition, statement of the problem should explain precisely what we want to measure. The purpose of the test should be specified: achievement, proficiency, diagnostic, and placement. Each type is suitable for certain purposes. For example, proficiency tests “measure people’s ability in a language, regardless of any training they may have had in that language” (Hughes, 2003, p. 11). On the other hand, achievement tests “measure how much a student has learned from a particular course or syllabus” (Richards et al. 1992, p. 292). Moreover, placement tests help to classify test takers and assign them to the suitable programme of study. Finally, diagnostic tests “show what skills or knowledge a learner knows and doesn’t know” (Richards et al. 1992, p. 106).

Content: The description of the test content is used in writing different versions of the same test. Therefore, it is important to describe the content precisely, but in a way that will not limit the choices for test developers. The following are some of the elements in describing the content:

- Operations: In reading comprehension for example, the test taker should be able to perform different operations (constructs) such as skimming the text for the main idea, identifying referents of pronouns, and making inferences or conclusions.

- Types of texts: This might include texts taken from newspapers, magazines, textbooks, or the internet.
- Addresses: This may include a description of the test taker's educational level, native language, age, ethnic group, gender, and any required training or experience.
- Length of texts: This describes the length of passages in the test. The approximate number of words is specified.
- Topics: In reading comprehension tests, for example, subject areas will have to be as neutral and non-specialized as possible.
- Readability: This reflects the difficulty of the passage. The text can be set within any range of difficulty that complies with test specifications. Different formulas can be used to measure text difficulty such as Spache Readability Formula or Dale-Chall Readability Formula.
- Structural/vocabulary range: This shows whether there are any limitations in using specific range of structures or sets of vocabulary.

Format and timing: A description of the different parts or sections of the test is given, and the number of passages and questions is specified. Moreover, the time for the whole test or each part of it is mentioned.

Criterial levels of performance: A clear classification of different levels of performance is given. In addition, the necessary score to pass or retake the test is specified.

Scoring procedures: This explains whether the test will be scored objectively or subjectively. In addition, it gives full description of the scoring process especially if the test is subjectively scored. The rating process was discussed previously in this chapter.

## 3.5 Testing Methods

Many techniques and methods have been advocated for testing language skills. In this part, I will briefly talk about some of the techniques and methods that are considered more appropriate for testing reading comprehension. These are:

Multiple-choice: This is the most common method of testing language skills. A clear example of this method is IELTS and TOEFL. The multiple-choice method is popular because it is objectively scored, fast and economical which make it possible to include more items in the same period of time and hence may increase test reliability and validity. Finally, multiple-choice does not require other language skills such as writing. A test taker may be a fluent reader and at the same time a bad writer, which will prevent him/her from performing well in the reading comprehension test. When writing a multiple-choice item

> the ‘correct’ answer must be genuinely correct….Each wrong alternative should be attractive to at least some of the students…. The correct alternative should not look so different from the distracters that it stands out from the rest. It should not be noticeably longer or shorter, nor be written in a different style. (Alderson et al., 1995, p. 47-49).

Short-answer: This method is also called ‘open-ended’. The answer is usually very short and it is objectively scored. Hughes (2003) pointed out that short-answer items have some advantages over multiple choice questions and vice versa. Short-answer items are easer to construct, but they take more time to score. Moreover, cheating and guessing are less likely to happen with short-answer items. On the other hand, multiple-choice items do not require the use of other language skills.

Yes/no and true/false: Although it is popular, this method is not as good as the ‘multiple-choice’ or the ‘short-answer’. There are not enough distracters. Test takers have a 50% possibility of guessing the right answer.

Gap filling: Items in gap filling method “work best if the missing words are to be found in the text or are straightforward” (Hughes, 2003, p. 80). It is important to inform test takers whether each single gap should be filled with one word or more, and whether contractions such as ‘It’s and they’ll’ are considered one word or more (Alderson et al., 1995; Hughes, 2003).

Cloze: In this method, “words are deleted mechanically. Each nth word is deleted regardless of what the function of the word is” (Alderson et al., 1995, p. 55). The disadvantage of this method is that we cannot control the choice of deleted words. Moreover, it is difficult to agree on an answer key for the test since each gap might have different possible answers. (Alderson et al., 1995).

## 3.6 The Rating Process

At the next stage, the role of the test taker ends and the job of the test developer or administrator starts again. Papers are collected and are ready to be marked. Luckily, in reading comprehension tests, most of the testing methods are objectively scored which makes test scores more reliable. A scoring method is considered objective when the test is marked without depending on the evaluator/s personal judgment or opinion. Clear examples of the objective tests are the multiple-choice, yes/no and true/false questions. On the other hand, subjective tests are dependent on the evaluator/s personal judgment which may affect the reliability of test scores. Therefore, certain considerations are suggested to reduce this negative effect.

Problematic answers happen and consequently different reactions from examiners are found. Good training helps examiners to deal with these problematic answers and hence saves examiners' time and increase test reliability. Training makes subjective rating more objective. Another important consideration is the availability of a detailed rating manual. Examiners need this manual as a reference wherever they are not sure of something and cannot decide what to do.

Sometimes, in testing reading comprehension, test takers are asked to write their answers. In this case, how can we deal with their spelling and grammatical mistakes? Hughes (2003) pointed out that "errors of grammar, spelling or punctuation should not be penalised…. To test productive skills at the same time … simply makes the measurement of reading ability less valid" (p. 155).

## 3.7 Item Analysis

In item analysis, three main characteristics of the test are analyzed, which are: item difficulty, item discrimination, and distractor analysis. The rationale of item analysis is "to examine the contribution that each item is making to the test. Items that are identified as faulty or inefficient can be modified or rejected" (Hughes, 2003, p. 225). In addition, item analysis helps in maintaining test reliability and validity. The following is a brief description of each characteristic.

### Item difficulty.

We cannot determine which item in the test is difficult and which is easy simply by reading the questions. Item difficulty analysis is carried out to make sure that a test will be neither too difficult nor too easy. Item difficulty ranges from 0 to 1, with (0) being a very difficult item and (1) being a very easy item. For example, if only 24 students out of 150 answered an item correctly, then the item difficulty of this item is 24/150, which is (.16). Therefore, we can say that this item is very difficult.

#### Item discrimination.

Item discrimination distinguishes between test takers who have a lot of the ability being measured and those who have only a little of it. A good item discrimination value means that more high-level students will have answered the item correctly. The maximum value of discrimination is 1. The higher the value is, the better it discriminates (Hughes, 2003).

#### Distractor analysis.

The last step in analyzing an item is distractor analysis. It is used with multiple-choice items to show which distractors were chosen more often than others. If good students choose the correct answer and the rest randomly choose among distractors besides the correct answer, then it is a good item. If a distractor was not chosen at all, then it adds no real value to the test and it should be revised (Hughes, 2003).

In this chapter, I discussed some of the main issues and considerations in language testing. I tried to cover issues such as: reliability and how it could be increased, the sources of variance in examinees' performance, validity and how to validate a test, and Messick's six aspects of validity. Furthermore, a brief history of validity was introduced, and a general discussion of test specifications was presented.

# CHAPTER FOUR: RESEARCH DESIGN

## 4.1 Introduction

Although much attention has been given to factors that might affect testing reading comprehension, as mentioned earlier in chapter two, there has been very little interest in investigating the effect of using of L1 in testing L2 reading comprehension. In fact, no studies have been carried out to investigate the effect of using Arabic (L1) as a language for questions and answers in testing reading comprehension in English (L2) by using both multiple-choice and short-answer questions for upper intermediate and post beginner students. This absence of attention to this important factor has kept the picture unclear. Therefore, Alderson (2000) argues that there are still questions without answers, and there are doubts that should be clarified. More research is needed in this area. In the present research, the researcher will examine the effect of using Arabic (L1) in testing English (L2) reading comprehension. Other factors such as testing method and proficiency level will also be investigated.

Shohamy (1984) said that “presenting the questions in L1 may be considered more ethical, since the decision maker obtains information on the test taker’s ability to understand the L2 text, without a carry-over from the language of the questions” (p. 158). According to Hughes (2003), the use of L1 in testing L2 reading comprehension becomes more appropriate in monolingual situations. He said that “where candidates share a single native language, this can be used both for items and for responses” (p. 153). Moreover, he believes that the language of the questions should not add any difficulty to the test. However, he is concerned that the use of L1 might give clues to the test taker. Nevertheless, Alderson (2000) emphasized that more research should be carried out to examine these doubts.

## 4.2 Research questions

The following research questions were designed to guide the present study:

1- Does using Arabic (L1) in testing English (L2) reading comprehension affect the levels of performance of upper-intermediate and post-beginner students in multiple-choice and short answer questions?

2- When Arabic (L1) is used as the language of the questions and answers of an English (L2) reading comprehension test, how would gender, testing method, proficiency level, and reading comprehension sub-skills affect the level of performance of test-takers?

3- What do university and secondary school students and their English-language teachers think of using Arabic (L1) in the questions and answers of the reading comprehension test?

To investigate the research questions, four studies were carried out:

- The first study requires the development of the two original reading comprehension tests, and piloting them.
- The second study includes four case-II independent sample t-test studies which aim at determining whether the language of the test affects the performance of students in a reading comprehension test.
- The third study includes ninety case-II independent sample t-test studies which aim to trace any significant differences among the study variables that might have an effect on students performance. The variables are language of the test, proficiency level, testing method, gender and five sub-skills of reading comprehension.

- The forth study requires conducting sixteen semi-structured interviews to explore students and English-language teachers opinions about the use of Arabic (L1) as a language for questions and answers in reading comprehension tests.

## 4.3 Constructing the Reading Comprehension Test

As mentioned above, the first study in the present research requires the constructing of two original reading comprehension tests. The following are the test specifications of these two reading comprehension tests.

### 4.3.1 Test Specifications

Based on Hughes's (2003) classification, the test specifications of the reading comprehension test in the present research are as follows.

Statement of problem: From primary 6 to graduate level in Saudi Arabia, English is taught to all students as a foreign language, and is compulsory. Every year, thousands of secondary school students apply to the English departments in Saudi Arabia. One of the main admission requirements is to pass an English proficiency test. However, I met a large number of the students who did not pass the test and who complained and claimed that they did not pass because they faced difficulties in understanding test questions, which were written in English (L2). They argued that if questions were translated for them into Arabic, they would have answered more questions correctly.

There is an urgent need for reliable and valid language tests. Unfortunately, language testing is not well-researched in Saudi Arabia. There has been very little interest among researchers in the use of Arabic (L1) in testing English (L2) reading comprehension. This study aims to contribute towards this important area.

### 4.3.1.1 Content

Types of texts: The passages used in this study were non-specialized texts taken from encyclopaedias; however, these passages received minor modifications to meet test specifications.

Subjects: This reading test addressed two different groups of test takers. Table 4 summarizes these two groups.

Table 4

| | Group (A) | Group (B) |
|---|---|---|
| Educational level | 2nd year undergraduate | Final year secondary school |
| Native language | Arabic | Arabic |
| Level of proficiency | Upper intermediate | Post beginner |
| Age | 19-22 | 17-19 |
| No. of years studying English | 8 | 6 |

Group A: 72 second-year English department undergraduate Saudi students who have studied English as a foreign language for eight years. Their native language is Arabic and they are from the same age group (19 – 22 years). Their level of reading proficiency could be close to (B2) as defined by the Council of Europe Common European Framework in their ‘Overall Reading Comprehension’ scale. Their level will be referred to in the study as ‘upper-intermediate’.

Group B: 72 final year high-school Saudi students who have studied English as a foreign language for four hours a week for six years, in public schools. Their native language is Arabic and they are from the same age group (17 – 19 years). Their level of reading proficiency could be close to (A2) as defined by the Council of

Europe Common European Framework in their ‘Overall Reading Comprehension’ scale. Their level will be referred to in the study as ‘post-beginner’.

#### Sub-skills

The test included items that attempted to assess five reading comprehension sub-skills which were discussed in detail in chapter two. They are:

1. Scanning a text to locate specific information.
2. Skimming a text for the main idea.
3. Making inferences or drawing conclusions.
4. Identifying referents of pronouns.
5. Guessing the meaning of unknown words from context.

### 4.3.1.2 Testing methods

#### Multiple-choice.

This is the most common method of testing language skills, and it is used in parts of some international language tests such as IELTS and TOEFL. The multiple-choice method is popular because it is objectively scored, fast and economical which make it possible to include more items in the same period of time and hence may increase test reliability and validity. Moreover, the multiple-choice method does not require other language skills such as writing. A test taker may be a fluent reader and at the same time a bad writer, which might prevent him/her from performing well in the reading comprehension test.

#### Short-answer.

The answer is usually very short and it is objectively scored. Hughes (2003) pointed out that short-answer items have some advantages over multiple choice questions and vice versa. Short-answer items are easer to construct, but they take more time to mark. Moreover, cheating and guessing are less likely to happen with short-answer items.

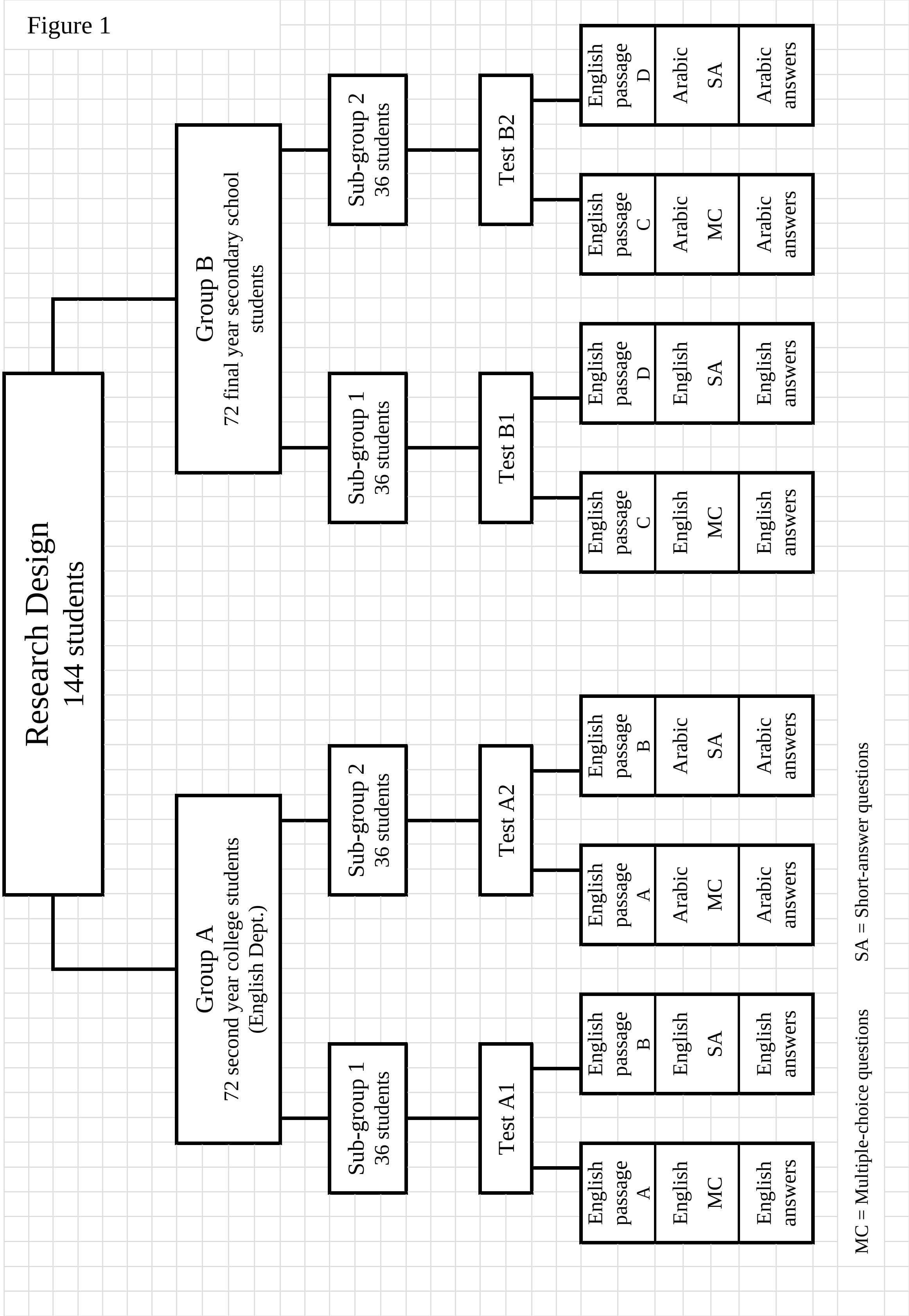
### 4.3.1.3 Format and timing

Figure 1 shows the research design of the main study in this research. This study consists of two similar parts. The first part will be administered to undergraduate English department students (Group A), while the second part will be administered to final year secondary school students (Group B). Both parts of the study are similar in the structure and the number of participants. The differences between them are the length and difficulty of the passages and the questions.

The following description of the test format is applicable to both parts of the study. In each part of the study, two authentic passages were used. The first passage was followed by ten multiple-choice questions and the second passage by ten short-answer questions. In each test, two testing methods were used. Furthermore, all the five reading sub-skills were assessed with equal number of questions in each test.

Students in each group (A) and (B) were divided randomly into two equal sub-groups. Both sub-groups took the same test. The only difference was the language of the questions and answers. The first sub-group had the questions in English and wrote their answers in English; the second sub-group had the questions in Arabic and were asked to write their answers in Arabic. Students were given thirty minutes to complete the test.

Figure 1



MC = Multiple-choice questions SA = Short-answer questions

Table 5 is a number key that shows which questions tested which sub-skills in the main study.

Table 5: The distribution of reading comprehension sub-skills in the test

| Reading Comprehension Sub-skills | | Test A1 (University) | | Test B1 (Secondary School) | |
|---|---|---|---|---|---|
| | | SA | MC | SA | MC |
| 1- | Scanning a text to locate specific information. | 1<br>6 | 15<br>18 | 4<br>6 | 12<br>16 |
| 2- | Skimming a text for the main idea. | 4<br>7 | 11<br>16 | 5<br>7 | 11<br>17 |
| 3- | Making inferences or drawing conclusions. | 5<br>8 | 12<br>17 | 1<br>8 | 13<br>18 |
| 4- | Identifying referents of pronouns. | 2<br>9 | 13<br>19 | 2<br>9 | 14<br>19 |
| 5- | Guessing the meaning of unknown words from context. | 3<br>10 | 14<br>20 | 3<br>10 | 15<br>20 |

SA = short-answer questions MC = multiple-choice questions

### 4.3.1.4 Scoring procedure

Since the tests are either in a multiple-choice or a short-answer format, they are objectively scored. Each correct answer will receive one point. A perfect score is 20 out of 20. All grammatical, spelling, and punctuation mistakes are ignored. Hughes (2003) argues that “errors of grammar, spelling or punctuation should not be penalised…. To test productive skills at the same time … simply makes the measurement of reading ability less valid” (p. 155).

### 4.3.1.5 Piloting

The drafts of the tests were vetted by native and non-native speakers of English, and the typographical errors were corrected. Moreover, after revision, the final draft of the tests were administered to a pilot group of undergraduate and secondary school Saudi students in order to identify any ambiguities in the test, to resolve unexpected problems, and to make sure that the specified time (40 minutes) was adequate.

## 4.4 Item Analysis

In the main study of the present research, a comprehensive item analysis was carried out. In this analysis, three main characteristics of the test were analysed, which were: item difficulty, item discrimination, and distractor analysis. “The purpose of item analysis is to examine the contribution that each item is making to the test. Items that are identified as faulty or inefficient can be modified or rejected” (Hughes, 2003, p. 225). The results of this comprehensive analysis helped in identifying the strengths and weaknesses of the tests. This will be discussed thoroughly in chapter five. According to Hughes (2003), the main aspects of item analysis are as follows:

### 4.4.1 Item Difficulty

It is not possible to determine which items in the test are difficult and which are easy simply by reading the questions. Item difficulty analysis is carried out to make sure that a test will be neither too difficult nor too easy. It is calculated by dividing the number of test takers answering an item correctly by the total number of test takers taking the test. Item difficulty ranges from 0 to 1, with (0) being a very difficult item and (1) being a very easy item.

### 4.4.2 Item Discrimination

Item discrimination is a measure of how much an item distinguishes between test takers who have much of the skill being measured and those who have only a little of it. The maximum value of discrimination is 1. The higher the value is, the better the item discriminates.

### 4.4.3 Distractor Analysis

The last step in analyzing an item is distractor analysis. It is used with multiple-choice items to show which distractors were chosen more often than others. If good students choose the correct answer and the rest randomly choose among distractors besides the correct answer, then it is a good item. If a distractor was not chosen at all, then it adds no real value to the test and it should be revised.

### 4.4.4 The standard error of measurement

All scores are only estimates, and they are not true scores. “Our tests are not perfectly reliable, and errors of measurement-error scores- cause observed scores to vary from true scores…. The more reliable the test is, the closer the obtained scores will cluster around the true score mean, resulting in a smaller standard deviation of errors” (Bachman, 1990, pp. 198-199). The standard error of measurement (SEM) shows the range of an individual’s true score within which he or she will score if a test were repeated again and again (Brown, 1988).

## 4.5 Statistical Analysis

The second study of the present research includes four case-II independent sample t-test studies which aim at investigating the effect of using Arabic (L1) as a language for questions and answers in testing reading comprehension in English (L2) by using both multiple-choice and short-answer questions for upper intermediate and post beginner students.

A case II independent sample t-test, with a one-tailed hypothesis will be used. The t-test was chosen because the comparison is between the means of the two independent groups. According to Hatch & Lazaraton (1991), case II t-test studies are more popular in applied linguistics research than are case I. “A case 1 study compares a sample mean with an established population mean” (p. 253), which is not the case in the present research where the comparison is between the means of the two groups. The assumptions for the t-test are met, and they are:

1. The independent variable (language of the test) is nominal and has only two levels: English and Arabic.
2. The dependent variable (test score) is interval.
3. The dependent variable is normally distributed.
4. Variances should be equal in each group. This is not important since the groups’ sizes are equal (Bachman, 2004).

The four main t-test studies of the present research are as follows:

1. First t-test study: test A1 (upper intermediate - English MC) and test A2 (upper intermediate - Arabic MC)
2. Second t-test study: test A1 (upper intermediate - English SA) and test A2 (upper intermediate - Arabic SA)
3. Third t-test study: test B1 (post beginner - English MC) and test B2 (post beginner - Arabic MC)
4. Forth t-test study: test B1 (post beginner - English SA) and test B2 (post beginner - Arabic SA)

The t-test observed value is compared with the critical value. Meaningfulness is determined by calculating eta$^2$ to show what percent of the dependent variable (test score) is accounted for by the independent variable (language of the test).

### 4.5.1 Research Hypotheses

H1: There is a positive effect of using L1 in testing L2 reading comprehension on the levels of performance of upper intermediate students in a multiple-choice reading comprehension test.

H2: There is a positive effect of using L1 in testing L2 reading comprehension on the levels of performance of upper intermediate students in a short-answer reading comprehension test.

H3: There is a positive effect of using L1 in testing L2 reading comprehension on the levels of performance of post beginner students in a multiple-choice reading comprehension test.

H4: There is a positive effect of using L1 in testing L2 reading comprehension on the levels of performance of post beginner students in a short-answer reading comprehension test.

The following are sub-hypotheses that consider all the possible combinations of the five sub-skills of reading in relation to gender and testing method. The sub-hypotheses were:

H5: There is a positive effect of using L1 in testing the L2 reading sub-skill of scanning on the levels of performance of M, F, M & F university students in MC, SA, MC & SA reading comprehension tests.

H6: There is a positive effect of using L1 in testing the L2 reading sub-skill of skimming on the levels of performance of M, F, M & F university students in MC, SA, MC & SA reading comprehension tests.

H7: There is a positive effect of using L1 in testing the L2 reading sub-skill of inferring on the levels of performance of M, F, M & F university students in MC, SA, MC & SA reading comprehension tests.

H8: There is a positive effect of using L1 in testing the L2 reading sub-skill of identifying references on the levels of performance of M, F, M & F university students in MC, SA, MC & SA reading comprehension tests.

H9: There is a positive effect of using L1 in testing the L2 reading sub-skill of guessing new words on the levels of performance of M, F, M & F university students in MC, SA, MC & SA reading comprehension tests.

H10: There is a positive effect of using L1 in testing the L2 reading sub-skill of scanning on the levels of performance of M, F, M & F secondary school students in MC, SA, MC & SA reading comprehension tests.

H11: There is a positive effect of using L1 in testing the L2 reading sub-skill of skimming on the levels of performance of M, F, M & F secondary school students in MC, SA, MC & SA reading comprehension tests.

H12: There is a positive effect of using L1 in testing the L2 reading sub-skill of inferring on the levels of performance of M, F, M & F secondary school students in MC, SA, MC & SA reading comprehension tests.

H13: There is a positive effect of using L1 in testing the L2 reading sub-skill of identifying references on the levels of performance of M, F, M & F secondary school students in MC, SA, MC & SA reading comprehension tests.

H14: There is a positive effect of using L1 in testing the L2 reading sub-skill of guessing new words on the levels of performance of M, F, M & F secondary school students in MC, SA, MC & SA reading comprehension tests.

## 4.6 Interviews

The interviews are part of the present study which aims at investigating the effect of using Arabic (L1) in testing English (L2) reading comprehension. They were used to help the researcher understand some of the students answers in the test, and to listen directly from them instead of relying only on his own interpretations. Furthermore, the interviews were expected to be a rich resource of students' feedback about their experience in this test, previous tests, or the difficulties they face in reading comprehension in general.

Sixteen interviews were carried out in this study. The participants were two teachers and six students from the College of Languages and Translation , Al-Imam University, in Riyadh, and two teachers and six students are from Al-Rowad Secondary School in Riyadh. I conducted all interviews myself, and tried to adhere to the following techniques and principles recommended in the literature:

1. Effective listening: To be a good interviewer I tried to be a good listener. Being so, enables the interviewer to use prompt and probes in the right time and consequently getting richer responses (Jones, 1991).
2. Facial expressions: “The face is the main communicator but our expression is often more impassive or ambiguous than we realize” (Gillham 2000, p. 30).
3. Eye contact: Eyes can convey what a tongue cannot. However, I was careful in using this technique and followed the advice of Gillham (2000). “Too much eye contact makes people feel embarrassed or ‘dominated’” (p. 31).
4. Gestures and head nods: The fewer the gestures and head nods, the more powerful and effective they are. A balanced use of them will encourage interviewees to speak and express their views with confidence (Dörnyei, 2007, Gillham 2000).

5. Proximity: Usually, interviewees do not fell comfortable if the interviewer is too close to them (Gillham, 2000). In this study, I used a small table between me and the interviewees.

6. Negative reinforcement: Sometimes, the interview loses its focus. Therefore, the interviewer should interrupt the interviewee in a polite way by a word or a comment in order to put the interview back to the right track (Dörnyei, 2007). Nevertheless, this should be done with caution because the interviewee may have an important point but he or she wants to give a brief introduction first.

7. Transition announcement: I tried to inform the interviewee "before a new issue is addressed because this advanced notice will help to create the appropriate mindset schema" (Dörnyei 2007, p. 143).

### 4.6.1 Interview Questions

#### 4.6.1.1 Opening questions

At the beginning of the interview, I started with questions about "noncontroversial present behaviours, activities, and experiences.... they require minimal recall and interpretation" (Patton 1990, p. 294). Many writers stress the importance of the first few questions because "they set the tone and create initial rapport" (Dörnyei 2007, p. 137). They are like a foundation for the interview. Therefore, all complex questions which require recall and interpretation were deferred until trust was established between the researcher and the interviewee (Jones, 1991).

#### 4.6.1.2 Content questions

Patton (1990) classifies interview questions into six types: 1) behaviour / experience questions 2) opinion / values questions 3) feeling questions 4) knowledge questions 5) sensory questions 6) background / demographic questions. Moreover, any of the above questions can be asked in the present, past, or future tense, which makes it possible to investigate a wide variety of issues.

"There are no fixed rules of sequence in organizing an interview" (Patton 1990, p. 294). Question sequences might differ from one interview to another depending on the circumstances, with the exception of the structured interview. The use of probes and prompts is essential in this part of the interview. A probe is a question generated from the interviewee's response and used to direct or explain the interviewee response (Dörnyei, 2007, Gillham 2000); while a prompt is used to remind the interviewees of issues they have not talked about (Gillham, 2000).

#### 4.6.1.3 Closing questions

When the interview came to its end, the researcher told the interviewee that he had no further questions. Moreover, he asked his interviewee if he would like to add anything. After that, the researcher expressed his thanks to the interviewee (Dörnyei, 2007).

### 4.6.2 Recording the Interview

Interview recording is obviously an important research tool. "Taking notes is simply not enough as we are unlikely to be able to catch all the details of the nuances of personal meaning; furthermore, note-taking also disrupts the interviewing process" (Dörnyei 2007, p. 139). However, Jones (1991) argues that "field notes should in any case be made to supplement taped interviews" (p. 207).Therefore, a balanced combination of both ways might help in compensating the loss of some non-verbal features during the interview.

Before starting the recording, I gave a brief introduction about my research topic and the purpose of the interview, which helped in building trust and rapport between us. Moreover, it was intended to encourage the interviewee to be a positive participant by giving deeper and richer responses (Dörnyei, 2007).

To ensure a good quality recording, I carefully chose the interview place to "avoid interruption background noise or intrusive curiosity (Gillham 2000, p. 7). Furthermore, I made sure that the recorder was working and that there were spare batteries. In addition, I made copies of the recorded materials and tagged them in a precise and clear way to avoid any confusion in the future (Dörnyei, 2007).

### 4.6.3 Students and Teachers Interviews

The interviews took place in clean, quiet, and well-lit rooms. All interviews were recorded. Students were selected randomly for the interviews which took place on the same day of the test or the very next day. Before the interview, I corrected students' papers to use them in the interview and ask every interviewee about his answers. The average time for each interview was about thirty minutes. Table 6 shows the student interview schedule.

Teacher interviews took place after the completion of student interviews The average time for each interview was about twenty-five minutes. Table 7 shows the teacher interview schedule. All interviewees in this study were cooperative and responsive but their answers were sometimes short.

Before the interview, I asked each interviewee about his language preference. Fifteen out of sixteen interviewees chose Arabic. Only one teacher chose English. Using interviewees' native language may have increased the reliability of the interviews. According to Silverman (1993), to achieve a higher reliability, “it is very important that each respondent understands the questions in the same way and that answers can be coded without the possibility of uncertainty” (p. 148). In this study both the researcher and the interviewees share the same native language, Arabic. This helped both sides to understand each other in a better way, and consequently increased the interview reliability.

Table 6: Student Interview Schedule

| | Questions | Prompts |
|---|---|---|
| Intro. | Greeting<br>Introducing myself and explaining the idea of my research<br>Taking permission to record the interview | |
| The Interview | What do you think of learning English as a foreign language? | ● purpose<br>● difficulties<br>● the easiest / most difficult skill |
| | What are the main difficulties in answering language tests? | ● written answers<br>● language of the questions<br>● time |
| | What is your preferred method of reading comprehension tests? Why? | ● MC / SA / …etc<br>● oral / written<br>● in class / take home |
| | What do you think of using Arabic in testing reading comprehension? | ● anxiety<br>● difficulty level<br>● any clues! |
| | Look at your corrected answer sheet and explain to me the wrong answers. | ● unclear instructions<br>● more than one correct answer<br>● time |
| | What is the strategy you used to answer each item in the test? Let us look again at your answer sheet. | |
| Closure | Is there any thing you want to add at the end of this interview?<br>Switch off the recorder<br>Appreciation comment | |

Table 7: Teacher Interview Schedule

| | Questions | Prompts |
|---|---|---|
| Intro. | Greeting<br>Introducing myself and explaining the idea of my research<br>Taking permission to record the interview | |
| The Interview | Could you please introduce yourself? | ● No. of years in teaching<br>● Grades you teach |
| | Could you please tell me about your experience in teaching reading skills? | ● Your preferred method<br>● The difficulties you face<br>● The language you use (L1, L2) |
| | How do you assess your students reading comprehension ability? | ● Testing methods (MC, SA …etc.)<br>● The preferred method to students<br>● The language you use (L1, L2) |
| | Why do some good students perform badly in reading comprehension tests? | ● Passage difficulty<br>● Questions difficulty (L1, L2)<br>● Instructions difficulty (L1, L2) |
| | What do you think of using Arabic (L1) in testing reading comprehension in English (L2)? | ● Validity<br>● Students' anxiety<br>● Testing method |
| | After the test, I interviewed some students and they gave me the following remarks: What do you think of each remark? | |
| Closure | Is there any thing you want to add at the end of this interview?<br>Switch off the recorder<br>Appreciation comment | |

I transcribed the substantive content of the interviews that were related to the research questions because I was interested in interviewees’ opinions, not their language. This accords with Kvale (2007, p. 94) view that “the amount and form of transcribing depends on such factors as the nature of the material and the purpose of the investigation”. Furthermore, Dörnyei (2007) pointed out that in some mixed method research “where the qualitative component is of secondary importance and is mainly intended to provide additional illustration or clarification….a possible compromise is to prepare a *partial transcription* of the sections that seem important” (p. 248-249). I translated the transcribed content and revised it twice. There were no difficulties in the translation process.

# CHAPTER FIVE: DATA COLLECTION

## 5.1 Introduction

The present study aims at investigating the effect of using Arabic (L1) in testing English (L2) reading comprehension. Other factors such as testing method and proficiency level will also be investigated. Two reading comprehension tests were developed and piloted in this study which was administered in Riyadh, Saudi Arabia, where all students share one native language, namely Arabic.

Eighty students participated in the pilot study, and one hundred and forty four students participated in the main study. They come from two groups, which are:

Group A: second-year English department undergraduate Saudi students who have studied English as a foreign language for eight years. Their native language is Arabic and they are from the same age group (19 – 22 years). Their level of proficiency could be close to upper intermediate.

Group B: final year high-school Saudi students who have studied English as a foreign language for four hours a week for six years, in public schools. Their native language is Arabic and they are from the same age group (17 – 19 years). Their level of proficiency could be close to post beginner.

The researcher travelled to Riyadh, Saudi Arabia, on 2-11-2006 to collect data for the study. The trip lasted for almost three months. The only difficulty faced the researcher in collecting data was getting the official written permission to conduct the study in public schools from the Saudi Ministry of Education, which took almost a month. The study took place at Al-Rowad Secondary School and at the College of Languages and Translation, Al-Imam University.

## 5.2 Test Development

### 5.2.1 Passage Selection

The passages used in this study were non-specialized texts taken from encyclopaedias; however, these passages received minor modifications to meet test specifications. The two passages for the second-year English Language Department undergraduate students were 350 words long, and were taken from *Microsoft Encarta Encyclopedia*. Their subjects were *Newspapers* and *Road Safety*. The two passages for the final year high-school students were 200 words long, and were taken from *1000 Questions and Answers Factfile* (Kerrod, R., Madgwick, W., Read, S., Collins, F., & Brooks, P., 2006). Their subjects were *Water* and *Deserts*.

### 5.2.2 Writing the Questions

The tests questions were written based on the test specifications. The reading sub-skills of the test were:

1. Scanning a text to locate specific information.
2. Skimming a text for the main idea.
3. Making inferences or drawing conclusions.
4. Identifying referents of pronouns.
5. Guessing the meaning of unknown words from context.

In the pilot study, two passages were used in each test. The first passage was followed by fifteen multiple-choice questions and the second passage by fifteen short-answer questions. Moreover, in every passage, the five reading comprehension sub-skills were tested three times. Afterwards, based on the results of the item analysis of the pilot study, the researcher chose ten questions for each passage to be used in the main study. More details about item selection are found at the end of the item analysis in this chapter. In the main study, each passage was followed by ten questions, where each construct was assessed by two questions. In both the pilot and main study, all the five reading comprehension sub-skills were assessed with the same number of questions.

### 5.2.3 Answer key

The researcher prepared an answer key for all the tests. Since the test is in either a multiple-choice or a short-answer format, it was objectively scored. Each correct answer received one point. A perfect score is 30 out of 30 in the pilot study, and 20 out of 20 in the main study. All grammatical, spelling, and punctuation mistakes were ignored. After correcting samples of the pilot-test papers, I added some possible answers to the answer key of the shorts answer questions. The answer key is included within the tests themselves in the appendices.

### 5.2.4 Translation

The researcher, whose native language is Arabic, translated all the questions and instructions of the tests into Arabic. Moreover, the translation was vetted by a University of Edinburgh PhD student whose major is translation, and whose native language is Arabic. The researcher faced some difficulties in translating the questions of the tests. However, he tried to minimize the effect of translating on the clarity, accuracy, and difficulty of the questions. Moreover, both the English and Arabic versions of the tests were identical in their arrangement and layout.

## 5.3 Data Collection

### 5.3.1 The First Phase

Before collecting data, the researcher coordinated with the following official Saudi authorities in order to get the permission for the study:

1- Saudi Arabian Cultural Bureau, London

2- College of Languages and Translation, Al-Imam University, Riyadh

3- Deanery of Graduate Studies and Scientific Research, Al-Imam University

4- Girls Education Centre, Al-Imam University, Riyadh

5- Ministry of Education, Riyadh

6- Al-Rowad Secondary School, Riyadh

The researcher explained his research plans and provided a copy of the tests. All of the bodies were cooperative and supportive, but bureaucracy at the Ministry of Education consumed almost a month of researcher's time before getting the official permission to conduct the study in public schools.

### 5.3.2 The Panel

While waiting for the official permission from the Saudi Ministry of Education, the researcher arranged a panel of five experienced English language teachers in order to vet the tests in the present study. One of the teachers is a teaching assistant at the Department of English and Literature, Al-Imam University. The second one is a senior supervisor of English language teachers in Saudi Arabia. The other three are English language teachers in Saudi public schools.

At the beginning of the meeting, the researcher explained the research idea and the role of the panel. After that, the tests for the final year high-school students were distributed among them. After fifteen minutes of silent reading, the discussion started. All questions were vetted by panel members who checked and confirmed that the five reading comprehension sub-skills were tested with equal number of items in each test. Subsequently, the researcher distributed the tests for the second-year English Language Department undergraduate students among the panel. After twenty minutes of silent reading, the discussion began. All questions were vetted by panel members. They checked and confirmed that the five reading comprehension sub-skills were tested with equal number of items in each test.

After that, there was an open discussion about the tests and the study in general. The panel suggested the following:

1- The researcher should start with the secondary school students for two reasons:

   a. Students have just finished the mid-term exams in secondary schools, and shortly they will be very busy with the final year examinations.

   b. Unlike the second-year English Language Department undergraduate students whose number is limited, there are large numbers of final year high-school students. Therefore, it would have been possible to find an alternative group of final year high-school students in case something went wrong.

2- Four members of the panel recommended conducting the study at Al-Rowad Secondary School because it is one of the biggest private schools in Riyadh. Moreover, its facilities might help the researcher to conduct his study in a better way.

## 5.4 The Pilot Study

The final drafts of the tests were administered to two pilot groups of students in order to identify any ambiguities in the test, to resolve unexpected problems, and to make sure that the specified time (40 minutes) was adequate.

Eighty Saudi students participated in the pilot study. The participants were told in the cover letter of the test that their compliance with this study was voluntary. The students were also told that if they chose to either not participate or to withdraw from the study at any time, there would be no penalty; they were also assured that their decision would not affect their grades. The cover letter was written in students' native language, Arabic, to ensure that they completely understood its content.

### 5.4.1 Format and Timing

Both the pilot and the main study are similar in their format. They consist of two similar parts. Tests (A1) & (A2) were administered to undergraduate English language department students (Group A), while test (B1) & (B2) were administered to final year secondary school students (Group B). Both parts of the study are similar in the structure and the number of participants. The differences between them are the length and difficulty of the passages and the questions.

In each part of the pilot and the main study, two passages have been used. The first passage is followed by ten multiple-choice questions and the second passage by ten short-answer questions. Students in each group (A) and (B) have been divided randomly into two equal sub-groups. Both sub-groups took the same test. The only difference was the language of the questions and answers. The first sub-group had the questions in English and wrote their answers in English; the second sub-group had the questions in Arabic and were asked to write their answers in Arabic. Students were given 40 minutes to complete the test.

### 5.4.2 Al-Rowad Secondary School

The researcher started the pilot study with the final year secondary school students as the panel suggested. Forty students from Al-Rowad Secondary School participated in the study. The test was given in large, clean, quiet, and well-lit classrooms.  The test lasted for forty minutes, but most of the students finished the test in thirty minutes. The researcher supervised the test by himself.

### 5.4.3 Department of English and Literature

Forty second-year English department undergraduate students participated in the study. The test was given in large, clean, quiet, and well-lighted classrooms inside the College of Languages and Translation. The test lasted for fifty minutes, but most of the students finished the test in forty minutes. The researcher supervised the test by himself.

### 5.4.4 Item Analysis

A comprehensive item analysis of the pilot study was carried out for all the four tests: English MC, English SA, Arabic MC, and Arabic SA. Based on the item analysis results, only ten questions out of the fifteen questions were chosen for each passage. More details about item selection are found at the end of the item analysis in this chapter. In this analysis, three main characteristics of the test were analysed, which are item difficulty, item discrimination, and distractor analysis.

### 5.4.4.1 Item difficulty

Item difficulty analysis was carried out to make sure that the tests would be neither too difficult nor too easy. Table 8 illustrates item difficulty for the secondary school Arabic short answer test. The table shows that all items of the test are of acceptable difficulty except items number (11) and (12) which are considered relatively easy, and item number (10) which is considered relatively difficult. Table 9 illustrates item difficulty for the secondary school English short answer test. The table shows that all items of the test are of acceptable difficulty except items number (2), (4) and (9) which are considered relatively easy, and items number (10) and (15) which are considered relatively difficult.

Table 8: Arabic SA - 22 Secondary school students

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Right | 17 | 16 | 17 | 17 | 12 | 11 | 13 | 16 | 17 | 1 | 18 | 18 | 7 | 13 | 6 |
| P | .77 | .73 | .77 | .77 | .55 | .50 | .59 | .73 | .77 | .05 | .82 | .82 | .32 | .59 | .27 |

Table 9: English SA - 18 Secondary school students

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Right | 12 | 16 | 11 | 15 | 5 | 12 | 11 | 7 | 16 | 1 | 8 | 12 | 4 | 13 | 2 |
| P | .67 | .89 | .61 | .83 | .28 | .67 | .61 | .39 | .89 | .06 | .44 | .67 | .22 | .72 | .11 |

Table 10 illustrates item difficulty for the secondary school Arabic multiple-choice test. The table shows that all items of the test are of acceptable difficulty except items number (19), (22) and (27) which are considered relatively easy. Table 11 illustrates item difficulty for the secondary school English multiple-choice test. The table shows that all items of the test are of acceptable difficulty except items number (19), (22), (27) and (29) which are considered relatively easy, and item number (18) which is considered relatively difficult.

Table 10: Arabic MC - 22 Secondary school students

| Item | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Right | 8 | 17 | 6 | 20 | 7 | 12 | 21 | 11 | 8 | 14 | 14 | 21 | 12 | 11 | 14 |
| P | .36 | .77 | .27 | .91 | .32 | .55 | .95 | .50 | .36 | .64 | .64 | .95 | .55 | .50 | .64 |

Table 11: English MC - 18 Secondary school students

| Item | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Right | 9 | 11 | 2 | 15 | 4 | 5 | 17 | 8 | 10 | 14 | 13 | 15 | 10 | 15 | 11 |
| P | .50 | .61 | .11 | .83 | .22 | .28 | .94 | .44 | .56 | .78 | .72 | .83 | .56 | .83 | .61 |

Table 12 illustrates item difficulty for the university Arabic short answer test. The table shows that all items of the test are of acceptable difficulty except item number (6) which is considered relatively difficult. Table 13 illustrates item difficulty for the university English short answer test. The table shows that all items of the test are of acceptable difficulty except items number (2), (3), (5), (6), (10), (13) and (15) which are considered relatively difficult.

Table 12: Arabic SA - 20 university students

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Right | 14 | 10 | 4 | 9 | 7 | 2 | 12 | 9 | 14 | 4 | 16 | 13 | 5 | 9 | 9 |
| P | .70 | .50 | .20 | .45 | .35 | .10 | .60 | .45 | .70 | .20 | .80 | .65 | .25 | .45 | .45 |

Table 13: English SA - 20 university students

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Right | 13 | 3 | 2 | 9 | 2 | 2 | 11 | 5 | 16 | 2 | 15 | 9 | 2 | 7 | 3 |
| P | .65 | .15 | .10 | .45 | .10 | .10 | .55 | .25 | .80 | .10 | .75 | .45 | .10 | .35 | .15 |

Table 14 illustrates item difficulty for the university Arabic multiple-choice test. The table shows that all items of the test are of acceptable difficulty except item number (22) which is considered relatively difficult. Table 15 illustrates item difficulty for the university English multiple-choice test. The table shows that all items of the test are of acceptable difficulty.

Table 14: Arabic MC - 20 university students

| Item | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Right | 13 | 13 | 7 | 10 | 7 | 12 | 2 | 6 | 14 | 5 | 8 | 8 | 5 | 7 | 6 |
| P | .65 | .65 | .35 | .50 | .35 | .60 | .10 | .30 | .70 | .25 | .40 | .40 | .25 | .35 | .30 |

Table 15: English MC - 20 university students

| Item | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Right | 13 | 14 | 9 | 8 | 8 | 10 | 5 | 10 | 10 | 8 | 12 | 5 | 5 | 13 | 4 |
| P | .65 | .70 | .45 | .40 | .40 | .50 | .25 | .50 | .50 | .40 | .60 | .25 | .25 | .65 | .20 |

### 5.4.4.2 Item discrimination

Item discrimination is carried out to measure how much each item distinguishes between test takers who have a high level of the skill being measured and those who have only a little of it. Values greater than .40 are considered very good, between .30 and .39 are acceptable; between .20 and .29 are marginal, and below .19 are poor and should be revised (McDonald, 2002). To calculate item discrimination, I arranged scores from the highest total score to the lowest total score. Then, I divided the scores into three equal parts. For each item in the upper third, I divided the number of examinees who answered correctly by the number of examinees in that third. Also, for each item in the lower third, I divided the number of examinees who answered correctly by the number of examinees in that third. Then, I calculated the difference: $D = P_h – P_L$.

Table 16 illustrates the item discrimination value (D) for each item in the secondary school Arabic short answer test. The table shows that eight items performed well and have good discrimination values. Items number (2), (3), and (4) have a marginal discrimination value, but items number (1), (7), (10), and (11) have a poor discrimination value. Table 17 illustrates the item discrimination value (D) for each item in the secondary school English short answer test. The table shows that seven items performed well and have good discrimination values. Items (5) and (15) have acceptable discrimination values, but items number (1), (2), (3), (4), (9) and (10) have a poor discrimination value.

Table 16: Arabic SA - 22 Secondary school students

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High group | 6 | 6 | 6 | 6 | 7 | 6 | 3 | 7 | 7 | 0 | 5 | 7 | 6 | 6 | 4 |
| P | .86 | .86 | .86 | .86 | 1 | .86 | .43 | 1 | 1 | 0 | .71 | 1 | .86 | .86 | .57 |
| Low group | 5 | 4 | 4 | 4 | 1 | 1 | 5 | 3 | 3 | 1 | 6 | 3 | 0 | 0 | 0 |
| P | .71 | .57 | .57 | .57 | .14 | .14 | .71 | .43 | .43 | .14 | .86 | .43 | 0 | 0 | 0 |
| $D = P_H - P_L$ | .15 | .29 | .29 | .29 | .86 | .72 | -.28 | .57 | .57 | -.14 | -.15 | .57 | .86 | .86 | .57 |

Table 17: English SA - 18 Secondary school students

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High group | 5 | 6 | 6 | 5 | 3 | 6 | 6 | 5 | 6 | 1 | 4 | 6 | 4 | 6 | 2 |
| P | .83 | 1 | 1 | .83 | .50 | 1 | 1 | .83 | 1 | .17 | .67 | 1 | .67 | 1 | .33 |
| Low group | 4 | 5 | 0 | 4 | 1 | 3 | 2 | 0 | 6 | 0 | 1 | 2 | 0 | 3 | 0 |
| P | .67 | .83 | 0 | .67 | .17 | .50 | .33 | 0 | 1 | 0 | .17 | .33 | 0 | .50 | 0 |
| $D = P_H - P_L$ | .16 | .17 | 1 | .16 | .33 | .50 | .67 | .83 | 0 | .17 | .50 | .67 | .67 | .50 | .33 |

Table 18 illustrates the item discrimination value (D) for each item in the secondary school Arabic multiple-choice test. The table shows that twelve items performed well and have good discrimination values. Items number (19), (22), and (27) have a poor discrimination value. Table 19 illustrates the item discrimination value (D) for each item in the secondary school English multiple-choice test. The table shows that seven items performed well and have good discrimination values. Items (18), (19), (26), and (27) have acceptable discrimination values, but items number (20), (22), (24), and (29) have a poor discrimination value.

Table 18: Arabic MC - 22 Secondary school students

| Item | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High group | 4 | 7 | 5 | 7 | 4 | 7 | 7 | 6 | 6 | 6 | 7 | 7 | 7 | 7 | 6 |
| P | .57 | 1 | .71 | 1 | .57 | 1 | 1 | .86 | .86 | .86 | 1 | 1 | 1 | 1 | .86 |
| Low group | 0 | 3 | 1 | 6 | 1 | 2 | 6 | 3 | 0 | 3 | 2 | 6 | 2 | 2 | 3 |
| P | 0 | .43 | .14 | .86 | .14 | .29 | .86 | .43 | 0 | .43 | .29 | .86 | .29 | .29 | .43 |
| $D = P_H - P_L$ | .57 | .57 | .57 | .14 | .43 | .71 | .14 | .43 | .86 | .43 | .71 | .14 | .71 | .71 | .43 |

Table 19: English MC - 18 Secondary school students

| Item | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High group | 4 | 6 | 2 | 6 | 2 | 4 | 6 | 4 | 6 | 6 | 6 | 6 | 6 | 5 | 6 |
| P | .67 | 1 | .33 | 1 | .33 | .67 | 1 | .67 | 1 | 1 | 1 | 1 | 1 | .83 | 1 |
| Low group | 1 | 1 | 0 | 4 | 1 | 0 | 5 | 0 | 0 | 3 | 4 | 4 | 1 | 6 | 1 |
| P | .17 | .17 | 0 | .67 | .17 | 0 | .83 | 0 | 0 | .50 | .67 | .67 | .17 | 1 | .17 |
| $D = P_H - P_L$ | .50 | .83 | .33 | .33 | .16 | .67 | .17 | .67 | 1 | .50 | .33 | .33 | .83 | -.17 | .83 |

Table 20 illustrates the item discrimination value (D) for each item in the university Arabic short answer test. The table shows that five items performed well and have good discrimination values. Items number (2), (3), (4), (6), (11), (12), and (13) have a marginal discrimination value, but items number (7), (10), and (15) have a poor discrimination value. Table 21 illustrates the item discrimination value (D) for each item in the university English short answer test. The table shows that four items performed well and have good discrimination values. Items (2), (6), (10), (11), (13), and (14) have marginal discrimination values, but items number (1), (3), (5), (8), and (9) have a poor discrimination value.

Table 20: Arabic SA - 20 university students

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High group | 7 | 4 | 3 | 4 | 5 | 2 | 4 | 5 | 7 | 2 | 7 | 5 | 3 | 5 | 3 |
| P | 1 | .57 | .43 | .57 | .71 | .29 | .57 | .71 | 1 | .29 | 1 | .71 | .43 | .71 | .43 |
| Low group | 3 | 2 | 1 | 2 | 0 | 0 | 5 | 2 | 2 | 1 | 5 | 3 | 1 | 1 | 2 |
| P | .43 | .29 | .14 | .29 | 0 | 0 | .71 | .29 | .29 | .14 | .71 | .43 | .14 | .14 | .29 |
| D = $P_H$ - $P_L$ | .57 | .28 | .29 | .28 | .71 | .29 | -.14 | .42 | .71 | .15 | .29 | .28 | .29 | .57 | .14 |

Table 21: English SA - 20 university students

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High group | 4 | 2 | 1 | 4 | 1 | 2 | 7 | 3 | 7 | 2 | 6 | 6 | 2 | 3 | 3 |
| P | .57 | .29 | .14 | .57 | .14 | .29 | 1 | .43 | 1 | .29 | .86 | .86 | .29 | .43 | .43 |
| Low group | 3 | 0 | 0 | 1 | 0 | 0 | 3 | 2 | 6 | 0 | 4 | 1 | 0 | 1 | 0 |
| P | .43 | 0 | 0 | .14 | 0 | 0 | .43 | .29 | .86 | 0 | .57 | .14 | 0 | .14 | 0 |
| D = $P_H$ - $P_L$ | .14 | .29 | .14 | .43 | .14 | .29 | .57 | .14 | .14 | .29 | .29 | .72 | .29 | .29 | .43 |

Table 22 illustrates the item discrimination value (D) for each item in the university Arabic multiple-choice test. The table shows that nine items performed well and have good discrimination values. Item (17) has marginal discrimination value, but items number (16), (22), (24), (25), and (27) have a poor discrimination value. Table 23 illustrates the item discrimination value (D) for each item in the university English multiple-choice test. The table shows that nine items performed well and have good discrimination values. Items (17), (22), (24), and (27) have marginal discrimination values, but items number (20) and (25) have a poor discrimination value.

Table 22: Arabic MC - 20 university students

| Item | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High group | 5 | 5 | 5 | 5 | 4 | 6 | 0 | 4 | 5 | 2 | 5 | 2 | 4 | 6 | 4 |
| P | .71 | .71 | .71 | .71 | .57 | .86 | 0 | .57 | .71 | .29 | .71 | .29 | .57 | .86 | .57 |
| Low group | 4 | 3 | 0 | 2 | 0 | 1 | 2 | 1 | 4 | 3 | 1 | 2 | 1 | 1 | 1 |
| P | .57 | .43 | 0 | .29 | 0 | .14 | .29 | .14 | .57 | .43 | .14 | .29 | .14 | .14 | .14 |
| $D = P_H - P_L$ | .14 | .28 | .71 | .71 | .57 | .72 | -.29 | .43 | .14 | -.14 | .57 | 0 | .43 | .72 | .43 |

Table 23: English MC - 20 university students

| Item | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| High group | 7 | 6 | 4 | 4 | 1 | 6 | 3 | 5 | 4 | 3 | 7 | 3 | 3 | 6 | 3 |
| P | 1 | .86 | .57 | .57 | .14 | .86 | .43 | .71 | .57 | .43 | 1 | .43 | .43 | .86 | .43 |
| Low group | 1 | 4 | 1 | 1 | 3 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 0 | 3 | 0 |
| P | .14 | .57 | .14 | .14 | .43 | .29 | .14 | .29 | .29 | .29 | .29 | .14 | 0 | .43 | 0 |
| $D = P_H - P_L$ | .86 | .29 | .43 | .43 | -.29 | .57 | .29 | .42 | .28 | .14 | .71 | .29 | .43 | .43 | .43 |

### 5.4.4.3 Distractor analysis

A distractor analysis of all the multiple-choice items was carried out. Table 24 shows that 48 out of the 60 distractors in the Arabic multiple-choice secondary school test were chosen by examinees at least once. The other twelve distractors were not chosen at all. They are number (a) in item (16), number (a) in item (17), number (d) in item (18), number (b) and (c) in item (19), number (a) and (c) in item (22), number (c) and (d) in item (23), number (b) and (d) in item (27), and number (c) in item (29). Table 25 shows that 52 out of the 60 distractors in the English multiple-choice secondary school test were chosen by examinees at least once. The other eight distractors were not chosen at all. They are number (d) in item (18), number (b) and (c) in item (22), number (d) in item (23), number (c) and (d) in item (24), number (c) in item (26), and number (c) in item (27).

Table 24: Arabic multiple-choice secondary school test (22 students)

| | Key | A | B | C | D |
|---|---|---|---|---|---|
| Q16 | C | 0 | 3 | 8 | 11 |
| Q17 | D | 0 | 3 | 2 | 17 |
| Q18 | B | 2 | 6 | 14 | 0 |
| Q19 | D | 2 | 0 | 0 | 20 |
| Q20 | C | 7 | 4 | 7 | 4 |
| Q21 | B | 2 | 12 | 4 | 4 |
| Q22 | D | 0 | 1 | 0 | 21 |
| Q23 | A | 11 | 11 | 0 | 0 |
| Q24 | B | 10 | 8 | 1 | 3 |
| Q25 | A | 14 | 4 | 3 | 1 |
| Q26 | B | 2 | 14 | 1 | 5 |
| Q27 | A | 21 | 0 | 1 | 0 |
| Q28 | C | 2 | 7 | 12 | 1 |
| Q29 | D | 3 | 8 | 0 | 11 |
| Q30 | C | 1 | 2 | 14 | 5 |

Table 25: English multiple-choice secondary school test (18 students)

| | Key | A | B | C | D |
|---|---|---|---|---|---|
| Q16 | C | 1 | 3 | 9 | 5 |
| Q17 | D | 2 | 3 | 2 | 11 |
| Q18 | B | 1 | 2 | 15 | 0 |
| Q19 | D | 1 | 1 | 1 | 15 |
| Q20 | C | 6 | 6 | 4 | 2 |
| Q21 | B | 3 | 5 | 7 | 3 |
| Q22 | D | 1 | 0 | 0 | 17 |
| Q23 | A | 8 | 8 | 2 | 0 |
| Q24 | B | 8 | 10 | 0 | 0 |
| Q25 | A | 14 | 1 | 1 | 2 |
| Q26 | B | 2 | 13 | 0 | 3 |
| Q27 | A | 15 | 1 | 0 | 2 |
| Q28 | C | 3 | 3 | 10 | 2 |
| Q29 | D | 1 | 1 | 1 | 15 |
| Q30 | C | 4 | 2 | 10 | 2 |

Table 26 shows that 52 out of the 60 distractors in the Arabic multiple-choice university test were chosen by examinees at least once. The other eight distractors were not chosen at all. They are number (a) in item (16), number (d) in item (17), number (b) and (d) in item (22), number (d) in item (23), number (b) and (d) in item (24), and number (c) in item (29). Table 27 shows that 57 out of the 60 distractors in the English multiple-choice university test were chosen by examinees at least once. The other three distractors were not chosen at all. They are number (b) in item (17), number (b) in item (22), and number (d) in item (23).

Table 26: Arabic multiple-choice university test (20 students)

| | Key | A | B | C | D |
|---|---|---|---|---|---|
| Q16 | D | 0 | 4 | 3 | 13 |
| Q17 | A | 13 | 4 | 3 | 0 |
| Q18 | B | 6 | 7 | 2 | 5 |
| Q19 | D | 1 | 6 | 3 | 10 |
| Q20 | C | 3 | 1 | 7 | 9 |

| | | | | | |
|---|---|---|---|---|---|
| Q21 | B | 2 | 12 | 1 | 5 |
| Q22 | C | 18 | 0 | 2 | 0 |
| Q23 | A | 6 | 7 | 7 | 0 |
| Q24 | C | 6 | 0 | 14 | 0 |
| Q25 | B | 4 | 5 | 8 | 3 |
| Q26 | C | 6 | 1 | 8 | 5 |
| Q27 | B | 3 | 8 | 7 | 2 |
| Q28 | B | 11 | 5 | 1 | 3 |
| Q29 | A | 7 | 8 | 0 | 5 |
| Q30 | D | 1 | 12 | 1 | 6 |

Table 27: English multiple-choice university test (20 students)

| | Key | A | B | C | D |
|---|---|---|---|---|---|
| Q16 | D | 1 | 2 | 4 | 13 |
| Q17 | A | 14 | 0 | 4 | 2 |
| Q18 | B | 5 | 9 | 2 | 4 |
| Q19 | D | 1 | 8 | 3 | 8 |
| Q20 | C | 5 | 1 | 8 | 6 |
| Q21 | B | 2 | 10 | 2 | 6 |
| Q22 | C | 12 | 0 | 5 | 3 |
| Q23 | A | 10 | 6 | 4 | 0 |
| Q24 | C | 3 | 4 | 10 | 3 |
| Q25 | B | 2 | 8 | 7 | 3 |
| Q26 | C | 2 | 1 | 12 | 5 |
| Q27 | B | 4 | 5 | 9 | 2 |
| Q28 | B | 9 | 5 | 2 | 4 |
| Q29 | A | 13 | 4 | 2 | 1 |
| Q30 | D | 5 | 7 | 4 | 4 |

## 5.5 The Main Study

One hundred and forty four Saudi students participated in the main study which took place at four different places. These were:

1- College of Languages and Translation, Al-Imam University, Riyadh

2- Girls Education Centre, Al-Imam University, Riyadh

3- Al-Rowad Secondary School for Boys, Riyadh

4- Al-Rowad Secondary School for Girls, Riyadh

The participants were told in the cover letter of the test that their compliance with this study is voluntary. The students were told that if they chose to either not participate or to withdraw from the study at any time, there would be no penalty; they were also assured that their decision would not affect their grades. The cover letter was written in students' native language, Arabic, to ensure that they completely understood its content.

### 5.5.1 Item selection

Based on the results of this comprehensive item analysis of all the tests in the pilot study, the researcher chose ten questions out of fifteen to use in the main study. Five items in each test were eliminated. In the secondary school test, for example, item number 10 was eliminated because it had a poor discrimination value and was considered to be relatively difficult in both versions, i.e. Arabic and English. Item number 19 had a poor discrimination value, its Arabic version was considered to be relatively easy and two of its distractors were not chosen at all. Item number 22 was considered to be relatively easy, had a poor discrimination value and two of its distractors were not chosen at all. At the university level, item number 3 was considered to be relatively difficult and had a poor discrimination value in its English version. Item number 6 was considered to be relatively difficult in both versions. The Arabic version of item number 24 had a poor discrimination value and two of its distractors were not chosen at all. Item number 25 had a poor discrimination value in both versions.

Moreover, some of the selected ten items for the main study received minor modifications. For example, in the secondary school pilot test, the word ‘synonym’ was unknown to most of the students. Therefore, items number 5 and 15 were changed from: ‘Write a synonym for the word “……” in line (…)?’ to ‘What does the word “……” in line (…) mean?’ Sometimes, one or two of the Arabic or English distractors in the multiple-choice tests were replaced because they did not work well. For example, in items 27 and 29, two distractors were changed in each item. The distractor ‘rain’ was replaced with ‘temperatures’; ‘weather’ with ‘desert animals’; ‘rainfall’ with ‘rain’; and ‘plants’ with ‘places’. In the university test, two distractors were changed in item number 22. The distractor ‘car accidents’ was replaced with ‘careless drivers’ and ‘seat belt’ with ‘road design’. Any change made to an item is also repeated in its parallel item in the other language.

After that, the researcher prepared the final version of all the tests, and edited them in similar layouts. Versions of both the pilot and main study tests are provided in the appendices. Time was changed from 40 to 20 minutes for the secondary school students and from 50 to 30 minutes for the university students. This reduction in time for the main study was based on the average time needed by most students to finish the test in the pilot study. In addition, time was reduced due to the reduction in the number of questions from 30 in the pilot study to 20 in the main study.

### 5.5.2 Secondary School Students

Seventy-two final year secondary school students (36 men & 36 women) completed the test and agreed to be part of the study. All of them scored over 70 out of 100 in their final-year English-language exam. The researcher himself supervised the study at the boys' school. Students were divided randomly into two equal sub-groups. They answered the same test. The only difference is the language of the questions and answers, which was Arabic for the first sub-group and English for the second one. At the girls' school, an experienced English language teacher supervised the tests on behalf of the researcher. I explained the study to her in detail. In addition, I was available by phone to answer any questions or to solve any problems in case something went wrong, but everything went smoothly.

### 5.5.3 University Students

Seventy-two second-year English department students (36 men & 36 women) completed the test and agreed to be part of the study. All of them passed the entrance test of the Department of English and Literature and also passed all of the first-year English Department courses. The researcher himself supervised the study at the English department. Students were divided randomly into two equal sub-groups. They answered the same test. The only difference is the language of the questions and answers, which was Arabic for the first sub-group and English for the second one. At the Girls Education Centre, an English-language lecturer supervised the tests on behalf of the researcher.

### 5.5.4 Students Interviews

After the test, six final year secondary school students and six second year English language undergraduate students were selected randomly for the interviews which took place on the same day of the test or the very next day. All interviews were recorded. Before the interview, the researcher marked students papers to use them in the interview and ask every student about his answers. Interviews took place in clean, quiet, and well-lit rooms. The average time for each interview was about thirty minutes. Table 6 shows the pattern of students interviews (see page 85).

### 5.5.5 Teachers Interviews

After the completion of students' interviews, the researcher coordinated with two English language teachers at Al-Rowad Secondary School and two English language teachers at the Department of English and Literature, Al-Imam University. The interviews took place in clean, quiet, and well-lit rooms. All interviews were recorded. The average time for each interview was about twenty-five minutes. Table 7 shows the pattern of teachers interviews (see page 86).

# CHAPTER SIX: QUANTITATIVE RESULTS AND DISCUSSION

## 6.1 Introduction

This chapter presents the findings related to the quantitative research questions. As discussed in more detail in chapter four, the total number of participants in this study is one hundred forty four students, and they come from two different groups. Group A consists of seventy-two second-year English department students (36 men & 36 women), and group B consists of seventy-two final year secondary school students (36 men & 36 women). All participants share the same native language, namely Arabic.

The current study seeks to examine the effect of using Arabic (L1) as a language for questions and answers in testing reading comprehension in English (L2) by using both multiple-choice and short-answer questions for university students (upper intermediate) and secondary school students (post beginner). Other factors including gender and five reading sub-skills were considered. Each sub-skill was explored in relation to gender (M, F, M+F) and testing method (MC, SA, MC+SA). All possible combinations were examined in order to make sure that all the controlled variables were considered, and to trace any possible source of significant difference.

The study utilizes a quantitative method (case II independent sample t-tests) to gain information about students performance. Ninety-four case-II independent sample t-test studies were carried out. The hypotheses were:

H1: There is a positive effect of using L1 in testing L2 reading comprehension on the levels of performance of upper intermediate students in a multiple-choice reading comprehension test.

H2: There is a positive effect of using L1 in testing L2 reading comprehension on the levels of performance of upper intermediate students in a short-answer reading comprehension test.

H3: There is a positive effect of using L1 in testing L2 reading comprehension on the levels of performance of post beginner students in a multiple-choice reading comprehension test.

H4: There is a positive effect of using L1 in testing L2 reading comprehension on the levels of performance of post beginner students in a short-answer reading comprehension test.

The following are sub-hypotheses that consider all the possible combinations of the five sub-skills of reading in relation to gender and testing method. The sub-hypotheses were:

H5: There is a positive effect of using L1 in testing the L2 reading sub-skill of scanning on the levels of performance of M, F, M & F university students in MC, SA, MC & SA reading comprehension tests.

H6: There is a positive effect of using L1 in testing the L2 reading sub-skill of skimming on the levels of performance of M, F, M & F university students in MC, SA, MC & SA reading comprehension tests.

H7: There is a positive effect of using L1 in testing the L2 reading sub-skill of inferring on the levels of performance of M, F, M & F university students in MC, SA, MC & SA reading comprehension tests.

H8: There is a positive effect of using L1 in testing the L2 reading sub-skill of identifying references on the levels of performance of M, F, M & F university students in MC, SA, MC & SA reading comprehension tests.

H9: There is a positive effect of using L1 in testing the L2 reading sub-skill of guessing new words on the levels of performance of M, F, M & F university students in MC, SA, MC & SA reading comprehension tests.

H10: There is a positive effect of using L1 in testing the L2 reading sub-skill of scanning on the levels of performance of M, F, M & F secondary school students in MC, SA, MC & SA reading comprehension tests.

H11: There is a positive effect of using L1 in testing the L2 reading sub-skill of skimming on the levels of performance of M, F, M & F secondary school students in MC, SA, MC & SA reading comprehension tests.

H12: There is a positive effect of using L1 in testing the L2 reading sub-skill of inferring on the levels of performance of M, F, M & F secondary school students in MC, SA, MC & SA reading comprehension tests.

H13: There is a positive effect of using L1 in testing the L2 reading sub-skill of identifying references on the levels of performance of M, F, M & F secondary school students in MC, SA, MC & SA reading comprehension tests.

H14: There is a positive effect of using L1 in testing the L2 reading sub-skill of guessing new words on the levels of performance of M, F, M & F secondary school students in MC, SA, MC & SA reading comprehension tests.

## 6.2 University Students (Multiple-choice Test)

The first main study examined the effect of using L1 in testing L2 reading comprehension on the levels of performance of university students in a multiple-choice test. The observed value (1.218) of the t-test study in table 28 was less than its critical value (1.671). The first hypothesis was rejected at p < .05.

Table 28: University Students (Multiple-choice Test)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 4.5278 | 5.1389 | 1.218 |
| s | 2.235891 | 2.016401 | |

* P < .05 , df = 70

## 6.3 University Students (Short Answer Test)

The second main study examined the effect of using L1 in testing L2 reading comprehension on the levels of performance of university students in a short-answer test. The observed value (1.372) of the t-test study in table 29 was less than its critical value (1.671). The second hypothesis was rejected at p < .05.

Table 29: University Students (Short answer Test)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 5.75 | 6.4167 | 1.372 |
| s | 2.075366 | 2.047647 | |

* P < .05 , df = 70

## 6.4 Secondary School Students (Multiple-choice Test)

The third main study examined the effect of using L1 in testing L2 reading comprehension on the levels of performance of secondary school students in a multiple-choice test. The observed value (0.702) of the t-test study in table 30 was less than its critical value (1.671). The third hypothesis was rejected at p < .05.

Table 30: Secondary School Students (Multiple-choice Test)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 7.3889 | 7 | 0.702 |
| s | 2.405681 | 2.292846 | |

* P < .05 , df = 70

## 6.5 Secondary School Students (Short Answer Test)

The fourth main study examined the effect of using L1 in testing L2 reading comprehension on the levels of performance of secondary school students in a short-answer test. The observed value (0.115) of the t-test study in table 31 was less than its critical value (1.671). The fourth hypothesis was rejected at p < .05.

Table 31: Secondary School Students (Short answer Test)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 7.0833 | 7.1389 | 0.115 |
| s | 2.182724 | 1.914647 | |

* P < .05 , df = 70

## 6.6 University Students

### 6.6.1 Scanning a Text to Locate Specific Information

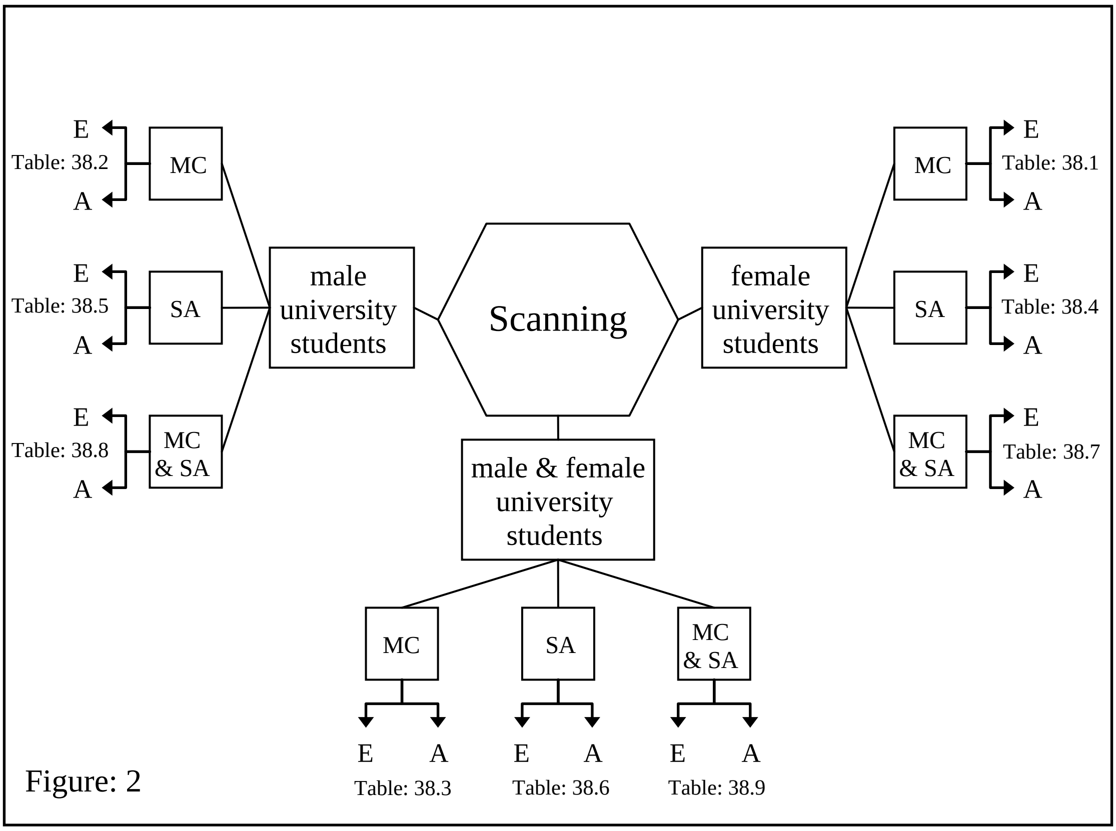


Figure: 2

The observed values of the t-test studies in table 32.1 (1.2761), table 32.2 (-.6216), table 32.4 (.3776), table 32.5 (.1925), table 32.7 (.8017), and table 32.8 (-.2048) are all less than their critical value (3.646); and the observed values of the t-test studies in table 32.3 (.3264), table 32.6 (.4102), and table 32.9 (.3873) are less than their critical value (3.460). Therefore, hypothesis 5 with all its variables as in figure 2 above is rejected at p < .0005.

Table 32.1: Multiple-choice: university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .778 | 1.333 | 1.2761 |
| s | 1.085 | 1.495 | |

* P < .0005 , df = 34

Table 32.2: Multiple-choice: university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.611 | 1.278 | - .6216 |
| s | 1.766 | 1.435 | |

* P < .0005 , df = 34

Table 32.3: Multiple-choice: all university students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.194 | 1.306 | .3264 |
| s | 1.444 | 1.444 | |

* P < .0005 , df = 70

Table 32.4: Short answer: university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.5 | 1.722 | .3776 |
| s | 1.663 | 1.863 | |

* P < .0005 , df = 34

Table 32.5: Short answer: university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.5 | 1.611 | .1925 |
| s | 1.732 | 1.732 | |

* P < .0005 , df = 34

Table 32.6: Short answer: all university students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.5 | 1.667 | .4102 |
| s | 1.673 | 1.773 | |

* P < .0005 , df = 70

Table 32.7: The whole test (MC+SA): university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 2.278 | 3.056 | .8017 |
| s | 2.556 | 3.227 | |

* P < .0005 , df = 34

Table 32.8: The whole test (MC+SA): university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 3.111 | 2.889 | - .2048 |
| s | 3.413 | 3.087 | |

* P < .0005 , df = 34

Table 32.9:
The whole test (MC+SA): all university students (female + male)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 2.694 | 2.972 | .3873 |
| s | 2.971 | 3.112 | |

* P < .0005 , df = 70

## 6.6.2 Skimming a Text for the Main Idea

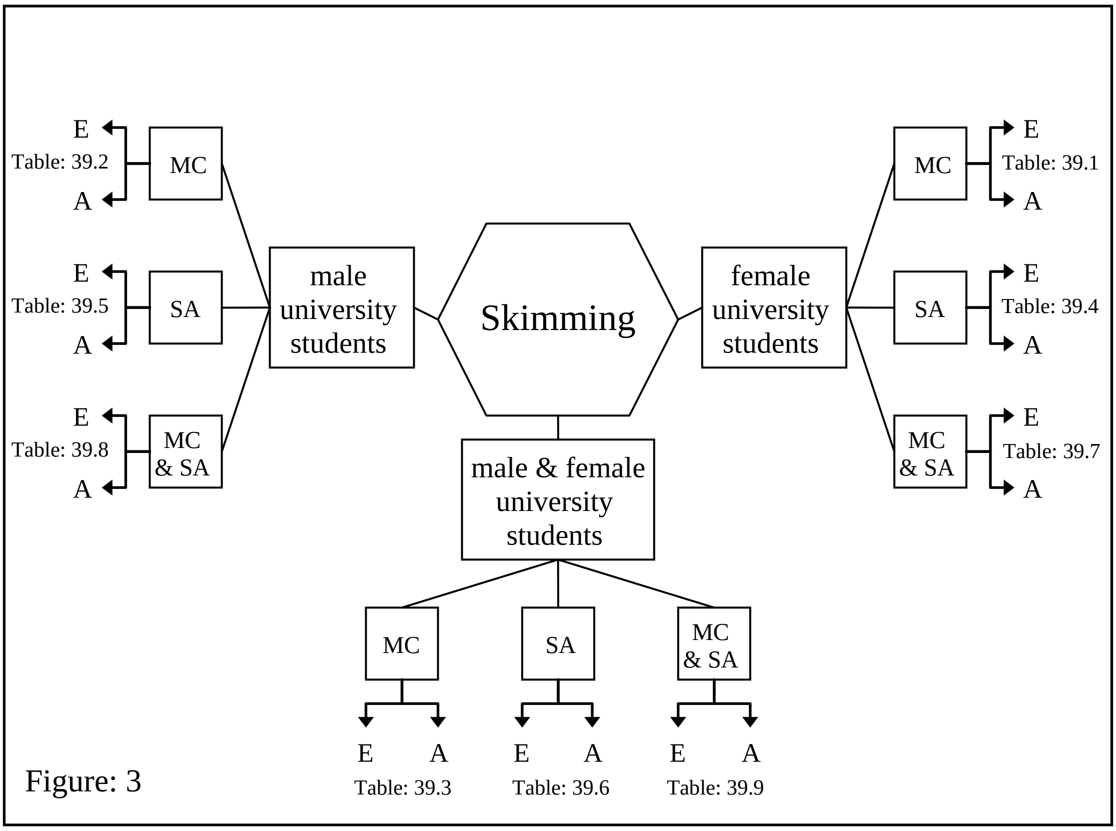


Figure: 3

The observed values of the t-test studies in table 33.1 (1.0037), table 33.2 (.6439), table 33.4 (- .2873), table 33.5 (- .1008), table 33.7 (.2888), and table 33.8 (.2695) are less than their critical value (3.646); and the observed values of the t-test studies in table 33.3 (1.1584), table 33.6 (- .2817), and table 33.9 (.3999) are less than their critical value (3.460). Therefore, hypothesis 6 with all its variables as in figure 3 above is rejected at p < .0005.

Table 33.1: Multiple-choice: university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .833 | 1.278 | 1.0037 |
| s | 1.111 | 1.515 | |

* P < .0005 , df = 34

Table 33.2: Multiple-choice: university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.167 | 1.5 | .6439 |
| s | 1.435 | 1.663 | |

* P < .0005 , df = 34

Table 33.3: Multiple-choice: all university students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1 | 1.389 | 1.1584 |
| s | 1.265 | 1.568 | |

* P < .0005 , df = 70

Table 33.4: Short answer: university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.667 | 1.5 | - .2873 |
| s | 1.815 | 1.663 | |

* P < .0005 , df = 34

Table 33.5: Short answer: university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.5 | 1.444 | - .1008 |
| s | 1.627 | 1.680 | |

* P < .0005 , df = 34

Table 33.6: Short answer: all university students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.583 | 1.472 | - .2817 |
| s | 1.699 | 1.648 | |

* P < .0005 , df = 70

Table 33.7: The whole test (MC+SA): university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 2.5 | 2.778 | .2888 |
| s | 2.776 | 2.990 | |

* P < .0005 , df = 34

Table 33.8: The whole test (MC+SA): university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 2.667 | 2.944 | .2695 |
| s | 2.890 | 3.281 | |

* P < .0005 , df = 34

Table 33.9:
The whole test (MC+SA): all university students (female + male)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 2.583 | 2.861 | .3999 |
| s | 2.793 | 3.094 | |

* P < .0005 , df = 70

### 6.6.3 Making Inferences or Drawing Conclusions

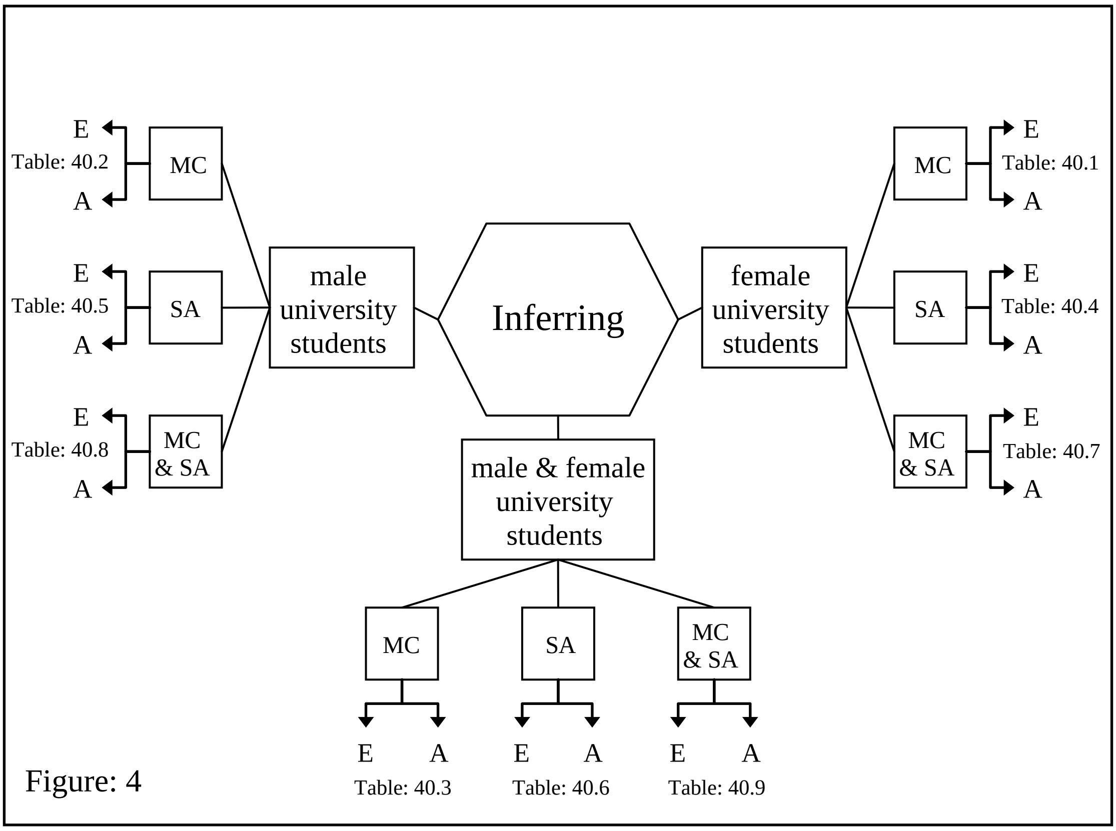


Figure: 4

The observed values of the t-test studies in table 34.1 (- 1.0308), table 34.2 (.1244), table 34.4 (0), table 34.5 (1.3477), table 34.7 (- .5195), and table 34.8 (.7655) are less than their critical value (3.646); and the observed values of the t-test studies in table 34.3 (- .5112), table 34.6 (.9860), and table 34.9 (.2752) are less than their critical value (3.460). Therefore, hypothesis 7 with all its variables as in figure 4 above is rejected at p < .0005.

Table 34.1: Multiple-choice: university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .833 | .5 | - 1.0308 |
| s | 1.163 | .728 | |

* P < .0005 , df = 34

Table 34.2: Multiple-choice: university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.056 | 1.111 | .1244 |
| s | 1.306 | 1.372 | |

* P < .0005 , df = 34

Table 34.3: Multiple-choice: all university students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | .944 | .806 | - .5112 |
| s | 1.219 | 1.082 | |

* P < .0005 , df = 70

Table 34.4: Short answer: university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .889 | .889 | 0 |
| s | 1.188 | 1.188 | |

* P < .0005 , df = 34

Table 34.5: Short answer: university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .667 | 1.222 | 1.3477 |
| s | .970 | 1.455 | |

* P < .0005 , df = 34

Table 34.6: Short answer: all university students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | .778 | 1.056 | .9860 |
| s | 1.069 | 1.309 | |

* P < .0005 , df = 70

Table 34.7: The whole test (MC+SA): university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.722 | 1.389 | - .5195 |
| s | 2.072 | 1.766 | |

* P < .0005 , df = 34

Table 34.8: The whole test (MC+SA): university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.722 | 2.333 | .7655 |
| s | 2.100 | 2.657 | |

* P < .0005 , df = 34

Table 34.9:
The whole test (MC+SA): all university students (female + male)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.722 | 1.861 | .2752 |
| s | 2.056 | 2.223 | |

* P < .0005 , df = 70

### 6.6.4 Identifying Referents of Pronouns

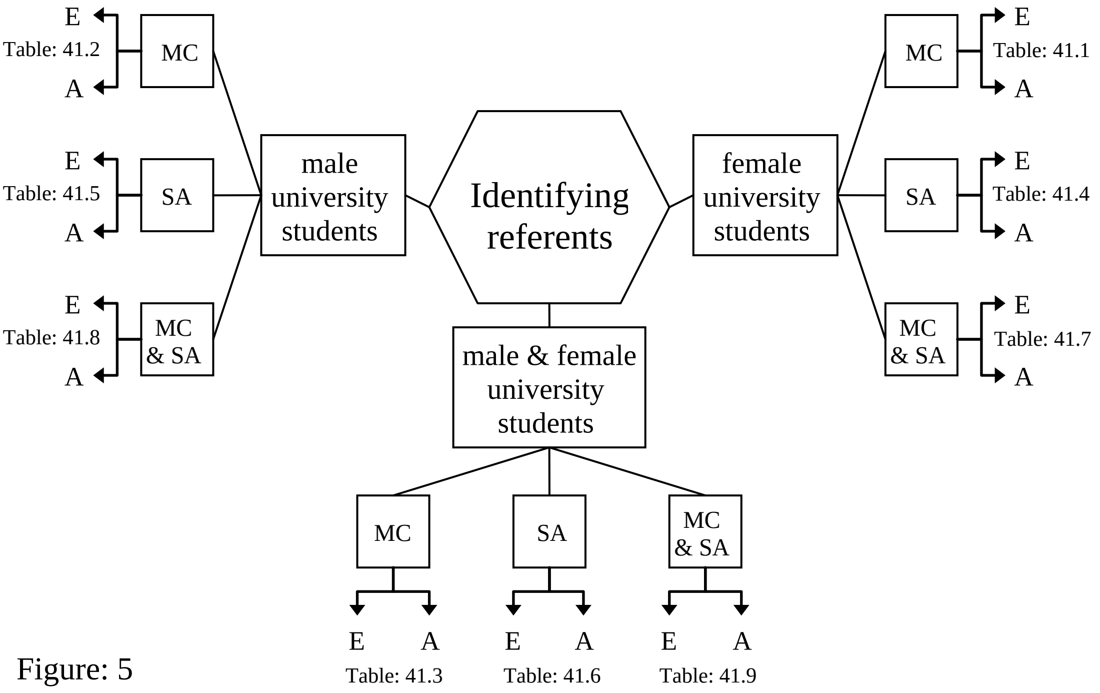


Figure: 5

The observed values of the t-test studies in table 35.1 (1.0624), table 35.2 (- .5936), table 35.4 (- .1388), table 35.5 (1.1199), table 35.7 (.4761), and table 35.8 (.2633) are less than their critical value (3.646); and the observed values of the t-test studies in table 35.3 (.1898), table 35.6 (.7521), and table 35.9 (.5140) are less than their critical value (3.460). Therefore, hypothesis 8 with all its variables as in figure 5 above is rejected at p < .0005.

Table 35.1: Multiple-choice: university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .667 | 1.056 | 1.0624 |
| s | .907 | 1.260 | |

* P < .0005 , df = 34

Table 35.2: Multiple-choice: university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.333 | 1.056 | - .5936 |
| s | 1.495 | 1.306 | |

* P < .0005 , df = 34

Table 35.3: Multiple-choice: all university students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1 | 1.056 | .1898 |
| s | 1.219 | 1.265 | |

* P < .0005 , df = 70

Table 35.4: Short answer: university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .889 | .833 | - .1388 |
| s | 1.188 | 1.213 | |

* P < .0005 , df = 34

Table 35.5: Short answer: university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .778 | 1.278 | 1.1199 |
| s | 1.138 | 1.515 | |

* P < .0005 , df = 34

Table 35.6: Short answer: all university students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | .833 | 1.056 | .7521 |
| s | 1.146 | 1.352 | |

* P < .0005 , df = 70

Table 35.7: The whole test (MC+SA): university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.556 | 1.889 | .4761 |
| s | 1.910 | 2.275 | |

* P < .0005 , df = 34

Table 35.8: The whole test (MC+SA): university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 2.111 | 2.333 | .2633 |
| s | 2.473 | 2.590 | |

* P < .0005 , df = 34

Table 35.9:
The whole test (MC+SA): all university students (female + male)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.833 | 2.111 | .5140 |
| s | 2.178 | 2.402 | |

* P < .0005 , df = 70

6.6.5 Guessing the Meaning of Unknown Words from Context

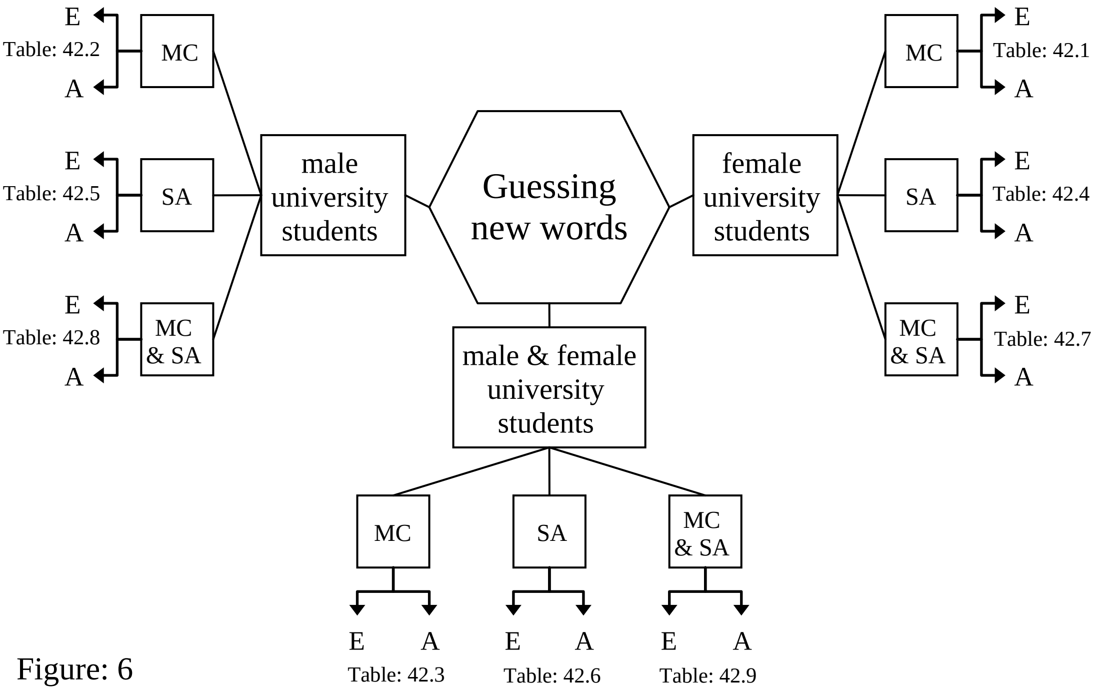


Figure: 6

The observed values of the t-test studies in table 36.1 (.6362), table 36.2 (1.2432), table 36.4 (1.0209), table 36.5 (- .3562), table 36.7 (.9639), and table 36.8 (.2491) are less than their critical value (3.646); and the observed values of the t-test studies in table 36.3 (1.3533), table 36.6 (.4410), and table 36.9 (.8561) are less than their critical value (3.460). Therefore, hypothesis 9 with all its variables as in figure 6 above is rejected at p < .0005.

Table 36.1: Multiple-choice: university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .389 | .556 | .6362 |
| s | .728 | .840 | |

* P < .0005 , df = 34

Table 36.2: Multiple-choice: university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .278 | .611 | 1.2432 |
| s | .642 | .940 | |

* P < .0005 , df = 34

Table 36.3: Multiple-choice: all university students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | .333 | .583 | 1.3533 |
| s | .676 | .878 | |

* P < .0005 , df = 70

Table 36.4: Short answer: university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .833 | 1.278 | 1.0209 |
| s | 1.163 | 1.435 | |

* P < .0005 , df = 34

Table 36.5: Short answer: university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.278 | 1.111 | - .3562 |
| s | 1.475 | 1.328 | |

* P < .0005 , df = 34

Table 36.6: Short answer: all university students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.056 | 1.194 | .4410 |
| s | 1.309 | 1.363 | |

* P < .0005 , df = 70

Table 36.7: The whole test (MC+SA): university female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.222 | 1.833 | .9639 |
| s | 1.645 | 2.128 | |

* P < .0005 , df = 34

Table 36.8: The whole test (MC+SA): university male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.556 | 1.722 | .2491 |
| s | 1.879 | 2.128 | |

* P < .0005 , df = 34

Table 36.9: The whole test (MC+SA): all university students (female + male)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.389 | 1.778 | .8561 |
| s | 1.740 | 2.098 | |

* P < .0005 , df = 70

## 6.7 Secondary School Students

### 6.7.1 Scanning a Text to Locate Specific Information

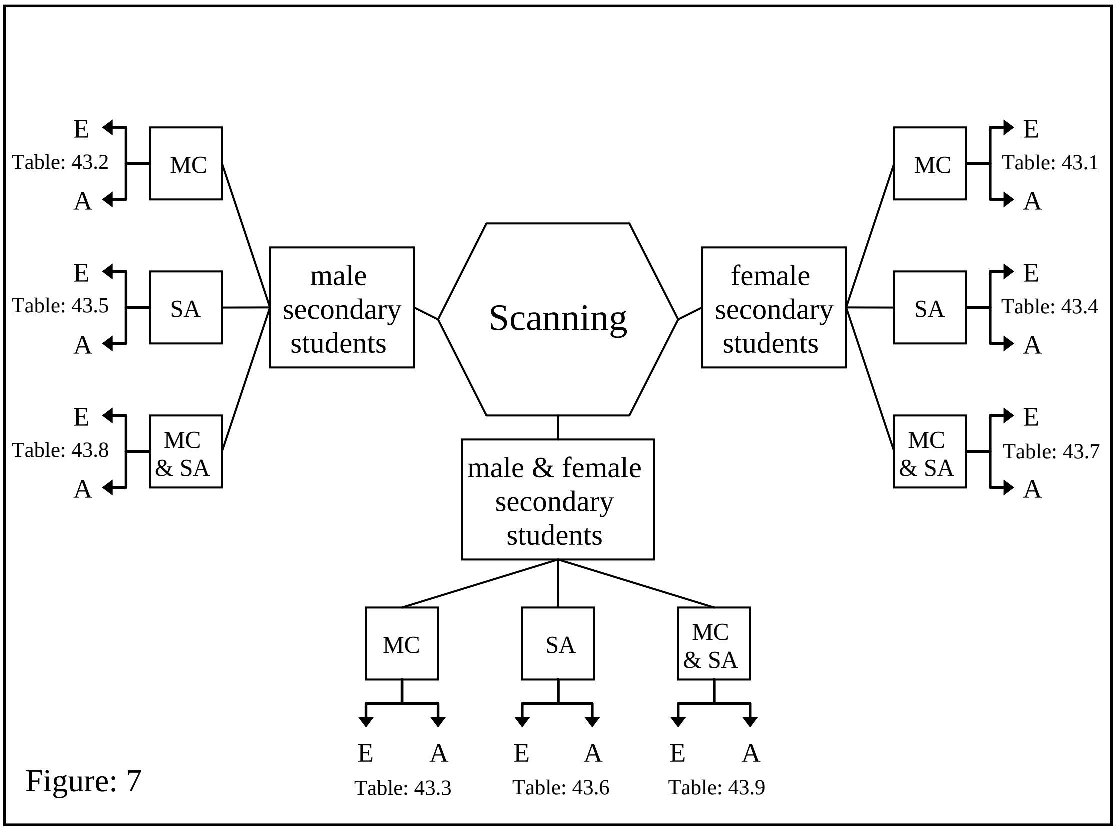


Figure: 7

The observed values of the t-test studies in table 37.1 (.2026), table 37.2 (- .5774), table 37.4 (- .0898), table 37.5 (.7413), table 37.7 (.0488), and table 37.8 (.0972) are less than their critical value (3.646); and the observed values of the t-test studies in table 37.3 (- .2832), table 37.6 (.4581), and table 37.9 (.1048) are less than their critical value (3.460). Therefore, hypothesis 10 with all its variables as in figure 7 above is rejected at p < .0005.

Table 37.1: Multiple-choice: secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.389 | 1.5 | .2026 |
| s | 1.627 | 1.663 | |

* P < .0005 , df = 34

Table 37.2: Multiple-choice: secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.722 | 1.389 | - .5774 |
| s | 1.831 | 1.627 | |

* P < .0005 , df = 34

Table 37.3: Multiple-choice: all secondary school students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.556 | 1.444 | - .2832 |
| s | 1.707 | 1.621 | |

* P < .0005 , df = 70

Table 37.4: Short answer: secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.778 | 1.722 | - .0898 |
| s | 1.879 | 1.831 | |

* P < .0005 , df = 34

Table 37.5: Short answer: secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.389 | 1.833 | .7413 |
| s | 1.663 | 1.925 | |

* P < .0005 , df = 34

Table 37.6: Short answer: all secondary school students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.583 | 1.778 | .4581 |
| s | 1.748 | 1.852 | |

* P < .0005 , df = 70

Table 37.7: The whole test (MC+SA): secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 3.167 | 3.222 | .0488 |
| s | 3.387 | 3.447 | |

* P < .0005 , df = 34

Table 37.8: The whole test (MC+SA): secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 3.111 | 3.222 | .0972 |
| s | 3.395 | 3.464 | |

* P < .0005 , df = 34

Table 37.9:
The whole test (MC+SA): all secondary school students (female + male)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 3.139 | 3.222 | .1048 |
| s | 3.342 | 3.406 | |

* P < .0005 , df = 70

### 6.7.2 Skimming a Text for the Main Idea

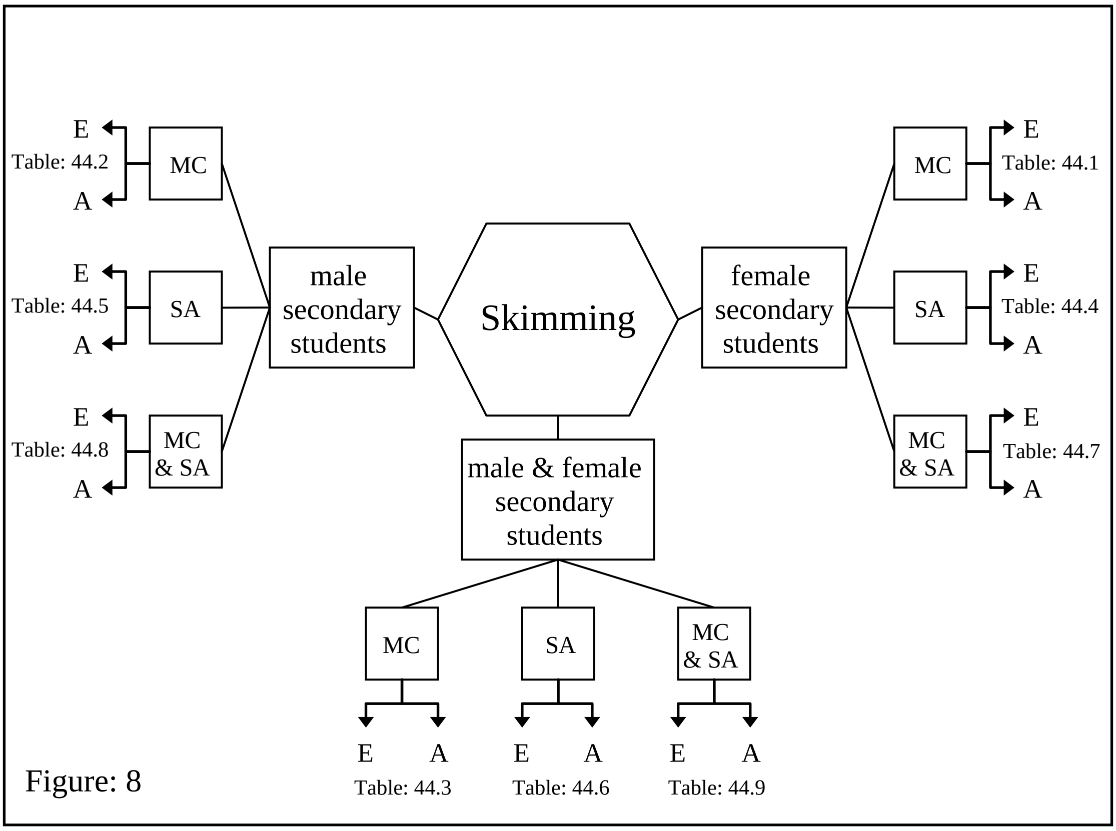


Figure: 8

The observed values of the t-test studies in table 38.1 (- .2845), table 38.2 (- .3812), table 38.4 (- .6014), table 38.5 (- .1030), table 38.7 (- .4566), and table 38.8 (- .2523) are less than their critical value (3.646); and the observed values of the t-test studies in table 38.3 (- .4774), table 38.6 (- .5102), and table 38.9 (- .5081) are less than their critical value (3.460). Therefore, hypothesis 11 with all its variables as in figure 8 above is rejected at p < .0005.

Table 38.1: Multiple-choice: secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.667 | 1.5 | - .2845 |
| s | 1.815 | 1.697 | |

* P < .0005 , df = 34

Table 38.2: Multiple-choice: secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.722 | 1.5 | - .3812 |
| s | 1.863 | 1.627 | |

* P < .0005 , df = 34

Table 38.3: Multiple-choice: all secondary school students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.694 | 1.5 | - .4774 |
| s | 1.813 | 1.639 | |

* P < .0005 , df = 70

Table 38.4: Short answer: secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.611 | 1.278 | - .6014 |
| s | 1.799 | 1.515 | |

* P < .0005 , df = 34

Table 38.5: Short answer: secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.5 | 1.444 | - .1030 |
| s | 1.663 | 1.572 | |

* P < .0005 , df = 34

Table 38.6: Short answer: all secondary school students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.556 | 1.361 | - .5102 |
| s | 1.707 | 1.521 | |

* P < .0005 , df = 70

Table 38.7: The whole test (MC+SA): secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 3.278 | 2.778 | - .4566 |
| s | 3.506 | 3.049 | |

* P < .0005 , df = 34

Table 38.8: The whole test (MC+SA): secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 3.222 | 2.944 | - .2523 |
| s | 3.464 | 3.134 | |

* P < .0005 , df = 34

Table 38.9:
The whole test (MC+SA): all secondary school students (female + male)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 3.25 | 2.861 | - .5081 |
| s | 3.435 | 3.047 | |

* P < .0005 , df = 70

### 6.7.3 Making Inferences or Drawing Conclusions

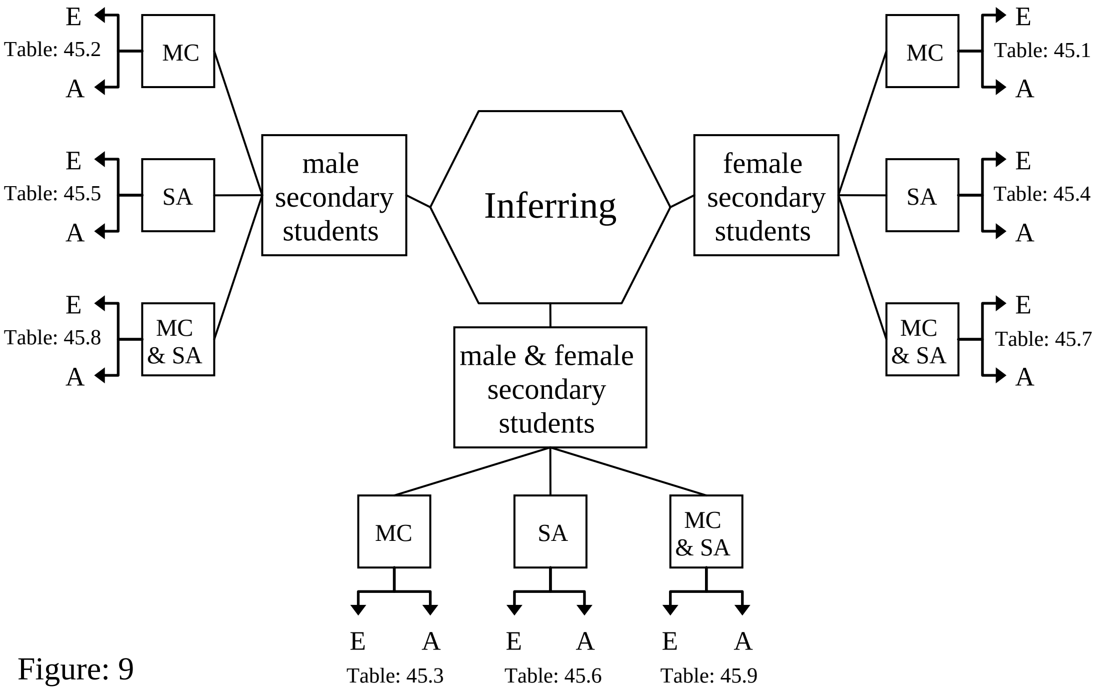


Figure: 9

The observed values of the t-test studies in table 39.1 (.1388), table 39.2 (0), table 39.4 (- .6969), table 39.5 (- .1080), table 39.7 (- .3314), and table 39.8 (- .0602) are less than their critical value (3.646); and the observed values of the t-test studies in table 39.3 (.0936), table 39.6 (- .5617), and table 39.9 (- .2712) are less than their critical value (3.460). Therefore, hypothesis 12 with all its variables as in figure 9 above is rejected at p < .0005.

Table 39.1: Multiple-choice: secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | .889 | .944 | .1388 |
| s | 1.188 | 1.213 | |

* P < .0005 , df = 34

Table 39.2: Multiple-choice: secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.111 | 1.111 | 0 |
| s | 1.372 | 1.328 | |

* P < .0005 , df = 34

Table 39.3: Multiple-choice: all secondary school students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1 | 1.028 | .0936 |
| s | 1.265 | 1.254 | |

* P < .0005 , df = 70

Table 39.4: Short answer: secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.333 | 1 | - .6969 |
| s | 1.572 | 1.283 | |

* P < .0005 , df = 34

Table 39.5: Short answer: secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.389 | 1.333 | - .1080 |
| s | 1.590 | 1.495 | |

* P < .0005 , df = 34

Table 39.6: Short answer: all secondary school students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.361 | 1.167 | - .5617 |
| s | 1.558 | 1.373 | |

* P < .0005 , df = 70

Table 39.7: The whole test (MC+SA): secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 2.222 | 1.944 | - .3314 |
| s | 2.612 | 2.413 | |

* P < .0005 , df = 34

Table 39.8: The whole test (MC+SA): secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 2.5 | 2.444 | - .0602 |
| s | 2.839 | 2.701 | |

* P < .0005 , df = 34

Table 39.9:
The whole test (MC+SA): all secondary school students (female + male)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 2.361 | 2.194 | - .2712 |
| s | 2.689 | 2.524 | |

* P < .0005 , df = 70

### 6.7.4 Identifying Referents of Pronouns

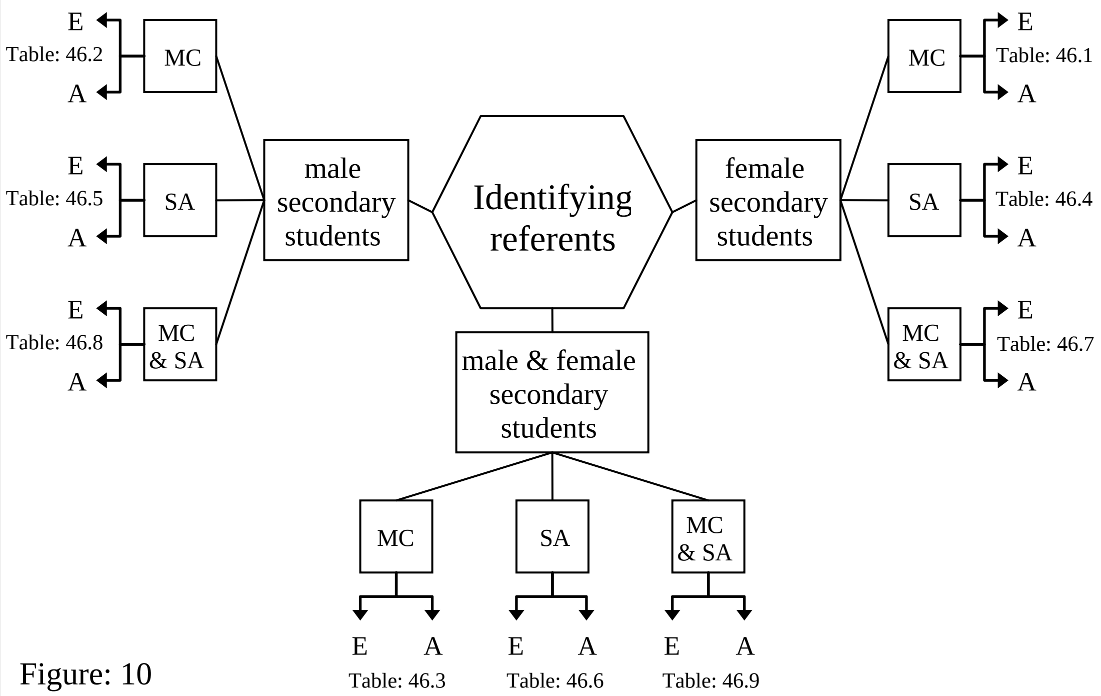


Figure: 10

The observed values of the t-test studies in table 40.1 (- .0939), table 40.2 (- .6689), table 40.4 (- .5611), table 40.5 (- .0922), table 40.7 (- .3347), and table 40.8 (- .3684) are less than their critical value (3.646); and the observed values of the t-test studies in table 40.3 (- .5102), table 40.6 (- .4664), and table 40.9 (- .5027) are less than their critical value (3.460). Therefore, hypothesis 13 with all its variables as in figure 10 above is rejected at p < .0005.

Table 40.1: Multiple-choice: secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.611 | 1.556 | - .0939 |
| s | 1.799 | 1.749 | |

* P < .0005 , df = 34

Table 40.2: Multiple-choice: secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.444 | 1.111 | - .6689 |
| s | 1.645 | 1.328 | |

* P < .0005 , df = 34

Table 40.3: Multiple-choice: all secondary school students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.528 | 1.333 | - .5102 |
| s | 1.699 | 1.531 | |

* P < .0005 , df = 70

Table 40.4: Short answer: secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.778 | 1.444 | - .5611 |
| s | 1.910 | 1.645 | |

* P < .0005 , df = 34

Table 40.5: Short answer: secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.722 | 1.667 | - .0922 |
| s | 1.831 | 1.782 | |

* P < .0005 , df = 34

Table 40.6: Short answer: all secondary school students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 1.75 | 1.556 | - .4664 |
| s | 1.844 | 1.690 | |

* P < .0005 , df = 70

Table 40.7: The whole test (MC+SA): secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 3.389 | 3 | - .3347 |
| s | 3.638 | 3.325 | |

* P < .0005 , df = 34

Table 40.8: The whole test (MC+SA): secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 3.167 | 2.778 | - .3684 |
| s | 3.370 | 2.951 | |

* P < .0005 , df = 34

Table 40.9:
The whole test (MC+SA): all secondary school students (female + male)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 3.278 | 2.889 | - .5027 |
| s | 3.456 | 3.098 | |

* P < .0005 , df = 70

## 6.7.5 Guessing the Meaning of Unknown Words from Context

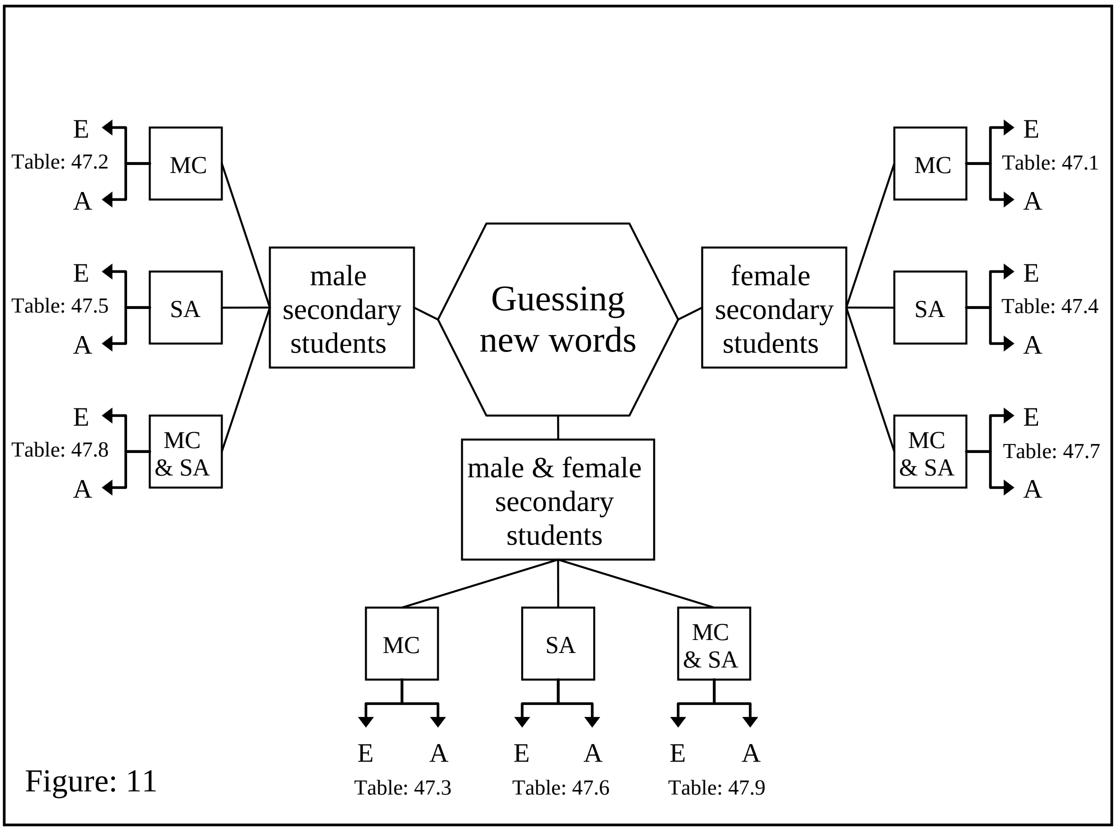


Figure: 11

The observed values of the t-test studies in table 41.1 (.3641), table 41.2 (-.0948), table 41.4 (1.2546), table 41.5 (.7792), table 41.7 (.7630), and table 41.8 (.2888) are less than their critical value (3.646); and the observed values of the t-test studies in table 41.3 (.1999), table 41.6 (1.4648), and table 41.9 (.7642) are less than their critical value (3.460). Therefore, hypothesis 14 with all its variables as in figure 11 above is rejected at p < .0005.

Table 41.1: Multiple-choice: secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 1.556 | 1.778 | .3641 |
| s | 1.749 | 1.910 | |

* P < .0005  , df = 34

Table 41.2: Multiple-choice: secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\bar{X}$ | 1.667 | 1.611 | - .0948 |
| s | 1.782 | 1.732 | |

* P < .0005 , df = 34

Table 41.3: Multiple-choice: all secondary school students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\bar{X}$ | 1.611 | 1.694 | .1999 |
| s | 1.740 | 1.797 | |

* P < .0005 , df = 70

Table 41.4: Short answer: secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\bar{X}$ | .833 | 1.389 | 1.2546 |
| s | 1.057 | 1.553 | |

* P < .0005 , df = 34

Table 41.5: Short answer: secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\bar{X}$ | .833 | 1.167 | .7792 |
| s | 1.111 | 1.435 | |

* P < .0005 , df = 34

Table 41.6: Short answer: all secondary school students (female + male )

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | .833 | 1.278 | 1.4648 |
| s | 1.070 | 1.474 | |

* P < .0005 , df = 70

Table 41.7: The whole test (MC+SA): secondary school female students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 2.389 | 3.167 | .7630 |
| s | 2.646 | 3.421 | |

* P < .0005 , df = 34

Table 41.8: The whole test (MC+SA): secondary school male students

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 18 | 18 | |
| $\overline{X}$ | 2.5 | 2.778 | .2888 |
| s | 2.733 | 3.029 | |

* P < .0005 , df = 34

Table 41.9:
The whole test (MC+SA): all secondary school students (female + male)

| Statistics | English | Arabic | $t_{obs}$ |
|---|---|---|---|
| N | 36 | 36 | |
| $\overline{X}$ | 2.444 | 2.972 | .7642 |
| s | 2.651 | 3.185 | |

* P < .0005 , df = 70

## 6. 8 Summary

In this chapter, I have attempted to analyse the quantitative data of the present study which seeks to examine the effect of using Arabic (L1) as a language for questions and answers in testing reading comprehension in English (L2) for university and secondary school students by using both multiple-choice and short answer questions. Besides the two main variables, the language of questions (L1, L2) and the testing method, there were other variables which were included, namely: gender and five sub-skills of reading comprehension. Each sub-skill was statistically analysed in relation to gender (M, F, “M +F”), testing method (MC, SA, “MC+SA”), and the proficiency level (post-beginner & upper-intermediate). A total of ninety-four t-test studies were carried out.

Unlike Shohamy (1984) who found that “the testing methods – MC, OE, and language of questions (L1, L2) – can make a difference in the assessment of the trait and can affect the scores that students obtain on reading comprehension tests” (p.157), the results of the four main t-test studies in the present research show that there are no significant differences in the performance of both university and secondary school students in both multiple-choice and short answer questions. In other words, the use of Arabic (L1) as a language for questions and answers did not improve students’ performance. However, the same findings of the present research study confirmed those of Hamdan & Diab (1997) where the use of Arabic (L1) did not help the high-level students, but only helped the low-level students, which was not the case in my present study. A possible reason for this might be the simplicity and clarity of the language used to form the questions in the English version of the test in the present research which might have made the two versions of the test (English and Arabic) of similar difficulty and therefore led to a similar performance.

The remaining ninety t-test studies covered the five reading sub-skills with their different combinations of the gender (M, F, “M +F”), testing method (MC, SA, “MC+SA”), and the proficiency level (post-beginner + upper-intermediate). The findings of all the ninety t-test studies showed that there were no significant differences in any of them. This confirms the results of the previous four main t-test studies where the use of Arabic (L1) as a language for questions and answers did not improve students’ performance.

The fact that students did not benefit from the use of L1 in the questions and answers might indicate that they had problems with understanding the passage itself. In other words, if the students were able to understand the passage, they would have been able to answer the questions regardless of their language, whether it is Arabic or English. This underlines the importance of teaching students the strategies and sub-skills of reading comprehension to enable them to understand and deal with different types of texts and questions.

# CHAPTER SEVEN: INTERVIEW ANALYSIS

## 7.1 Introduction

The interview is a traditional method of obtaining information, but “the term *interview*, is… of recent origin; it came into use in the seventeenth century” (Kvale, 2007, p. 5). Charles Booth is considered to be the first to use interviewing in social surveys (Fontana A., Frey J., 2000). The interview as Kvale (2007) defines it is “a conversation that has a structure and a purpose determined by the one party – the interviewer” (p.7). “The form and style of an interview is determined by its purpose” (Gillham 2000, p. 1).

In qualitative research, the interview is the most popular technique (Dörnyei, 2007). Fontana A. and Frey J. (2000) argue that the expansion of using interviews is due to two trends: psychological testing during World War I and clinical counselling. Sometimes, quantitative methods are not enough to investigate some research questions. Therefore, interviews can help in solving this difficulty (Banister, Burman, Parker, Taylor, and Tindall, 1994). Interviews give the study a human touch instead of pure dry numbers.

For some people, based on their culture, the term ‘interview’ may have negative implications and they might not accept to participate in the interview. To some of them it is usually connected with police investigations about state security or big crimes. Therefore, people do not feel comfortable in taking part in interviews. Thus, the interviewer can choose any other positive word such as ‘discussion’ or ‘chat’ to solve this linguistic problem (Gillham, 2000). In this study, the researcher explained to both students and teachers that it was not a formal interview, only a short ‘discussion’ or ‘chat’ about the test they had just taken. This was intended to encourage them to act as informants.

## 7.2 Interview Types

The interview is often classified into three different types based on their degree of structure (Dörnyei, 2007; Gillham, 2000; Kvale, 2007; Patton, 1990). These three types are: structured interview, unstructured interview and semi-structured interview. The following is a brief description of each type.

### 7.2.1 Structured Interview

In this first type of interview, “nothing is left to chance” (Fontana & Frey, 2000, p. 650). Everything is set in advance. That includes the organization, procedures, topics, questions wording and sequence (Richards, Platt, and Platt, 1992). The interviewer knows exactly what and how he or she will tackle the subject, and the interviewer has very limited freedom during the interview. A structured interview can be considered as a spoken questionnaire. According to Dörnyei (2007), “structured interviews are used in situations where a written questionnaire would in theory be adequate except that for some reason the written format is not feasible” (p. 135).

A limited range of interviewees’ responses is considered an advantage. It enables the researcher to compare responses easily. On the other hand, the restricted flexibility in the interviewer role and the limited variation of the interviewees’ responses are considered a disadvantage and have a negative effect on the interview richness and depth (Dörnyei, 2007).

### 7.2.2 Unstructured Interview

At the other end of the scale is the unstructured interview where there is only a simple interview schedule to help the interviewer especially with the opening questions (Dörnyei, 2007). Its open and flexible nature makes it a richer source of data than the other two types (Fontana A., Frey J., 2000). The unstructured interview "allows maximum flexibility to follow the interviewee in unpredictable directions, with only minimal interference from the research agenda" (Dörnyei 2007, p. 135). However, this flexibility makes it harder to analyze.

### 7.2.3 Semi-structured Interview

According to Dörnyei (2007), the semi-structured interview is the most used type in the filed of applied linguistics. It has a clear schedule with open-ended questions and prompts, yet there is flexibility in the questions wording and order. The interviewees are encouraged to express their viewpoints in a relatively open way. This type of interview is used especially when the researcher has a good background of his or her research questions and can develop an interview schedule that encourages interviewees to speak more about the topic which is not possible in the structured interview. In this study, I chose this type of interview because it suits my research needs and design.

The following abbreviations for both secondary and university teachers and students will be used in the rest of this chapter. They are:

- SS: Secondary school students (SS1, SS2, SS3, SS4, SS5, SS6)
- US: University students (US1, US2, US3, US4, US5, US6)
- ST: Secondary school teachers (ST1, ST2)
- UT: University teachers (UT1, UT2)

## 7.3 Student Interviews

The transcribed content was categorised based on the interview schedule. In other words, the researcher used the interview questions and prompts as a main source for categorization. A total of ten categories were used. They are:

- purpose of learning English
- difficulties in learning English
- easiest / most difficult language skill
- difficulties in English language tests
- students' preferred testing method in reading
- oral or written tests
- students' impression of using Arabic in the test
- anxiety in the Arabic test
- difficulty in the Arabic test
- reasons for mistakes

### 7.3.1 Purpose of learning English

English language has a special status among Saudi students. They believe that mastering English can help them in different aspects of their life. Therefore, large numbers of them enrol in private English-language teaching centres which are scattered all over the country.

The first category was an introductory question to the interview. Students were asked about their purpose in learning English. In their responses, they emphasized that English language is the key for success in life, study, job, and pleasure. Some of their answers were:

> SS1: Well, it is a job requirement. All jobs from low to high positions require the mastery of English … and it (English) can be considered as a cheque in your pocket.

SS4: Nowadays, English is an international language. In all distinguished universities if you want to enter the college of medicine, engineering, computer science, you will need English. So, at the end you have to learn it. To succeed in your life, you have to master English. It is the language of our time.

SS5: First, for the sake of my study. Learning English is the most important thing to learn. Second, it is the language of the world. When I buy any new gadget or anything else, its catalogues or literature are in English.
I: ok
SS5: In addition, all the new terminologies are in English. Also, When you travel abroad, you will need it.

US1: Frankly, for several reasons. First, I love the language.
I: ok
US1: My hobby is to learn new words whether in English or any other language.
I: nice
US1: Second, as you know, nowadays anyone who does not know English is considered illiterate.

US2: First, because I like it. Also, I feel that English is like a key to the other world. Through English, you can know anything you want.

US3: I wanted to do so (learn it) for a long time. Since I was in the secondary school, I wanted to master English. I have an ambition to learn English. Therefore, I built my skills according to that (ambition) and joined the English department.

### 7.3.2 Difficulties in leaning English

Vocabulary was the first difficulty in learning English for eight out of the twelve students. The rest mentioned grammar, pronunciation, listening, and writing.

The question was: What are the main difficulties you faced in learning English?

Some of their answers were:

SS1: It might be grammar.
I: Grammar?
SS1: Yes, because speaking and writing skills can be learned through practice, but in grammar … Sometimes, we are used to use some sentences which are grammatically incorrect.

SS3: The biggest difficulty we always face is the change of (English-language) teachers. Every year a new teacher. This means a new method of teaching , which eventually affects us negatively. For example, the teacher at the first grade of the intermediate stage does not explain well the basics (of the language), then the next teacher at the second grade of the intermediate stage explains things deeply. Therefore, we get lost.

SS4: Based on my own experience, learning English is easier than learning Arabic especially the grammar, but the problem of English is that I live in Saudi Arabia and I use Arabic all the time, and therefore I face difficulties mostly in the English vocabulary. When some one knows the vocabulary, he can communicate with others even though he does not know grammar that well.

SS5: The interaction with the teacher, I mean the teacher lectures all the time while we are only just listeners. Therefore, in regard to speaking, our language is weak. It affects our speaking skill. We need time to speak.

US2: In reading, for example, if I cannot understand the text, I stop immediately, I do not keep trying.
I: Why do you not understand the text … Is it because of the vocabulary, sentence structure or what?

US2: The problem is not from the structure. I do not understand because of the vocabulary.

US3: The difficulties … well, the understanding of what is said to me.

I: Do you mean listening?

US3: Yes, listening, also some times writing.

I: What is the problem with writing?

US3: The problem is that you cannot arrange your ideas. You just write whatever comes to your mind. That is the problem.

I: What about spelling?

US3: No, I used to memorize twenty words daily even before I joined the department. Most of the students ask me about the correct spelling.

US6: The difficulty I face in learning English is pronunciation, it is the most difficult thing.

I: What else?

US6: Another thing is forgetting new words. I do not use them much. There are words that I rarely use, therefore I forget them.

### 7.3.3 Easiest / most difficult language skill

Reading was chosen seven times as the easiest language skill, while speaking was chosen six times as the most difficult language skill. Listening and writing were chosen by almost equal number of students. The question was: of the four English language skills; reading, writing, listening, and speaking, which one is the easiest and which one is the most difficult? Some of their answers were:

SS1: Reading is the most difficult

I: Why?

SS1: Maybe because it sometimes requires high speed … you do not have the time to read carefully.

SS2: Reading is easy… speaking is the most difficult, I always try to speak with different people. However, even though I know high number of words, I found that I cannot understand what they are saying to me.

I: What about writing?

SS2: I do not know, but if you mean composition, I consider it the most difficult part in the textbook because I do not memorize writing conjunctions.

SS4: Well, reading is good to some extent. It is one of the easiest.

I: ok, what is the most difficult?

SS4: listening is not difficult … writing like composition is somehow difficult.

I: In what way it is difficult

SS4: writing requires that you know grammar and how to build a perfect sentence, correct spelling and also vocabulary knowledge. You might lose a whole sentence because you forgot one word.

SS6: Reading is the easiest but writing is not that much easy. Listening is a little bit difficult because in listening the language is fast. English is fast, it is difficult to compose it in your mind.

US1: I think the difficult skill, not the most difficult, but it needs practice is speaking. It needs dare, it needs breaking the language barrier (phobia). You have to speak with native speakers. You have to get used to it.

US2: The easiest is listening and speaking. The most difficult is writing.

I: Why is writing the most difficult?

US2: I have a spelling problem

US5: The most difficult skills for me might be listening and reading. In listening, you cannot cope with the recording questions. It is too fast.

I: What about reading?

US5: If I know the words then I will understand, but If I do not know some of the words, I will not understand.

I: What is the easiest skill?

US5: Writing?

I: Why is writing the easiest?

US5: We had a great teacher. He taught us for two consecutive semesters. He built a solid foundation.

I: You benefited a lot from him?

US5: Yes, we benefited a lot from him. At the beginning, we did not know how to write correctly. At the first semester, we wrote about thirty pages, each about different subject. Now, we finished writing about one hundred and fourteen pages together. Now we are able to write easily about any subject. We do not have any difficulties.

### 7.3.4 Difficulties in English language tests

This is an important category that is related directly to the main research question. Nine out of the twelve students were facing problems with the language of the questions. Their main concern was with the question structure and words. The question was: What are the major difficulties you face in English language tests? Some of their answers were as follows:

SS1: composition because I am used to writing compositions in my own style which I believe is correct but I receive low marks.

SS2: The biggest difficulty for me is time…. (Also) there is a difficulty when you ask me to write about a pure scientific subject which needs scientific terms…. Another problem is when the question contains key-words that I do not know. These key-words help in understanding the meaning. There is difficulty in such things.

SS4: Sometimes, I face difficulties with some questions even though the passage is clear, but the questions are indirect or they might have two correct answers based on my own point of view … it originally has only one answer but based on my understanding it has two answers.

SS6: Understanding the question, I do not understand it fully. For example, passage questions; he (the teacher) gives me a passage followed by questions. I might not understand what he wants from me in the question. If I understand the question, I would be able to answer it …. If I can translate it into Arabic, I would understand it.

US1: We suffer a lot from the passage.

I: The reading passage?

US1: A good example is the final test of the (college) English-language intensive course. We suffered a lot. It took most of our time. In fact, reading the passage took half of the total time of the intensive course test.

I: Was it difficult?

US1: It was difficult and, frankly, it was higher than our level…. Sometimes, the problem is that I do not know what they want. I read the questions, but I do not understand what is required.

US2: Well, sometimes, my problem is that I do not understand the question. I do not know what does it ask for.

I: The language of the questions?

US2: Yes, I do not understand it. That is the problem.

I: Writing the answers, is it sometimes a difficulty for you?

US2: Yes, I know the subject but I cannot write it down.

US3: In reading tests, you cannot arrange your time to read the passage quickly and answer the questions. Reading the passage takes most of the time and I cannot answer the questions. You cannot understand, there are several ideas in the same passage, there are a lot of ideas. You cannot decide on which you should concentrate. So, if you do not answer fast, time will finish.

### 7.3.5 Students' preferred testing method in reading

Multiple-choice and short answer methods were the most popular among students. Only two students chose the gap filling method. According to them, the multiple choice method is fast but confusing. In the short answer method, they felt they could write what they believe is the correct answer, but they are afraid of writing mistakes. The question was: What testing method do you prefer in reading tests; multiple-choice, short answer, true/false, gap filling, …etc? Some of their answers were:

SS1: Maybe it is gap filling because it is usually a sentence and you just need to fill the space with a word.

I: but, sometimes it is more than a word, you might need to write a sentence. It is not necessarily a single word.

SS1: a sentence is a different case, but we usually have a one-word space… Gap filling is the easiest.

SS4: Frankly, I hate true/false questions because it causes me a problem.

I: What is the problem?

SS4: In true/false?

I: Yes

SS4: Sometimes, test writers mean different things. Some of them concentrate on tiny details. So, he wants you to write a very specific answer. Others, no, they just want a general answer. Therefore, you give a very specific answer but the teacher considers it wrong because he wants a general answer.

I: ok, what about other methods?

SS4: I like short answer question

I: Why?

SS4: I can write whatever I want. I can guarantee to write what I think is the right answer, but multiple-choice, as in this test, you might get confused between two choices.

SS5: The multiple-choice is the easiest.

I: Why?

SS5: Because the answer is clear, and even if I do not know the answer, I can infer it, while in short-answer questions you cannot.

US1: Well, multiple-choice is easier for me because writing an answer requires many things. First, you need to be very careful with the grammar. In addition, you have to make sure of the spelling. Therefore, multiple-choice is better for me.

US2: Well, I think I prefer short answer…because your answer reflects your understanding, but in the multiple-choice you get confused. Sometimes, two answers are similar to each other.

US3: The best method is short answer.

I: Why?

US3: you can write your ideas fast, but in the multiple-choice you think of several ideas, this idea and this idea and that idea, so you get lost and you cannot decide which one the teacher wants, but in short answer you know what he wants.

US6: I prefer short answer questions more than multiple-choice because, sometimes, it confuses me. I feel confused, but (in the short answer) I write what I believe is the correct answer. Sometimes, I know the correct answer in the multiple-choice test but there is also another possible answer. I say to myself that the teacher will put the most difficult, he knows how students think, therefore he must have made the answer more difficult, and therefore I choose it.

### 7.3.6 Oral or written tests

Only two students preferred the oral test. Their reason is that they can express themselves in a better way, and can explain their answers more easily. On the other hand, ten students preferred the written test because it gives them enough time to think and revise their answers. The question was: If you have the passage with you, do you prefer oral or written reading questions? Some of their answers were:

SS1: Well, I prefer oral tests.
I: Why?
SS1: Maybe because there is a chance to express what you want.
I: Do you think that mistakes are more common in written tests?
SS1: Off course, one word can cause the mistake.
I: but the same thing can happen in the oral test.
SS1: Yes, but the teacher might correct the mistake and I will benefit from that.

SS4: Well, I prefer to write.
I: Why?
SS4: I do not know, but I feel that I might make mistakes when I speak. I prefer to write and revise what I have written. Something in front of me.

SS5: I prefer oral.
I: Why?
SS5: because the teacher might understand me better. You can explain your answer. Writing might be more difficult.

US1: Well, the written test is better.

I: Why?

US1: In the oral test, you have to answer the question immediately, but in the written test, you can think.

US3: Written test is better.

I: Why?

US3: You can gather your ideas, but in the oral test, you ask me a question, I cannot understand the question, but with the written question you can re-read the question again and again, you can understand.

US6: The written test

I: Why?

US6: The oral test, sometimes, confuses me. In the written test, you can see it. In the oral test, you hear only once, and that is it.

### 7.3.7 Students' impression of using Arabic in the test

Eight students liked the idea of using Arabic in reading comprehension tests. They felt that it is easier and more relaxing. In addition, it saves their time and guarantees that questions will be understood. On the other hand, four students preferred the English version of the test. It is easier for them to think, read and answer in one language. The question was: What do you think of using Arabic language in the test questions and answers? Some of the answers were:

SS1: Well, I believe it might be good and acceptable because students, in general, are more competent in Arabic … and they can write the answer in a better way.

SS2: The difficulty is how to translate the information or how to convey it.

I: ok, but if you have the choice, what would you choose, the Arabic or the English version of the test?

SS2: The Arabic …. When the question are in Arabic, they might be some confusion, because you need to go back to the English passage and deal with it,, then go back to the questions and deal with them in Arabic. If my mastery of English is high, I will prefer it.

SS3: I think that English is better, anyone who can understand the passage will be able to answer the questions, but if cannot understand the passage, then he will not answer the questions correctly even in Arabic.

I: ok, but if had the choice, what would you choose?

SS3: If I am given the choice, I will choose Arabic because I can guarantee that I will understand the questions, but I prefer English.

I: In what test can you get higher score?

SS3: I can get higher score in the Arabic test.

SS4: As an Arab, I prefer it in Arabic because I will understand the questions, and I can find the appropriate expression for my answer in the Arabic language.

SS5: I think it is brilliant because it enable the student to understand the question better. The problem is that there are students who understand the passage but cannot understand the questions. If he understands the question, he will answer in a better way.

SS6: Outstanding.

I: Why?

SS6: As I said before, I cannot understand what he (the teacher) wants from me in the English questions. In the Arabic questions, I understood what he wanted, and therefore I answered immediately. No problem at all even if I was asked to write the answer in English.

US2: Well, not bad, but it would be better if the questions were in English.

I: Why?

US2: You feel that you can get inside the test; there is a harmony between reading and the answering.

US3: I feel that the use of English language is better because one can know how to arrange his ideas as it has been written, but in Arabic he needs to translate. Sometimes, some words come in the test which you know their meaning in English but not in Arabic. Sometimes, it causes difficulty to you. The English test is better.

### 7.3.8 Anxiety in the Arabic test

Three students said that using Arabic will not make them feel comfortable. The other nine students said that using Arabic will reduce their anxiety because it is their native language. In addition, there will be no more writing mistakes. The question was: If you have two tests, one in Arabic and the other in English, which one will make you more comfortable and less anxious? Some of their answers were:

SS1: Well, I believe it is the English version because we are used to it. Arabic might take more time to get accustomed to.

SS3: It is more relaxing than the English test. However, the English test has some advantages. Sometimes, I cannot figure out the answer, but from the structure of the (English) question, I can tell what he wants to ask about in the passage. Therefore, I can answer.

SS4: Words that I do not know in English, I supposedly know them in Arabic. Therefore, I will be more confident with my answers in Arabic.

US3: As a student in the English department, I think I will feel less anxious in the English test.

US6: Well, I am relaxed with both of them. However, the English test is definitely more difficult than the Arabic, because you studied Arabic in the intermediate and secondary school, but English is new to you and needs a lot of effort.

### 7.3.9 Difficulty in the Arabic test

Most of the students felt that the Arabic test is easier. They said that they could perform better and get higher scores in Arabic tests. The question was: Which one do you think is easier: the English or the Arabic version of the test?

SS3: I can guarantee higher performance in Arabic tests.

SS4: Arabic is easier, but I think it will not be good for the language learning.

US3: Arabic might be a solution for low-level students, but high-level students will not benefit from the use of Arabic.

US4: If the questions were in English, I would have left some questions without an answer.

### 7.3.10 Reasons for mistakes

When students were asked about their mistakes in the test, they mentioned different reasons such as: lack of concentration, new vocabulary in the question, indirect questions, similar words, guessing, and time pressure.

SS1: We are used to direct questions only.

SS2: The order of the questions is different from the order of the information in the passage.

US1: I only concentrated on the first paragraph.

US6: There are inferential questions. I didn’t expect that.

## 7.4 Teacher Interviews

The four teachers in this study are Arabs but not Saudis. They were chosen because of their long teaching experience. In addition, they represent the majority compared to the Saudi English language teachers. The teaching experience for the four teachers in this study ranges from 12 to 27 years. The transcribed content of the interviews was arranged into four categories. They are:

- Language used in teaching reading
- Language used in assessing reading
- Good students … bad performance in reading tests
- What teachers think of using Arabic in the test

### 7.4.1 Language used in teaching reading

The four teachers have similar views in regard to the language used in teaching reading. They use English nearly all the time, but sometimes they are forced to use Arabic when they find it difficult to explain a new vocabulary item. The question was: Do you use Arabic in teaching? Some of their answers were:

ST1: Well, we are trying our best to minimize the use of Arabic in teaching, but sometimes it is a must. Always, inside the class, you will find students, if you do not use the native language, most of them will not follow you because their levels of English vary a lot.

ST2: We use Arabic when we deal with vocabulary and sometimes when we give instructions. Actually, students always really tend to translate them into their native language.

UT1: The language I use in teaching English in class is 99.5% English. I use Arabic only when I explain a word and I find that students cannot understand what is said in all ways, in this situation only I say it very fast in Arabic.

UT2: In regard to the reading skill in particular, it is one of the skills where I never use the native language, however, when I teach vocabulary, grammar, or writing I might be forced to use the native language because of the skill nature.

### 7.4.2 Language used in assessing reading

None of the four teachers thought of using Arabic in assessing English language. It was a new idea for them. The question was: Have you ever used Arabic in testing English (L2)? Some of their answers were:

ST1: I never used it before

ST2: No, no, we did not use Arabic language in testing reading skill in English

UT1: It never happened that I used Arabic in testing

UT2: Never, we teach English language and all the tests are always in English. That is how we used to do. It never came to my mind to use the Arabic language. It is a new idea.

### 7.4.3 Good students .. bad performance in reading tests

Two teachers said that good students usually don’t have problems in the test. The other two teachers mentioned some possible reasons for that. The question was: Why do some good students perform badly in reading comprehension tests? Some of their answers were:

ST1: Well, this is sometimes because of the vocabulary. They cannot understand the passage 100% … but usually the distinguished students do not face problems in this matter.

ST2: I do not agree with you, good students are always good even in reading. I rarely face students who are good in other skills and at the same time face difficulties in reading.

I: They are good in reading, but in the reading test they perform badly, just in the test. Why?

ST2: Sometimes, it might be due to the test itself. How teachers design their tests is a factor which affects this matter. Also, the variety of questions might be a factor, and the difficulty of the passage, sometimes if you choose a difficult passage or a complicated passage, it might also affect them. So, there are many reasons for that.

UT1: Until now, I have never seen a student who is good or excellent in class and his performance is bad in the test, never happened.

UT2: That is true, it might be due to non-linguistic reasons, but based on my experience, those students who might be distinguished in class but perform badly in the test, this might be due to his personal characteristics, he might have test phobia. Reasons are related to the student character, but if he is a normal student who does not have test phobia, if he is distinguished in class, he will be distinguished in the test.

### 7.4.4 What teachers think of using Arabic in the test

Teachers did not undertake the test. They only looked at it for few minutes, and listened to a brief introduction about the research idea. The question was: What do you think of using Arabic L1 in testing reading comprehension in English L2? Some of their comments were:

ST1: Well, it is a new experience to me, so my opinion might not be that good, we are used to ask question in English, and we prefer to use English. I have some concerns about it. I am afraid it might be applied to other skills.

ST2: I think it might have a psychological effect.

I: Good or bad?

ST2: I think it may be good because they can express themselves and they can explain their ideas and thoughts in Arabic more than expressing them in English. So, this will support them psychologically and may reduce some stress or so.

UT1: Well, The use of Arabic in the questions of reading tests, personally, after seventeen years of teaching, I don’t like this method. In regard to the students’ psychological comfort, it is important but there should be some kind of challenge because the student is learning a new language. It is not an easy task, he should make a lot of effort. Challenge is preferable. The student should not be stress-free 100%; the student should be challenged and work hard.

UT2: Well, I don’t prefer this idea because the use of Arabic language, even in teaching, based on my humble readings. It is a necessity to use L2 the target language in teaching and assessment.

I: Ok, the use of Arabic in testing, does it make the student more or less anxious?

UT2: It makes him less anxious, for sure.

I: Why?

UT2: The use of the native language, I myself, because I speak with you in Arabic I am more relaxed than if I were speaking in English even though I have been teaching English for twelve years. It is different. The native language is different from that we learn.

## 7.5 Summary

In this chapter, I have tried to analyse the qualitative data of the present research which aims to explore the opinions of students and teachers regarding the use of Arabic (L1) as a language for questions and answers in reading comprehension tests. Sixteen semi-structured interviews were carried out. The participants were twelve students and four English-language teachers of university and secondary school level.

The interview responses showed that most of the students felt that the use of Arabic (L1) in the reading test would make it easier and more relaxing. Moreover, they emphasised that using Arabic would help them to understand the questions and to answer them. However, the results of the present research showed that providing the questions in the students' native language did not make the test any easier, and did not improve the students' performance. In other words, the use of Arabic did not help students to avoid making mistakes or to write the correct answer.

In general, the interview responses were mixed, and there were good arguments on both sides. The qualitative analysis gave a more multi-faceted picture, but there was no consensus in favour of the use of Arabic (L1) as a language for questions and answers in reading comprehension tests. The conclusion chapter discusses the results of the interviews in further detail.

# CHAPTER EIGHT: CONCLUSION

## 8.1 Introduction

Reading is a complex cognitive skill that can be affected by various text, reader, and test variables. A considerable amount of research has been devoted to examining these variables. This study is part of the effort to understanding some aspects of reading comprehension and its assessment.

The present research is probably very unusual in the extent to which its initial hypotheses have not been supported. My informal general hypothesis was that providing reading comprehension questions and answers in Arabic might in some cases lead to better, and in some sense more valid tests that would be more acceptable to students and teachers. This rather loose idea was refined into a series of precise hypotheses which were tested in the main quantitative part of the research, and was also the main issue, though not the only one, which I tried to explore in the interviews.

As reported in chapter six, the quantitative findings were uniformly and rather astonishingly negative, in the sense that no significant differences were found. The interview results gave a more multi-faceted picture, presented in chapter seven and commented on further below, but again the main findings could be described as 'negative' in that there was no clear support for a change to questions and answers in Arabic; responses were mixed, with good arguments on both sides.

Some readers might conclude from the above that this research has been unsuccessful, but I would argue the exact opposite. It is well established, and often mentioned in the literature on research methods, that 'negative' results can be of great value. There is even some suggestion that they may be more useful in certain senses than ‘positive’ results, because the tendency only to report significant correlations and confirmed hypotheses, and simply not to publish the others, can introduce a distortion. Ioannidis (2005), although in a different field (medicine), makes important general points about the danger of ‘false positives’.

> Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs…. “Negative” research is also very useful. “Negative” is actually a misnomer, and the misinterpretation is widespread (p. 0696).

Of course, absence of evidence is not evidence of absence, and the non-confirmation of a hypothesis does not confirm its opposite. In this case, then, I have not demonstrated that use of Arabic in English-L2 reading comprehension questions and answers would never, or even rarely, make them easier or ‘better’; in fact I still suspect it quite often would, and I will return to this later in this chapter. Basically, however, the overwhelmingly ‘negative’ results should not be brushed aside: my research has made me more aware of arguments in favour of English in questions, and of the need to be cautious in advocating change.

The remainder of this chapter is organised as follows:

Section 9.2 is a brief restatement of, and commentary on, the results of the quantitative study.

Section 9.3 is a synopsis of the interview results highlighting a few points especially relevant to my initial questions.

Section 9.4 is a reflection on the successes and limitations of my research procedures: what seemed to work well, what I could change in any future similar research, and advice to others who may wish to follow in my footsteps.

Section 9.5 suggests some specific questions for future research in areas related to the present study.

## 8.2 Quantitative study summary and comments

The present study aimed at investigating the effect of using Arabic (L1) as a language of questions and answers in testing reading comprehension in English (L2) by using both multiple-choice and short answer questions for upper-intermediate and post-beginner students. Four main case-II independent sample t-test studies were carried out, namely:

| | |
|---|---|
| First t-test study: | Test A1 (Upper intermediate - English MC) and |
| | Test A2 (Upper intermediate - Arabic MC) |
| Second t-test study: | Test A1 (Upper intermediate - English SA) and |
| | Test A2 (Upper intermediate - Arabic SA) |
| Third t-test study: | Test B1 (Post beginner - English MC) and |
| | Test B2 (Post beginner - Arabic MC) |
| Forth t-test study: | Test B1 (Post beginner - English SA) and |
| | Test B2 (Post beginner - Arabic SA) |

For the second year English department students (upper-intermediate), there was no change in their level of performance in either the multiple-choice or the short answer tests. The statistical analysis of each of the two t-test studies showed that there was no significant difference between the two groups in either study. The use of Arabic (L1) as the language for questions and answers did not improve their performance. These similar results could be due to students' experience in English language tests as they are majoring in English in the College of Languages and Translation. Moreover, the simplicity and clarity of the questions in English (L2) might make them of similar difficulty with the questions in Arabic (L1). In other words, to them, using Arabic did not make the test easier, and the two versions of the test were of similar difficulty. Therefore their performance in the two tests did not change.

Similarly, there was no significant difference in the performance of the final-year secondary school students (post-beginner) in either the multiple-choice or the short answer tests regardless of the test language. Providing the reading comprehension questions and answers in Arabic (L1) did not improve students levels of performance. These similar results could also be due to the simplicity and clarity of the language of the questions in English (L2). However, if they were given more difficult questions with longer answers like wh-questions, their levels of performance might show significant difference between the two tests, in favour of Arabic. To me, this is an issue that I would like to investigate in my future research.

Besides the two main variables, the use of L1 and the testing method, there were other variables which were included, namely: gender and the following five sub-skills of reading comprehension.

1. Scanning a text to locate specific information.
2. Skimming a text for the main idea.
3. Making inferences or drawing conclusions.
4. Identifying referents of pronouns.
5. Guessing the meaning of unknown words.

Each sub-skill was statistically analyzed in relation to gender (M, F, M+F), testing method (MC, SA, MC+SA), and the proficiency level (post-beginner + upper-intermediate). All possible combinations were explored in order to make sure that all the possibilities of the controlled variables were considered in order to trace any possible source of significant difference. These combinations for each reading sub-skill were as follows:

Multiple-choice – female students
Multiple-choice – male students
Multiple-choice – female + male students

Short answer – female students
Short answer – male students
Short answer – female + male students

The whole test (MC+SA) – female students
The whole test (MC+SA) – male students
The whole test (MC+SA) – female + male students

A total of ninety t-test studies covering both university and secondary school students were carried out. Surprisingly, the results of all of these ninety studies were similar to those of the main study, and there were no significant differences in any of them. In other words, the use of Arabic in the questions and answers of the test did not improve or increase the performance of students in any of the five sub-skills with their different combination with gender, testing method, and the proficiency level. In a way, these results strengthen the results of the main study, and also they show how complex the reading skill is. The use of Arabic (L1), although it seems helpful in overcoming English (L2) difficulties in questions and answers, has not been proven to have a positive effect on students' performance. Therefore, in societies where Arabic is students' native language, test writers and developers should be cautious about using this language in testing reading comprehension.

## 8.3 Interviews summary and comments

A total of sixteen interviews were conducted in this study. The participants were six students and two English language teachers from Al-Rowad Secondary

School in Riyadh, and also six students and two English language teachers from the English Department at Al-Imam University in Riyadh. The interviews are part of the present study in which the use of Arabic (L1) in testing reading comprehension in English (L2) is investigated. The following summarizes the main remarks of the interviews.

1. The interviews with both university and secondary school students showed that they have a very positive attitude towards English (L2). They believe that mastering English will give them an advantage in their academic life, and will almost guarantee them a good job in the future. This positive attitude towards learning English (L2) made some of them say, during the interview, that they prefer to take the test in English (L2) just because they believe it could be part of their L2 learning, while taking the test in Arabic (L1) will not add anything to their L2 learning. However, when I asked them about the language of the test in which they might perform better and get higher marks, their answer was Arabic (L1).

2. The reading skill is considered, among the majority of students in the interview, as the easiest language skill. However, I noticed as a teacher that, to them, the term ‘reading’ is not necessarily related to comprehension, but it means the ability to pronounce letters and words correctly. I still remember that our English-language teachers in the intermediate school used to teach us phonics without any real concentration on comprehension. This may explain why most of them consider reading as the easiest language skill.

3. The category entitled ‘Difficulties in English language test’, in the students interview part, has some of the students complaints about the language of the English-language tests. This reminded me of the background of the present study

which I talked about in chapter one, where students used to complain about the entrance test of the College of Languages and Translation in Riyadh; in particular, the language of the test questions. However, results of the present research showed that providing the questions in students native language did not make the test any easier, and did not improve students' performance when compared to the English version of the test.

4. While most secondary school students in the interview chose multiple-choice as their preferred method of testing, most of the university- level students chose short answer. This might be related to students' proficiency level. It is possible that the university students preferred to write down their answers because their writing skill is better than the secondary school students who chose the multiple-choice method to avoid their weaknesses in writing.

5. Most of the university and secondary school students in the interview preferred a written test to an oral one. Their speaking ability might be a major factor here. They do not prefer oral tests because they, sometimes, face difficulties in understanding what is being said to them and in expressing their answers. In addition, they are under the pressure of time which is short and limited when compared to the time they have in the written test. Moreover, the ability to read the question again and again until they understand it is not available in the oral test. Their choice here is consistent with what is mentioned before that all university and secondary school students prefer to have the interview in Arabic (L1), which might reflect their anxiety of making mistakes or not understanding some of the questions in the interview. L1 was their choice because they are more competent in Arabic,

which may give them the power to finish the interview in a relaxing and positive way.

6. When students in the interview were asked about their impression of using Arabic (L1) in the reading test, most of them felt that it is easier and more relaxing. Moreover, they emphasized that using Arabic would save their time and would help them in understanding the questions and answering them. However, as stated before, the results of the present study showed that even though the use of L1 might have a positive psychological effect on the test-taker; nevertheless it did not improve their levels of performance. Since the performance is almost equal in the two versions of the test, it might be a good idea to investigate the use of L1 in reading tests just for the sake of reducing the anxiety level of test-takers.

7. At the other end of the line, a few students in the interview did not like the idea of using L1 in testing. They think that it is easier for them to read, think, and answer in one language. To them, only low-proficiency level students would prefer a test in their native language. In addition, they said that it is sometimes difficult to translate an English answer into Arabic. Furthermore, they mentioned the distortion they might have when dealing with two languages in one test. It is difficult for them to read an English text to write an Arabic answer. These interesting views reveal some illuminating remarks that support the use of L2 in testing L2 reading comprehension. Besides the results of the main study, these interesting remarks made me more aware of the need to be more cautious in advocating change in the language of the test even though it might look more valid.

8. When students were asked about their wrong answers in the test, they mentioned different reasons like unfamiliar vocabulary items in the questions, lack of

concentration, guessing, time pressure, and indirect questions. However, apart from the unfamiliar vocabulary items, all the other reasons could be found in both the Arabic and the English versions of the test. Therefore, it might be said that the language of the test in the present study (Arabic/English) was not a direct reason for either the wrong answers nor the correct ones. In other words, the use of Arabic did not help students to avoid mistakes or to write the correct answer. This, again, goes with the results of the present research which have shown that the levels of performance of both the university and secondary school students in the Arabic test are similar to those in the English one.

9. During the interviews, I noticed that, sometimes, there was a kind of confusion between the role of the test as a measure and the role of the test as part of the learning process. For example, some of the students and teachers reject the idea of using Arabic in testing reading comprehension just because it might have a negative effect on the learning process. They do not consider its positive or negative effect on the reliability and validity of the measurement itself. This might reflect an inadequacy of some of the EFL teacher preparation programs in Saudi Arabia which Al-Hazmi (2003) describes as “non-systematic and inadequate” (p. 341). This was discussed in more detail in chapter one.

## 8.4 Successes and Limitations

### 8.4.1 Sample

Most language research in Saudi Arabia is limited to male students for cultural and official reasons. However, this study succeeded in including equal numbers of female and male students at both the university and secondary level. The only part of the study where the female students were not included was the interviews, which would have been extremely difficult to conduct for cultural reasons. However, it is really important to include females in future research in order to obtain a broader and deeper view of the research questions. The only possible way to have interviews with female students whether at the university or secondary school level is through a female teacher or researcher, which emphasises the importance of training for both sexes in order to make sure that interviews are done in a professional way.

### 8.4.2 Administration

Before distributing the test papers, I gave the students a short introduction to my research and told them that I have two versions of the same test; one in English and the other in Arabic. Then I asked them about their language preference; but surprisingly most of them could not decide or said that both versions of the test were equally acceptable to them. Only a few students preferred one language to the other. Therefore, I distributed the test randomly with an equal number of each language. However, it might be useful in future similar research to distribute the two versions of the test according to students' preference and not randomly. This might help in obtaining more distinctive results in relation to the use of Arabic (L1) in testing reading comprehension in English (L2).

### 8.4.3 Test content

I developed two reading comprehension tests and piloted them before the study. The two tests were translated into Arabic, and were piloted at the same time as the English version of the test. The four tests were designed to match the needs of this study. They were limited in length and were developed to assess specific reading sub-skills. I was careful to use simple language in both the structure and vocabulary because, according to Hughes (2003), the language of the questions should not add any difficulty to the test. However, that simplicity might have made the questions very easy for students, and therefore the use of Arabic (L1) did not make the questions clearer or easier. This might be, as mentioned before, a possible reason for the similar performance by students in the two versions of the test. Translating and using the reading part of a well established test like the IELTS or TOFEL might make a difference in students' performance when compared with the English version of the test.

### 8.4.4 Length

For practical reasons, the length of the test was set to fit the class time which was forty five minutes in the secondary school and fifty minutes at the university. However, increasing the length of the test would make it more reliable, and might help in measuring students ability more accurately. The addition of a third passage, for example, with ten questions using a third testing method such as gap filling would increase test reliability and minimize the effect of the testing method in the test which would still be of acceptable length even after this addition.

## 8.5 Further research

This study is the first of its kind in Saudi Arabia in which Arabic (L1) was used in testing reading comprehension in English (L2), therefore, additional research could take this forward. However, in view of the limitations of the present study, the following recommendations for further research are offered for consideration.

1. Further research is needed to investigate the use of L1 in testing the other receptive skill, namely listening. Both reading and listening are receptive skills that could be assessed through the use of L1. Furthermore, a comparative study between the tests of the two skills might lead to better understanding of their assessment.

2. Saudi Arabia and other Arab countries have different strategies and policies in teaching and testing English as a foreign language. These differences might lead to different levels of performance if L1 is used in the questions or answers or both. Extension of similar research to include some of these countries might help educators and test developers to produce more reliable and valid reading tests.

3. The present study concentrated on only two levels; post beginner and upper intermediate. However, other levels such as advanced or beginner should also be investigated. Covering a wider range of levels of performance might help in generalizing the results.

4. In most classroom situations, Saudi teachers tend to assess their students' reading ability by written tests only. However, it might be illuminating to examine the use of L1 in oral questions and answers. Moreover, a comparative study is needed to compare the oral and written tests to show the differences and similarities when using L1 in testing reading. Such studies would have important implications for teaching and testing.

5. The current polices in Saudi Arabia do not allow the use of Arabic (L1) in teaching English as a foreign language. However, there seems to be a need for studies in which teachers are allowed to use Arabic in teaching English language,

and how this affects the performance of students as compared to that of a control group where English is the only medium of teaching. Similar tests to those in this study can be used where the questions and answers are in Arabic.

6. Lastly, although numerous studies on reading comprehension test taking strategies have been carried out, further research is needed to include reading tests where the questions, answers or both are in L1. Furthermore, it would be useful to cover several languages, different reading sub-skills, with various specialised and non-specialised texts. Such comprehensive treatment would enrich our understanding of this important field.

# REFERENCES

Abdan, A. (1991). An exploratory study of teaching English in the Saudi elementary public schools. System, 19, 253- 266.

Aebersold, J. A. & Field, M. L. (1997). From reader to reading teacher: Issues and strategies for second language classrooms. New York: Cambridge University Press.

Alabdelwahab, S. Q. (2002). Portfolio assessment: A qualitative investigation of portfolio self-assessment practices in an intermediate EFL classroom, Saudi Arabia. Unpublished doctoral dissertation, The Ohio State University, OH.

Al-Abed al Haq, F., and Smadi, O. (1996). The status of English in the Kingdom of Saudi Arabia (KSA) from 1940-1990. In JA Fishman, AW Conrad, & A. Rubal-Lopez (Eds.), Post-imperial English: Status change in former British and American colonies, 1940-1990 (pp. 457-84). Berlin, New York: Mouton de Gruyter.

Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J.C. Alderson & A.H. Urquhart (Eds.), Reading in a foreign language (pp.1-27). New York: Longman.

Alderson, J. C. (1990). Testing reading comprehension skills (part one). Reading in a Foreign Language, 6(2), 425–438.

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

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). Language Teaching, 35, 79-113.

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

Alderson, J. C. & Lukmani, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. Reading in a Foreign Language 5 (2), 253-70.

Alderson, J. C. & Urquhart, A. (1988). This test is unfair: I'm not an economist. In Carrell, P., Devine, J., & Eskey, D. (Eds.), Interactive approaches to second language reading (pp. 168-182). Cambridge: Cambridge University Press.

Al-Hazmi, S. (2003). EFL teacher preparation programs in Saudi Arabia: Trends and challenges. TESOL Quarterly, 37 (2), 341-344.

Al-Jarf, R. (2001). Processing of cohesive ties by EFL Arab college students. Foreign Language Annals, 32(2), 141-151.

Al-Seghayer, K. (2005). Teaching English in the Kingdom of Saudi Arabia: Slowly but steadily changing. In Braine, G. (Ed.) (2005). Teaching English to the World: History, Curriculum, and Practice. Mahwah, NJ: Lawrence Erlbaum.

Armbruster, B., Anderson, T., & Ostertag, J. (1987). Does text structure/summarization instruction facilitate learning from expository text? Reading Research Quarterly. 22, 331-346.

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

Bachman, L. (1993). What does language testing have to offer? In Silberstein S. (Ed.), State of the art TESOL essays (pp. 169-202). Bloomington, IL: Pantagraph Printing.

Bachman, L. (2004). Statistical analyses for language assessment. Cambridge: Cambridge University Press.

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. Journal of Experimental Psychology: Learning, Memory and Cognition, 28 (4), 748-769.

Banister, P., Burman, E., Parker, I., Taylor, M., & Tindall, C. (1994). Qualitative methods in psychology: A research guide. Buckingham: Open University Press.

Barnett, M. (1988). Reading through context: How real and perceived strategy use affects L2 comprehension. The Modern Language Journal, 72 (2), 150-162.

Berkemeyer, V. C. (1994). Anaphoric resolution and text comprehension for readers of German. Die Unterrichtspraxis/Teaching German 27(2), 15–22.

Bernhardt, E. (2005). Progress and procrastination in second language reading. Annual Review of Applied Linguistics, 25, 133-150.

Brantmeier, C. (2003). Does gender make a difference? Passage content and comprehension in second language reading. Reading in a Foreign Language, 15(1), 1-27.

Brown, J. D. (2005). Testing in language programs: A comprehensive guide to English language assessment. (New edition). New York: McGraw-Hill.

Bügel, K. & Buunk, B. P. (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. Modern Language Journal, 80(1), 15-31.

Cain, K. & Oakhill, J. V. (1999). Inference-making ability and its relation to comprehension failure in young children. Reading and Writing, 11, 489-503.

Cain, K. & Oakhill, J. V. (2006). Assessment matters: Issues in the measurement of reading comprehension. British Journal of Educational Psychology,76, 697-708.

Cain, K. & Towse, A. S. (2008). To get hold of the wrong end of the stick: Reasons for poor idiom understanding in children with reading comprehension difficulties. Journal of Speech, Language, and Hearing Research, 51, 1538-1549.

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference making ability and their relation to knowledge. Memory and Cognition, 29 (6), 850-859.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. Journal of Educational Psychology, 96(4), 671-681.

Carrell, P. (1985). Facilitating ESL reading by teaching text structure. TESOL Quarterly, 19(4), 727-752.

Carrell, P. (1987). Content and formal schemata in ESL reading. TESOL Quarterly, 21(3), 461-481.

Casteel, M. A. (1993). Effects of inference necessity and reading goal on children's inferential generation. Developmental Psychology, 29(2), 346–357.

Chang, C. (2006). Effects of topic familiarity and linguistic difficulty on the reading strategies and mental representations of nonnative readers of Chinese. Journal of Language and Learning, 4(2), 172-198.

Chapelle, C. (1999). Validity in language assessment. Annual Review of Applied Linguistics, 19, 254-272.

Clapham, C. (1998). The effect of language proficiency and background knowledge on EAP students' reading comprehension. In Kunnan, A. J. (Ed.), validation in language assessment (pp. 141-168). Mahwah, NJ: Lawrence Erlbaum Associates.

Cooper, M. (1984). Linguistic competence of practised and unpractised non-native speakers of English. In J.A. Alderson & A.H. Urquhart (Eds.), Reading in a foreign language (pp. 122-138). London: Longman.

Crawley, R. A., & Stevenson, R. J. (1990). Reference in single sentences and in texts. Journal of Psycholinguistic Research, 19(3), 191-210.

Demel, M. C. (1990). The relationship between overall reading comprehension and comprehension of coreferential ties for second language readers of English, TESOL Quarterly, 24(2), 267-292.

Demel, M. C. (1994). The relationship between overall comprehension and coreferential tie comprehension for second language readers of Spanish literature. Linguistics and Education, 6(3), 289-309.

Dörnyei, Z. (2007). Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies. Oxford: Oxford University Press

Duggan, G. B., & Payne, S. J. (2009). Text skimming: The process and effectiveness of foraging through text under time pressure. Journal of Experimental Psychology: Applied, 15(3), 228-242.

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? Educational Researcher, 36, 449-455.

Eskey, D. & Grabe, W. (1988). Interactive models for second language reading: perspectives on instruction. In Carrell, P., Devine, J., & Eskey, D. (Eds.), Interactive approaches to second language reading (pp. 223-238). Cambridge: Cambridge University Press.

Fontana, A. & Frey, J. (2000). The Interview: From structured questions to negotiated text. In Denzin, N. & Lincoln, Y. (Eds.), Handbook of qualitative research (2nd Ed.) (pp. 645-672). Thousand Oakes, CA: Sage Publications, Inc.

Geva, E. (1983). Facilitating reading comprehension through flowcharting. Reading Research Quarterly, 18, 384-405.

Gillham, B. (2000). The Research Interview. London: Continuum

Gomez, P. G. , Noah, A. , Schedl, M. , Wright, C. & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. Language Testing 24(3), 417-444.

Goodman, K. S. (1969). Analysis of oral reading miscues: Applied psycholinguistics. Reading Research Quarterly 5, 9-30.

Goodman, K. S. (1970). Reading: A psycholinguistic guessing game. In Gunderson, D.V., Language and learning: An interdisciplinary approach. Washington: Center for Applied Linguistics.

Goodman, K. S. (1976). Reading: A psycholinguistic guessing game. In Ruddell, R., Ruddell, M., Singer, H. (Eds.), Theoretical models and processes of reading (pp. 497-508). Newark: International Reading Association.

Goodman, K. S. (1982). Process, theory, research. (Vol. 1) London: Routledge and Kegan Paul.

Gorin, J. S. (2007). Reconsidering issues in validity theory. Educational Researcher, 36, 456-462.

Gough, P. (1972). One second of reading. In Kavanagh, J., Mattingly, I. (Eds.), Language by ear and by eye: The relationships between speech and reading (pp. 331-358). Cambridge: MIT Press.

Grabe, W. (1991). Current development in second language reading research. TESOL Quarterly, 25,3, 375-406.

Hatch, E., & Lazaraton, A. (1991). The research manual: Design and statistics for applied linguistics. Boston: Heinle & Heinle.

Haenggi, D., Gernsbacher, M. A. & Bolliger, C. A. (1994). Individual differences in situation-based inferencing during narrative text comprehension. In H. van Oostendorp & R. A. Zwaan (Eds.), Naturalistic text comprehension. (pp. 79-96). Norwood, NJ: Ablex.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48, 1-47.

Hamdan, J., & Diab, T. (1997). Using Arabic in testing reading comprehension in English. Journal of the Educational Research Centre, Qatar University 6,12, 1-19

Hsu, P. & Yang, W. (2007). Print and image integration of science texts and reading comprehension: A systemic functional linguistics perspective. International Journal of Science and Mathematics Education, 5, 639-659.

Huang, S. H. (2005). Assessing the relationship between referential understanding and academic reading comprehension among EFL college students.

Unpublished master's thesis, National Yunlin University of Science and Technology, Taiwan.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. The Journal of General Psychology, 123 (3), 207-215.

Hudson, T. (2007). Teaching second language reading. Oxford: Oxford University Press

Hughes, A. (2003). Testing for language teachers. Cambridge: Cambridge University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine 2(8), e124.

Ionescu, I. (2008). Approaches to reading. background knowledge and textual information. "Valahia" University Press, Târgovişte, 147-152. Retrieved May 18, 2010, from http://fsu.valahia.ro/user/image/annaleslettre2008.doc

Johnson, P. (1981). Effects on reading comprehension of language complexity and cultural background of a text. TESOL Quarterly, 15(2), 169-181.

Jones, C. (1991), Qualitative interviewing, in Allen,. G. and Skinner, C. (Editors), Handbook for Research. Students in the Social Sciences. London: The Falmer Press

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87, 329–354.

Kaivanpanah, S. & Alavi, S. M. (2008). Deriving unknown word meaning from context: Is it reliable? RELC Journal, 39(1), 77-95.

Kennison, S. M. (2003). Comprehending the pronouns her, him, and his: Implications for theories of referential processing. Journal of Memory and Language, 49, 335-352.

Kerrod, R., Madgwick, W., Read, S., Collins, F., & Brooks, P. (2006). Deserts. In 1000 Questions and Answers Factfile. London: Kingfisher.

Kerrod, R., Madgwick, W., Read, S., Collins, F., & Brooks, P. (2006). Water. In 1000 Questions and Answers Factfile. London: Kingfisher.

Khaldieh, S. A. (2001). The relationship between knowledge of icraab, lexical knowledge, and reading comprehension of nonnative readers of Arabic. The Modern Language Journal, 85(3), 416-431.

Kitao, S., & Kitao, K. (1999). Essentials of English language testing. Tokyo: Eichosha Co., Ltd.

Kunnan, A. J. (Ed.) (1998). Validation in language assessment. Mahwah, NJ: Lawrence Erlbaum Associates.

Kvale, S. (2007). Doing interviews. London: SAGE Publications Ltd

LaBerge, D. Samuels, S.J. (1974). Toward a theory of automatic information processing in reading. Cognitive Psychology, 6, 293-323.

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. Language Testing, 21(1), 74-100.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. Educational Researcher, 36, 437-448.

Lonsdale, M. d. S., Dyson, M. C., & Reynolds, L. (2006). Reading in examination-type situations: The effects of text layout on performance. Journal of Research in Reading , 29(4), 433-453.

Lorch, R. F., Jr., Lorch, E. P., & Klusewitz, M. A. (1995). Effects of typographical cues on reading and recall of text. Contemporary Educational Psychology, 20, 51-64.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. Language Testing 10(3), 211-234.

Martinez, E., & Godev, C. (1994). Should reading comprehension be tested in the target or the native language? A pilot study. (ERIC Document Reproduction Services No. ED 390 288)

Matthews, A. and Chodorow, M. S. (1988). Pronoun resolution in two-clause sentences: Effects of ambiguity, antecedent location, and depth of embedding. Journal of Memory and Language, 27, 245-260.

McDonald, M. (2002). Systematic assessment of learning outcomes: Developing multiple-choice exams. Boston: Jones and Bartlett Publishers.

McNamara, D. S., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 14(1), 1-43.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18 (2), 5-11.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5-8.

Messick, S. (1998). Test validity: A matter of consequence. Social Indicators Research, 45, 35-44.

Mitchell, D.C., (1982). The process of reading: A cognitive analysis of fluent reading and learning to read. New York: John Wiley & Sons.

Newspapers. [CD-ROM]. (2004). Encarta  Encyclopedia Plus (UK & Ireland Edition). Redmond, WA: Microsoft Corporation.

Norvig, P. (1989). Marker passing as a weak method for text inferencing. Cognitive Science, 13, 569-620.

Nuttall, C. (2005). Teaching Reading Skills in a Foreign Language (2nd ed.). Oxford: Macmillan.

Oakhill, J. (1993). Children's difficulties in reading comprehension. Educational Psychology Review, 5, 223-237.

Oh, S. (2001). Two types of input modification and EFL reading comprehension: simplification versus elaboration. TESOL Quarterly, 35,1, 69-96

Patton, M. Q. (1990). Qualitative Evaluation and Research Methods (2nd ed.). Newbury Park, CA: Sage Publications, Inc.

Pearson, P. D. & Stephens, D. (1994). Learning about literacy: A 30-year journey. In Ruddell, R., Ruddell, M., Singer, H. (Eds.), Theoretical models and processes of reading (pp. 22-42). Newark: International Reading Association.

Pearson, P. D. & Tierney, R. (1984). On becoming a thoughtful reader: Learning to read like a writer. In A.C. Purves, & O. Niles (Eds.), Becoming readers in a complex society (p.144-173). Chicago: University of Chicago Press.

Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Theory and research into practice: vocabulary assessment: What we know and what we need to learn. Reading Research Quarterly, 42(2), 282-296.

Perfetti, C. A. (1985). Reading ability. New York: Oxford University Press.

Perfetti, C. A. (1991). Representations and awareness in the acquisition of reading competence. In L. Rieben & C. A. Perfetti (Eds.), Learning to read: Basic research and its implications (pp. 33-44). Hillsdale, NJ: Lawrence Erlbaum Associates.

Phakiti, A. (2003). A closer look at gender and strategy use in L2 reading. Language Learning, 53(4), 649-702.

Powers, D. E., (1995). Performance by gender on an unconventional verbal reasoning task: Answering reading comprehension questions without the passages. (College Board Report No. 95-2, ETS RR No. 95-14). College Entrance Examination Board, New York.

Pretorius, E. J. (2005). English as a second language learner differences in anaphoric resolution: Reading to learn in the academic context. Applied Psycholinguistics, 26, 521-539.

Pritchard, R. (1990). The effect of cultural schemata on reading processing strategies. Reading Research Quarterly, 25(4), 273-295.

Richards, J.C., Platt, J., & Platt, H. (1992). Longman dictionary of language teaching and applied linguistics. Singapore: Longman.

Road Safety. [CD-ROM]. (2004). Encarta  Encyclopedia Plus (UK & Ireland Edition). Redmond, WA: Microsoft Corporation.

Rose, R. (2010). What's in a pronoun. Manuscript in preparation. Retrieved May 26, 2010, from http://www.roselab.sci.waseda.ac.jp/resources/file/rose_whats _in_a_pronoun.pdf

Rosowsky, A. (2000) Reading and culture: the Experience of some of our Bilingual Pupils. English in Education, 34(2), 45-53.

Rumelhart, D. E. (1977). Understanding and summarizing brief stories. In D. Laberge & S. J. Samuels (Eds.), Basic process in reading: Perception and comprehension (pp. 265-303). Hillsdale, N. J.: Lawrence Erlbaum Associates.

Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. Language Testing, 23 (4), 441-474.

Samuels, S.J., & Kamil, M. (1984). Models of the reading process. In Pearson, P., Barr, R., Kamil, M., and Mosenthal, P. (Eds.), Handbook of reading research (pp. 185-224). New York: Longman.

Samuels, S.J., & Kamil, M. (1988). Models of the reading process. In Carrell, P., Devine, J., & Eskey, D. (Eds.), Interactive approaches to second language reading (pp.22-36). New York: Cambridge University Press.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. Language Testing, 1,2, 147-170

Silverman, D. (1993). Interpreting qualitative data: Methods for analyzing talk, text, and interaction. London: SAGE Publications Ltd.

Sireci, S. G. (1998). Gathering and analyzing content validity data. Educational Assessment, 5, 299–321.

Smith, F. (1971). Understanding reading. New York: Holt, Rinehart & Winston.

Smith, F. (1985). Reading. Cambridge: Cambridge University Press.

Song, M. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. Language Testing, 25(4), 435-464.

Stobart, G. (2008). The validity of ability tests – A case of over-interpretation? Institute of Education, University of London.

Tierney, R., & Cunningham, J. (1984). Research on teaching reading comprehension. In Pearson, P., Barr, R., Kamil, M., and Mosenthal, P. (Eds.), Handbook of reading research (pp. 609-655). New York: Longman.

Tighezza, A. B. (2008, April). Modern validity and its developmental implications on assessment [نظرية الصدق الحديثة ومتضمناتها التطويرية لواقع القياس]. In F. A. Al-Saud (Chair), Psychology and issues of individual and community development. Symposium conducted at the College of Education, King Saud University, Riyadh, Saudi Arabia.

Urquhart, S. & C. Weir (1998). Reading in a Second Language: Process, Product and Practice. London: Addison Wesley Longman Ltd.

Wolf, F., Gibson, E. & Desmet, T. (2004). Coherence and pronoun resolution. Language and Cognitive Processes, 19(6), 665-675.

Yano, Y., Long, M. H., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language reading comprehension. Language Learning, 44(2), 189-219.

Zhang, L. J., & Anual, S. B. (2008). The role of vocabulary in reading comprehension: The case of secondary school students learning English in Singapore. RELC Journal, 39(1), 51-76.

Zwaan, R. A. & Brown, C. M. (1996). The influence of language proficiency and comprehension skill on situation-model construction. Discourse Process 21, 289-327.

# APPENDIX 1

## University Students: Main Study Test

# A)

Before the development of movable metal type in the mid-15th century and for some time thereafter, news was disseminated by word of mouth, by written letters, or by public notices. Not until 1609 were the earliest known newspapers published. The word *news* was not coined until a century later. By 1645 Stockholm had a court paper, which is still published. Early newspapers were small in size, usually consisting of only one page. They had neither headlines nor advertising and looked more like newsletters than today's broadsheet papers with their bold headlines and numerous pictures. The first daily newspaper in the United States, the *Pennsylvania Evening Post* and *Daily Advertiser*, had begun daily publication in 1783 in Philadelphia. By 1800, 20 daily papers were in operation, and the number continued to increase in the first three decades of the 19th century as the Industrial Revolution spread, spawning a new working class in the nation's growing cities.

In the United Kingdom there are 11 national and about 90 regional daily newspapers with a total daily circulation of almost 20 million copies. The paper with the largest circulation is the Sunday tabloid *News of the World*, which sells more than five million copies a week. Of the dailies, *The Sun* is the sales leader, selling around 4 million copies. In the United States, about 1,700 daily newspapers print a total of 63 million copies, and almost every copy is read by at least two people. Some 6,800 weekly newspapers are also published, with a combined circulation of approximately 40 million. In the United States, the *Wall Street Journal* has the largest circulation, with about 1.9 million copies sold daily.

Newspaper publishers are now experimenting with the use of computers and television to transmit news, advertising, and other information directly into homes. Some people believe that the newspaper of the future will not be printed but will be an electronic information service instantly available in every home. Many publishers already include an on-line version of their newspaper on the Internet, which can be accessed by anybody with a personal computer and a modem.

## Questions:

1) Where was the first American daily newspaper published?

[*Philadelphia*]

2) The pronoun “They” in line (6) refers to …………….

[*early newspapers*]

3) The word “spawning” in line (11) is closest in meaning to …………….

[*creating / making / starting*]

4) Write a title for this passage.

[*Newspapers / News / History of Newspapers / Newspapers Development*]

5) What is the total number of daily and weekly newspapers in America?

[*8500*]

6) How did people know news before 1609?

[*word of mouth, written letters, public notices*]

7) What does the third paragraph mainly discuss?

[*new trends*]

8) How many copies are circulated of American and British daily newspapers?

[*83 millions*]

9) The pronoun "their" in line (7) refers to …………….

[*today's newspapers / broadsheet papers*]

10) The word "disseminated" in line (2) is closest in meaning to …………….

[*spread / known / delivered*]

→

# B)

Throughout the world, at least half a million people are killed and about 15 million injured on the roads each year. Casualty rates vary widely. They depend on population and traffic density and the extent to which preventive and remedial measures have been applied. Typically, more deaths occur on rural roads, where speeds are higher than in urban areas, but serious injuries involving a stay in hospital are at least twice as numerous on the urban roads, where traffic faces more conflicts, especially at junctions.

Research studies have shown that human factors contribute to 95 per cent of accidents, road factors to about one quarter, and vehicle factors to fewer than five per cent. The main human errors are: going too fast for the conditions; failing to give way at junctions; following too closely; overtaking improperly; and misperceiving or misjudging the road situation ahead. Road deficiencies that are main contributory factors are: poor design of layout and control at junctions; inadequate signing, road marking, and lighting; slippery roads; and obstructions on the road, such as parked vehicles. The main vehicle factors are defects in tyres, brakes, and lights, arising from poor maintenance of the vehicle.

Extensive remedial measures aimed at improving road safety have been developed in the fields of engineering, education, and enforcement—the “three Es”. They address both primary safety (reduction of accidents) and secondary safety (alleviation of injury). Research has played a major part in testing and evaluating measures to ensure that resources are most effectively used in practice. The most dramatic effects have followed the use of seat belts and child restraints in cars. Use of seat belts reduces the risk of death or serious injury by about 45 per cent. Publicity has played a major part in increasing wearing rates, but for full effect it needs to be backed by legislation. Legislation for compulsory wearing was first introduced in the State of Victoria, Australia in 1971. Other protective measures that are gaining support are the wearing of helmets by cyclists (again led by the example of Victoria), and the use of crash protection barriers on the central reserve of high-speed motorways.

## Questions:

11) What does the second paragraph mainly discuss?
a. Reasons of Accidents
b. New Safety Studies
c. Human Errors in Driving
d. Road Construction Problems

12) How many people are killed because of car accidents during the last two years?
a. .5 million
b. 1 million
c. 1.5 million
d. 15 million

13) The pronoun “They” in line (2) refers to:
a. population
b. people
c. roads
d. rates

14) The word “remedial” in line (3) is closest in meaning to:
a. radical
b. new
c. corrective
d. accurate

15) What is “secondary safety”?
a. reduction of accidents
b. alleviation of injury
c. the “three Es”
d. none of the above

16) Which of the following would be the best title for this passage?
a. Traffic Laws
b. Road Design
c. Road Safety
d. Careless Drivers

17) Why there are fewer death rates on urban areas?
a. drivers do not drive fast
b. drivers are more cautious
c. drivers know their way
d. drivers are much older

18) Australia was the first country in:
a. building safe roads
b. protecting pedestrians
c. legislating seat belt laws
d. reducing highway speed limit

19) The pronoun “it” in line (24) refers to:
a. publicity
b. legislation
c. death
d. injury

20) The word “barriers” in line (28) could be replaced by:
a. sides
b. rules
c. signs
d. fences

# Answer Sheet

## B

| | | | | |
|---|---|---|---|---|
| **11** | a | b | c | d |
| **12** | a | b | c | d |
| **13** | a | b | c | d |
| **14** | a | b | c | d |
| **15** | a | b | c | d |
| **16** | a | b | c | d |
| **17** | a | b | c | d |
| **18** | a | b | c | d |
| **19** | a | b | c | d |
| **20** | a | b | c | d |

A)

Before the development of movable metal type in the mid-15th century and for some time thereafter, news was disseminated by word of mouth, by written letters, or by public notices. Not until 1609 were the earliest known newspapers published. The word *news* was not coined until a century later. By 1645 Stockholm had a court paper, which is still published. Early newspapers were small in size, usually consisting of only one page. They had neither headlines nor advertising and looked more like newsletters than today's broadsheet papers with their bold headlines and numerous pictures. The first daily newspaper in the United States, the *Pennsylvania Evening Post* and *Daily Advertiser*, had begun daily publication in 1783 in Philadelphia. By 1800, 20 daily papers were in operation, and the number continued to increase in the first three decades of the 19th century as the Industrial Revolution spread, spawning a new working class in the nation's growing cities.

In the United Kingdom there are 11 national and about 90 regional daily newspapers with a total daily circulation of almost 20 million copies. The paper with the largest circulation is the Sunday tabloid *News of the World*, which sells more than five million copies a week. Of the dailies, *The Sun* is the sales leader, selling around 4 million copies. In the United States, about 1,700 daily newspapers print a total of 63 million copies, and almost every copy is read by at least two people. Some 6,800 weekly newspapers are also published, with a combined circulation of approximately 40 million. In the United States, the *Wall Street Journal* has the largest circulation, with about 1.9 million copies sold daily.

Newspaper publishers are now experimenting with the use of computers and television to transmit news, advertising, and other information directly into homes. Some people believe that the newspaper of the future will not be printed but will be an electronic information service instantly available in every home. Many publishers already include an on-line version of their newspaper on the Internet, which can be accessed by anybody with a personal computer and a modem.

**الأسئلة:**

**1)** أين تمت طباعة أول جريدة أمريكية يومية؟

______________________________

**2)** إلى ماذا يعود الضمير "They" في السطر 6 ؟

______________________________

**3)** اكتب كلمة قريبة في معناها من كلمة "spawning" في السطر 11

______________________________

4) أكتب عنوانا مختصراً  لهذه القطعة

5) كم العدد الإجمالي للجرائد اليومية والأسبوعية في أمريكا؟

6) كيف كان الناس يعرفون الأخبار قبل عام 1609؟

7) عن ماذا يتحدث المقطع الثالث من هذه القطعة؟

8) كم عدد النسخ الموزعة للجرائد الأمريكية والبريطانية اليومية؟

9) إلى ماذا يعود الضمير "their" في السطر 7 ؟

10) اكتب كلمة قريبة في معناها من كلمة "disseminated" في السطر  2

→

B)

Throughout the world, at least half a million people are killed and about 15 million injured on the roads each year. Casualty rates vary widely. They depend on population and traffic density and the extent to which preventive and remedial measures have been applied. Typically, more deaths occur on rural roads, where speeds are higher than in urban areas, but serious injuries involving a stay in hospital are at least twice as numerous on the urban roads, where traffic faces more conflicts, especially at junctions.

Research studies have shown that human factors contribute to 95 per cent of accidents, road factors to about one quarter, and vehicle factors to fewer than five per cent. The main human errors are: going too fast for the conditions; failing to give way at junctions; following too closely; overtaking improperly; and misperceiving or misjudging the road situation ahead. Road deficiencies that are main contributory factors are: poor design of layout and control at junctions; inadequate signing, road marking, and lighting; slippery roads; and obstructions on the road, such as parked vehicles. The main vehicle factors are defects in tyres, brakes, and lights, arising from poor maintenance of the vehicle.

Extensive remedial measures aimed at improving road safety have been developed in the fields of engineering, education, and enforcement—the “three Es”. They address both primary safety (reduction of accidents) and secondary safety (alleviation of injury). Research has played a major part in testing and evaluating measures to ensure that resources are most effectively used in practice. The most dramatic effects have followed the use of seat belts and child restraints in cars. Use of seat belts reduces the risk of death or serious injury by about 45 per cent. Publicity has played a major part in increasing wearing rates, but for full effect it needs to be backed by legislation. Legislation for compulsory wearing was first introduced in the State of Victoria, Australia in 1971. Other protective measures that are gaining support are the wearing of helmets by cyclists (again led by the example of Victoria), and the use of crash protection barriers on the central reserve of high-speed motorways.

**الأسئلة:**

**11- عن ماذا يتحدث المقطع الثاني من هذه القطعة؟**

a) أسباب الحوادث
b) دراسات جديدة في السلامة
c) الأخطاء البشرية في القيادة
d) مشاكل بناء الطرق

**12- كم عدد الذين ماتوا في حوادث السيارات خلال العامين الأخيرين؟**

a) نصف مليون
b) مليون
c) 1.5 مليون
d) 15 مليون

**13- يعود الضمير "They" في السطر 2 إلى:**

a) السكان
b) الناس
c) الطرق
d)المعدلات

**14- كلمة "remedial" في السطر3 قريبة في معناها من كلمة:**

radical (a
new (b
corrective (c
accurate (d

**15- ما هي "secondary safety" ؟**

a) تقليل الحوادث
b) التخفيف من الإصابات
The "three Es" (c
d) الخيارات أعلاه غير صحيحة

**16- ما هو العنوان الأفضل لهذه القطعة من العناوين التالية؟**

a) قوانين المرور
b) تصميم الطريق
c) سلامة الطريق
d) السائقين المتهورين

**17- لماذا معدل الوفيات يقل في المدن؟**

a) لأن السائقين لا يسرعون
b) لأن السائقين حذرون أكثر
c) لأن السائقين معتادون على طرقهم
d) لأن السائقين أكبر سناً

**18- استراليا هي الدولة الأولى في:**

a) تشييد الطرق الآمنة
b) حماية المشاة في الطريق
c) سن قانون ضرورة لبس حزام الأمان
d) تخفيض السرعة في الطرق السريعة

**19- يعود الضمير "it" في السطر 24 إلى:**

a) الدعاية
b) القوانين
c) الموت
d) الإصابة

**20- كلمة "barriers" في السطر 28 قريبة في معناها من كلمة:**

sides (a
rules (b
signs (c
fences (d

# ورقة إجابة القطعة الثانية

B

| | | | | |
|---|---|---|---|---|
| **11** | a | b | c | d |
| **12** | a | b | c | d |
| **13** | a | b | c | d |
| **14** | a | b | c | d |
| **15** | a | b | c | d |
| **16** | a | b | c | d |
| **17** | a | b | c | d |
| **18** | a | b | c | d |
| **19** | a | b | c | d |
| **20** | a | b | c | d |

# APPENDIX 2

## University Students: Pilot Study Test

## A)

Before the development of movable metal type in the mid-15th century and for some time thereafter, news was disseminated by word of mouth, by written letters, or by public notices. Not until 1609 were the earliest known newspapers published. The word *news* was not coined until a century later. By 1645 Stockholm had a court paper, which is still published. Early newspapers were small in size, usually consisting of only one page. They had neither headlines nor advertising and looked more like newsletters than today's broadsheet papers with their bold headlines and numerous pictures. The first daily newspaper in the United States, the *Pennsylvania Evening Post* and *Daily Advertiser*, had begun daily publication in 1783 in Philadelphia. By 1800, 20 daily papers were in operation, and the number continued to increase in the first three decades of the 19th century as the Industrial Revolution spread, spawning a new working class in the nation's growing cities.

In the United Kingdom there are 11 national and about 90 regional daily newspapers with a total daily circulation of almost 20 million copies. The paper with the largest circulation is the Sunday tabloid *News of the World*, which sells more than five million copies a week. Of the dailies, *The Sun* is the sales leader, selling around 4 million copies. In the United States, about 1,700 daily newspapers print a total of 63 million copies, and almost every copy is read by at least two people. Some 6,800 weekly newspapers are also published, with a combined circulation of approximately 40 million. In the United States, the *Wall Street Journal* has the largest circulation, with about 1.9 million copies sold daily.

Newspaper publishers are now experimenting with the use of computers and television to transmit news, advertising, and other information directly into homes. Some people believe that the newspaper of the future will not be printed but will be an electronic information service instantly available in every home. Many publishers already include an on-line version of their newspaper on the Internet, which can be accessed by anybody with a personal computer and a modem.

## Questions:

1) Where was the first American daily newspaper published?

[*Philadelphia*]

2) Write a title for the second paragraph

[*newspaper circulation / newspapers in UK and USA*]

3) When was the first time the word “news” used?

[*around 1709*]

4) The pronoun “They” in line (6) refers to …………….

[*early newspapers*]

5) The word “spawning” in line (11) is closest in meaning to …………….

[*creating / making / starting*]

6) What is the highest circulated British daily newspaper?

[*The Sun*]

7) Write a title for this passage

[*Newspapers / News / History of Newspapers / Newspapers Development*]

8) What is the total number of daily and weekly newspapers in America?

[*8500*]

9) The pronoun “their” in line (26) refers to …………….

[*publishers*]

10) The word “numerous” in line (7) could be replaced by …………….

[*many / several / a lot of*]

11) How did people know news before 1609?

[*word of mouth, written letters, public notices*]

12) What does the third paragraph mainly discuss?

[*new trends*]

13) How many copies are circulated of American and British daily newspapers?

[*83 millions*]

14) The pronoun “their” in line (7) refers to …………….

[*today’s newspapers / broadsheet papers*]

15) The word “disseminated” in line (2) is closest in meaning to …………….

[*spread / known / delivered*]

B)

Throughout the world, at least half a million people are killed and about 15 million injured on the roads each year. Casualty rates vary widely. They depend on population and traffic density and the extent to which preventive and remedial measures have been applied. Typically, more deaths occur on rural roads, where speeds are higher than in urban areas, but serious injuries involving a stay in hospital are at least twice as numerous on the urban roads, where traffic faces more conflicts, especially at junctions.

Research studies have shown that human factors contribute to 95 per cent of accidents, road factors to about one quarter, and vehicle factors to fewer than five per cent. The main human errors are: going too fast for the conditions; failing to give way at junctions; following too closely; overtaking improperly; and misperceiving or misjudging the road situation ahead. Road deficiencies that are main contributory factors are: poor design of layout and control at junctions; inadequate signing, road marking, and lighting; slippery roads; and obstructions on the road, such as parked vehicles. The main vehicle factors are defects in tyres, brakes, and lights, arising from poor maintenance of the vehicle.

Extensive remedial measures aimed at improving road safety have been developed in the fields of engineering, education, and enforcement—the “three Es”. They address both primary safety (reduction of accidents) and secondary safety (alleviation of injury). Research has played a major part in testing and evaluating measures to ensure that resources are most effectively used in practice. The most dramatic effects have followed the use of seat belts and child restraints in cars. Use of seat belts reduces the risk of death or serious injury by about 45 per cent. Publicity has played a major part in increasing wearing rates, but for full effect it needs to be backed by legislation. Legislation for compulsory wearing was first introduced in the State of Victoria, Australia in 1971. Other protective measures that are gaining support are the wearing of helmets by cyclists (again led by the example of Victoria), and the use of crash protection barriers on the central reserve of high-speed motorways.

# Questions:

16) What is the main cause of car accidents?
a. bad weather
b. vehicle factors
c. road deficiencies
d. human errors

17) What does the second paragraph mainly discuss?
a. Reasons of Accidents
b. New Safety Studies
c. Human Errors in Driving
d. Road Construction Problems

18) How many people are killed because of car accidents during the last two years?
a. .5 million
b. 1 million
c. 1.5 million
d. 15 million

19) The pronoun “They” in line (2) refers to:
a. population
b. people
c. roads
d. rates

20) The word “remedial” in line (3) is closest in meaning to:
a. radical
b. new
c. corrective
d. accurate

21) What is “secondary safety”?
a. reduction of accidents
b. alleviation of injury
c. the “three Es”
d. none of the above

22) Which of the following would be the best title for this passage?
a. Car Accidents
b. Seat Belt
c. Road Safety
d. Traffic Laws

23) Why there are fewer death rates on urban areas?
a. drivers do not drive fast
b. drivers are more cautious
c. drivers know their way
d. drivers are much older

24) The pronoun “They” in line (18) refers to:
a. legislations
b. seat belts
c. remedial measures
d. car accidents

25) The word “inadequate” in line (13) is closest in meaning to:
a. extra
<u>b. not enough</u>
c. unclear
d. interactive

26) Australia was the first country in:
a. building safe roads
b. protecting pedestrians
<u>c. legislating seat belt laws</u>
d. reducing highway speed limit

27) Which of the following would be the best title for the third paragraph?
a. Seat Belts
<u>b. Protective Measures</u>
c. Safety Education
d. Car Accidents

28) What is the most successful safety measure?
a. speed limit
<u>b. seat belt</u>
c. car inspection
d. excellent roads

29) The pronoun “it” in line (24) refers to:
<u>a. publicity</u>
b. legislation
c. death
d. injury

30) The word “barriers” in line (28) could be replaced by:
a. sides
b. rules
c. signs
<u>d. fences</u>

# Answer Sheet

## B

| | | | | |
|---|---|---|---|---|
| **16** | a | b | c | d |
| **17** | a | b | c | d |
| **18** | a | b | c | d |
| **19** | a | b | c | d |
| **20** | a | b | c | d |
| **21** | a | b | c | d |
| **22** | a | b | c | d |
| **23** | a | b | c | d |
| **24** | a | b | c | d |
| **25** | a | b | c | d |
| **26** | a | b | c | d |
| **27** | a | b | c | d |
| **28** | a | b | c | d |
| **29** | a | b | c | d |
| **30** | a | b | c | d |

A)

Before the development of movable metal type in the mid-15th century and for some time thereafter, news was disseminated by word of mouth, by written letters, or by public notices. Not until 1609 were the earliest known newspapers published. The word *news* was not coined until a century later. By 1645 Stockholm had a court paper, which is still published. Early newspapers were small in size, usually consisting of only one page. They had neither headlines nor advertising and looked more like newsletters than today's broadsheet papers with their bold headlines and numerous pictures. The first daily newspaper in the United States, the *Pennsylvania Evening Post* and *Daily Advertiser,* had begun daily publication in 1783 in Philadelphia. By 1800, 20 daily papers were in operation, and the number continued to increase in the first three decades of the 19th century as the Industrial Revolution spread, spawning a new working class in the nation's growing cities.

In the United Kingdom there are 11 national and about 90 regional daily newspapers with a total daily circulation of almost 20 million copies. The paper with the largest circulation is the Sunday tabloid *News of the World,* which sells more than five million copies a week. Of the dailies, *The Sun* is the sales leader, selling around 4 million copies. In the United States, about 1,700 daily newspapers print a total of 63 million copies, and almost every copy is read by at least two people. Some 6,800 weekly newspapers are also published, with a combined circulation of approximately 40 million. In the United States, the *Wall Street Journal* has the largest circulation, with about 1.9 million copies sold daily.

Newspaper publishers are now experimenting with the use of computers and television to transmit news, advertising, and other information directly into homes. Some people believe that the newspaper of the future will not be printed but will be an electronic information service instantly available in every home. Many publishers already include an on-line version of their newspaper on the Internet, which can be accessed by anybody with a personal computer and a modem.

**الأسئلة:**

**1) أين تمت طباعة أول جريدة أمريكية يومية؟**

**2) أكتب عنوانا مناسبا للمقطع الثاني من هذه القطعة؟**

**3) متى استعملت كلمة "news" لأول مرة؟**

4) إلى ماذا يعود الضمير "They" في السطر 6 ؟

---

5) اكتب كلمة قريبة في معناها من كلمة "spawning" في السطر 11

---

6) ما هي الجريدة البريطانية اليومية ذات التوزيع الأعلى؟

---

7) أكتب عنوانا مناسبا لهذه القطعة

---

8) كم العدد الإجمالي للجرائد اليومية والأسبوعية في أمريكا؟

---

9) إلى ماذا يعود الضمير "their" في السطر 26 ؟

---

10) اكتب كلمة قريبة في معناها من كلمة "numerous" في السطر 7

---

11) كيف كان الناس يعرفون الأخبار قبل عام 1609؟

---

12) عن ماذا يتحدث المقطع الثالث من هذه القطعة؟

---

13) كم عدد النسخ الموزعة للجرائد الأمريكية والبريطانية اليومية؟

---

**14) إلى ماذا يعود الضمير "their" في السطر 7 ؟**

______________________________________________

**15) اكتب كلمة قريبة في معناها من كلمة "disseminated" في السطر 2**

______________________________________________

B)

Throughout the world, at least half a million people are killed and about 15 million injured on the roads each year. Casualty rates vary widely. They depend on population and traffic density and the extent to which preventive and remedial measures have been applied. Typically, more deaths occur on rural roads, where speeds are higher than in urban areas, but serious injuries involving a stay in hospital are at least twice as numerous on the urban roads, where traffic faces more conflicts, especially at junctions.

Research studies have shown that human factors contribute to 95 per cent of accidents, road factors to about one quarter, and vehicle factors to fewer than five per cent. The main human errors are: going too fast for the conditions; failing to give way at junctions; following too closely; overtaking improperly; and misperceiving or misjudging the road situation ahead. Road deficiencies that are main contributory factors are: poor design of layout and control at junctions; inadequate signing, road marking, and lighting; slippery roads; and obstructions on the road, such as parked vehicles. The main vehicle factors are defects in tyres, brakes, and lights, arising from poor maintenance of the vehicle.

Extensive remedial measures aimed at improving road safety have been developed in the fields of engineering, education, and enforcement—the “three Es”. They address both primary safety (reduction of accidents) and secondary safety (alleviation of injury). Research has played a major part in testing and evaluating measures to ensure that resources are most effectively used in practice. The most dramatic effects have followed the use of seat belts and child restraints in cars. Use of seat belts reduces the risk of death or serious injury by about 45 per cent. Publicity has played a major part in increasing wearing rates, but for full effect it needs to be backed by legislation. Legislation for compulsory wearing was first introduced in the State of Victoria, Australia in 1971. Other protective measures that are gaining support are the wearing of helmets by cyclists (again led by the example of Victoria), and the use of crash protection barriers on the central reserve of high-speed motorways.

## الأسئلة:

**16- ما هو السبب الرئيسي لحوادث السيارات؟**

a) الطقس السيئ
b) عوامل مرتبطة بالسيارة
c) عيوب في الطرق
d) أخطاء بشرية

**17- عن ماذا يتحدث المقطع الثاني من هذه القطعة؟**

a) أسباب الحوادث
b) دراسات جديدة في السلامة
c) الأخطاء البشرية في القيادة
d) مشاكل بناء الطرق

**18- كم عدد الذين ماتوا في حوادث السيارات خلال العامين الأخيرين؟**

a) نصف مليون
b) مليون
c) 1.5 مليون
d) 15 مليون

**19- يعود الضمير "They" في السطر 2 إلى:**

a) السكان
b) الناس
c) الطرق
d)المعدلات

**20- كلمة "remedial" في السطر3 قريبة في معناها من كلمة:**

radical (a
new (b
corrective (c
accurate (d

**21- ما هي "secondary safety" ؟**

a) تقليل الحوادث
b) التخفيف من الإصابات
The "three Es" (c
d) الخيارات أعلاه غير صحيحة

**22- ماهو العنوان الأفضل لهذه القطعة من العناوين التالية؟**

a) حوادث السيارات
b) حزام الأمان
c) سلامة الطريق
d) قوانين المرور

**23- لماذا معدل الوفيات يقل في المدن؟**

a) لأن السائقين لا يسرعون
b) لأن السائقين حذرون أكثر
c) لأن السائقين معتادون على طرقهم
d) لأن السائقين أكبر سناً

**24- يعود الضمير "They" في السطر 18 إلى:**

a) القوانين
b) حزام الأمان
c) أنظمة الحماية
d) حوادث السيارات

**25- كلمة "inadequate " في السطر13 قريبة في معناها من كلمة:**
extra (a
not enough (b
unclear (c
interactive (d

**26- استراليا هي الدولة الأولى في:**
a) تشييد الطرق الآمنة
b) حماية المشاة في الطريق
c) سن قانون ضرورة لبس حزام الأمان
d) تخفيض السرعة في الطرق السريعة

**27- ما هو العنوان الأفضل للمقطع الثالث من هذه القطعة؟**
a) حزام الأمان
b) أنظمة الحماية
c) ثقافة السلامة
d) حوادث السيارات

**28- ما هي أكثر طرق الحماية نجاحاً؟**
a) تحديد السرعة
b) حزام الأمان
c) فحص السيارة
d) الطرق المملؤة

**29- يعود الضمير "it" في السطر 24 إلى:**
a) الدعاية
b) القوانين
c) الموت
d) الإصابة

**30- كلمة "barriers" في السطر 28 قريبة في معناها من كلمة:**
sides (a
rules (b
signs (c
fences (d

# ورقة إجابة القطعة الثانية

B

| | | | | |
|---|---|---|---|---|
| **16** | a | b | c | d |
| **17** | a | b | c | d |
| **18** | a | b | c | d |
| **19** | a | b | c | d |
| **20** | a | b | c | d |
| **21** | a | b | c | d |
| **22** | a | b | c | d |
| **23** | a | b | c | d |
| **24** | a | b | c | d |
| **25** | a | b | c | d |
| **26** | a | b | c | d |
| **27** | a | b | c | d |
| **28** | a | b | c | d |
| **29** | a | b | c | d |
| **30** | a | b | c | d |

# APPENDIX 3

## Secondary School Students: Main Study Test

# A)

From space, Earth looks blue because 71 % of its surface is covered by water. About 97 % of the planet's water is in the sea and is salty. The remaining water is in the rivers, lakes and glaciers.

Sea water contains common salt and other minerals. On average, the sea is 3.5 % salt. However, the Dead Sea is 25 % salt. High salt content gives water great buoyancy, and so it is very easy for swimmers to float in the Dead Sea.

Oceans are huge areas of sea water. There are four oceans - the Pacific, the Atlantic, the Indian and the Arctic. The Pacific is the largest and the deepest. It is more than twice as big as the second largest ocean, the Atlantic. The Pacific is wide enough to fit all the continents, and deep enough to swallow Mount Everest, the world highest mountain.

The longest river in the world is the Nile in Africa. It runes 6695 kilometres from its source in Lake Victoria, Burundi, to the Mediterranean Sea. Rivers are a vital resource for humans, providing food and water for drinking and irrigation. The Nile is such a long river that it is even visible from space!

## Questions:

1) What is "land" percentage on Earth?

   [*29%*]

2) The pronoun "its" in line (12) refers to …………….

   [*the Nile / the river*]

3) What does the word "*visible*" in line (15) mean ?

   [*noticeable / clear / can be seen*]

4) Where does the longest river start?

   [*Lake Victoria*]

5) Write a title for this passage.

   [*Water*]

6) Why is it possible to see the Nile from space?

   [*long*]

7) What does the third paragraph mainly discuss?

[*oceans*]

8) What is the size of the Atlantic compared to the Pacific?

[*less than 50% of the Pacific / half / 50%*]

9) The pronoun “it” in line (8) refers to …………….

[*the Pacific*]

10) What does the word “*vital*” in line (13) mean ?

[*important / useful / essential*]

## B)

Deserts are regions where the annual rainfall is less than 250 millimetres, but some deserts have no rain for several years. They are the driest places on Earth and are often very windy. Few plants can survive in these conditions.

The three largest deserts in the world are the Sahara, the Arabian and the Gobi. The Sahara covers about one third of Africa. It is about 8.6 million square kilometres. The Arabian Desert and Gobi in Asia measure 2.3 and 1.2 million square kilometres respectively. Australia has several deserts that cover a huge area of land.

Daytime temperature can rise to a scoring 50˚C in some deserts. Once the sun sets, however, temperatures can drop dramatically because there are few clouds over deserts to keep the day’s heat in. The difference between day and night temperatures in the Western Sahara can be more than 45˚C, where as the Gobi has a more temperate climate.

Many desert animals, such as foxes, rabbits and coyotes, are nocturnal. They sleep during the scorching hot day and come out to find food at night when it is cooler. Some desert animals are able to survive with little or no water – camels can go for many days without drinking.

## Questions:

11) Which of the following would be the best title for the third paragraph?
a. Gobi Desert
b. Western Sahara
c. night temperature
d. desert temperatures

12) The size of Gobi Desert is:

a. 8.6 million km$^2$
b. 1.2 million km$^2$
c. 2.3 million km$^2$
d. 3.5 million km$^2$

13) The size of Africa is about:

a. 25.8 million km$^2$
b. 8.6 million km$^2$
c. 2.3 million km$^2$
d. 12.1 million km$^2$

14) The pronoun “It” in line (5) refers to:

a. Africa
b. Sahara Desert
c. Arabian Desert
d. Gobi Desert

15) The word “huge” in line (7) is closest in meaning to:

a. large
b. dry
c. little
d. hot

16) Sahara Desert is located in:

a. Asia
b. Africa
c. Australia
d. none of the above

17) Which of the following would be the best title for this passage?

a. Desert
b. Temperatures
c. Desert Animals
d. Earth

18) The second largest desert in the world is:

a. Sahara Desert
b. Gobi Desert
c. Arabian Desert
d. Australian Desert

19) The pronoun “They” in line (2) refers to:

a. years
b. rain
c. places
d. deserts

20) The word “survive” in line (15) could be replaced by:

a. sleep
b. eat
c. stay alive
d. move around

# Answer Sheet

## B

| | | | | |
|---|---|---|---|---|
| **11** | a | b | c | d |
| **12** | a | b | c | d |
| **13** | a | b | c | d |
| **14** | a | b | c | d |
| **15** | a | b | c | d |
| **16** | a | b | c | d |
| **17** | a | b | c | d |
| **18** | a | b | c | d |
| **19** | a | b | c | d |
| **20** | a | b | c | d |

A)

From space, Earth looks blue because 71 % of its surface is covered by water. About 97 % of the planet’s water is in the sea and is salty. The remaining water is in the rivers, lakes and glaciers.

Sea water contains common salt and other minerals. On average, the sea is 3.5 % salt. However, the Dead Sea is 25 % salt. High salt content gives water great buoyancy, and so it is very easy for swimmers to float in the Dead Sea.

Oceans are huge areas of sea water. There are four oceans - the Pacific, the Atlantic, the Indian and the Arctic. The Pacific is the largest and the deepest. It is more than twice as big as the second largest ocean, the Atlantic. The Pacific is wide enough to fit all the continents, and deep enough to swallow Mount Everest, the world highest mountain.

The longest river in the world is the Nile in Africa. It runes 6695 kilometres from its source in Lake Victoria, Burundi, to the Mediterranean Sea. Rivers are a vital resource for humans, providing food and water for drinking and irrigation. The Nile is such a long river that it is even visible from space!

## الأسئلة :

1) ما هي نسبة اليابسة في الكرة الأرضية ؟

2) إلى ماذا يعود الضمير "its" في السطر 12 ؟

3) ماذا تعني كلمة "*visible*" في السطر 15 ؟

4) من أين يبدأ أطول نـهر في العالم ؟

5) أكتب عنوانا معبراً لمحتوى هذه القطعة

6) لماذا من الممكن رؤية نـهر النيل من الفضاء الخارجي ؟

**7)** **عن ماذا يتحدث المقطع الثالث من هذه القطعة ؟**

______________________________________________

**8)** **كم تبلغ مساحة المحيط الأطلسي مقارنة بالمحيط الهادي ؟**

______________________________________________

**9)** **إلى ماذا يعود الضمير** "it" **في السطر** 8 ؟

______________________________________________

**10)** **ماذا تعني كلمة** "*vital*" **في السطر** 13 ؟

______________________________________________

# B)

Deserts are regions where the annual rainfall is less than 250 millimetres, but some deserts have no rain for several years. They are the driest places on Earth and are often very windy. Few plants can survive in these conditions.

The three largest deserts in the world are the Sahara, the Arabian and the Gobi. The Sahara covers about one third of Africa. It is about 8.6 million square kilometres. The Arabian Desert and Gobi in Asia measure 2.3 and 1.2 million square kilometres respectively. Australia has several deserts that cover a huge area of land.

Daytime temperature can rise to a scoring 50˚C in some deserts. Once the sun sets, however, temperatures can drop dramatically because there are few clouds over deserts to keep the day's heat in. The difference between day and night temperatures in the Western Sahara can be more than 45˚C, where as the Gobi has a more temperate climate.

Many desert animals, such as foxes, rabbits and coyotes, are nocturnal. They sleep during the scorching hot day and come out to find food at night when it is cooler. Some desert animals are able to survive with little or no water – camels can go for many days without drinking.

## الأسئلة :

**11- ماهو العنوان الأفضل للمقطع الثالث من هذه القطعة؟**

a) صحراء "قوبي"
b) "صحارى الغربية"
c) درجة حرارة المساء
d) درجة حرارة الصحراء

**12- مساحة صحراء "قوبي" هي:**

a) 8.6 مليون كم$^2$<br>
b) 1.2 مليون كم$^2$<br>
c) 2.3 مليون كم$^2$<br>
d) 3.5 مليون كم$^2$

**13- تبلغ مساحة أفريقيا حوالي:**

a) 25.8 مليون كم$^2$<br>
b) 8.6 مليون كم$^2$<br>
c) 2.3 مليون كم$^2$<br>
d) 12.1 مليون كم$^2$

**14- يعود الضمير "It" في السطر 5 إلى:**

a) قارة أفريقيا<br>
b) صحراء "صحارى"<br>
c) الصحراء العربية<br>
d) صحراء قوبي

**15- كلمة "huge" في السطر 7 قريبة في معناها من كلمة:**

a) large<br>
b) dry<br>
c) little<br>
d) hot

**16- تقع صحراء "صحارى" في:**

a) آسيا<br>
b) أفريقيا<br>
c) استراليا<br>
d) الخيارات أعلاه غير صحيحة

**17- ماهو العنوان الأفضل لهذه القطعة من العناوين التالية؟**

a) الصحراء<br>
b) درجات الحرارة<br>
c) حيوانات الصحراء<br>
d) الأرض

**18- ثاني أكبر صحراء في العالم هي:**

a) صحراء "صحارى"<br>
b) صحراء "قوبي"<br>
c) الصحراء العربية<br>
d) الصحراء الاسترالية

**19- يعود الضمير "They" في السطر 2 إلى:**

a) السنوات<br>
b) الأمطار<br>
c) الأماكن<br>
d) الصحاري

**20- يمكن استبدال كلمة "survive" في السطر 15 بكلمة:**

a) sleep<br>
b) eat<br>
c) stay alive<br>
d) move around

# ورقة إجابة القطعة الثانية

B

| | | | | |
|---|---|---|---|---|
| **11** | a | b | c | d |
| **12** | a | b | c | d |
| **13** | a | b | c | d |
| **14** | a | b | c | d |
| **15** | a | b | c | d |
| **16** | a | b | c | d |
| **17** | a | b | c | d |
| **18** | a | b | c | d |
| **19** | a | b | c | d |
| **20** | a | b | c | d |

# APPENDIX 4

## Secondary School Students: Pilot Study Test

## A)

From space, Earth looks blue because 71 % of its surface is covered by water. About 97 % of the planet's water is in the sea and is salty. The remaining water is in the rivers, lakes and glaciers.

Sea water contains common salt and other minerals. On average, the sea is 3.5 % salt. However, the Dead Sea is 25 % salt. High salt content gives water great buoyancy, and so it is very easy for swimmers to float in the Dead Sea.

Oceans are huge areas of sea water. There are four oceans - the Pacific, the Atlantic, the Indian and the Arctic. The Pacific is the largest and the deepest. It is more than twice as big as the second largest ocean, the Atlantic. The Pacific is wide enough to fit all the continents, and deep enough to swallow Mount Everest, the world highest mountain.

The longest river in the world is the Nile in Africa. It runes 6695 kilometres from its source in Lake Victoria, Burundi, to the Mediterranean Sea. Rivers are a vital resource for humans, providing food and water for drinking and irrigation. The Nile is such a long river that it is even visible from space!

## Questions:

1) Where is Lake Victoria located?

[*Burundi*]

2) Write a title for the forth paragraph

[*The Nile*]

3) What is "land" percentage on Earth?

[*29%*]

4) The pronoun "its" in line (12) refers to …………….

[*the Nile / the river*]

5) Write a synonym for the word "visible" in line (15)

[*noticeable / clear / can be seen*]

6) Where does the longest river start?

[*Lake Victoria*]

7) Write a title for this passage

[*Water*]

8) What is the percentage of “drinkable water” of the planet’s water?

[*3%*]

9) The pronoun “its” in line (1) refers to …………….

[*earth*]

10) Write a synonym for the word “buoyancy” in line (5)

[*remain afloat*]

11) Why is it possible to see the Nile from space?

[*long*]

12) What does the third paragraph mainly discuss?

[*oceans*]

13) What is the size of the Atlantic compared to the Pacific?

[*less than 50% of the Pacific / half / 50%*]

14) The pronoun “it” in line (8) refers to …………….

[*the Pacific*]

15) Write a synonym for the word “vital” in line (13)

[*important / useful / essential*]

# B)

Deserts are regions where the annual rainfall is less than 250 millimetres, but some deserts have no rain for several years. They are the driest places on Earth and are often very windy. Few plants can survive in these conditions.

The three largest deserts in the world are the Sahara, the Arabian and the Gobi. The Sahara covers about one third of Africa. It is about 8.6 million square kilometres. The Arabian Desert and Gobi in Asia measure 2.3 and 1.2 million square kilometres respectively. Australia has several deserts that cover a huge area of land.

Daytime temperature can rise to a scoring 50˚C in some deserts. Once the sun sets, however, temperatures can drop dramatically because there are few clouds over deserts to keep the day’s heat in. The difference between day and night temperatures in the Western Sahara can be more than 45˚C, where as the Gobi has a more temperate climate.

Many desert animals, such as foxes, rabbits and coyotes, are nocturnal. They sleep during the scorching hot day and come out to find food at night when it is cooler. Some desert animals are able to survive with little or no water – camels can go for many days without drinking.

## Questions:

16) Why does desert temperature go down at night?
a. no rain
b. too dark
c. no clouds
d. very windy

17) Which of the following would be the best title for the third paragraph?
a. Gobi Desert
b. Western Sahara
c. night temperature
d. desert temperatures

18) The night temperature in Western Sahara is around:
a. 50 ˚C
b. 5 ˚C
c. 45 ˚C
d. 30 ˚C

19) The pronoun “They” in line (13) refers to:
a. foxes
b. rabbits
c. coyotes
d. all of the above

20) The word “annual” in line (1) is closest in meaning to:
a. amount
b. heavy
c. yearly
d. unusual

21) The size of Gobi Desert is:
a. 8.6 million km$^2$
b. 1.2 million km$^2$
c. 2.3 million km$^2$
d. 3.5 million km$^2$

22) What does the forth paragraph mainly discuss?
a. camels
b. foxes
c. rabbits
d. desert animals

23) The size of Africa is about:
a. 25.8 million km$^2$
b. 8.6 million km$^2$
c. 2.3 million km$^2$
d. 1.2 million km$^2$

24) The pronoun “It” in line (5) refers to:
a. Africa
b. Sahara
c. Arabian
d. Gobi

25) The word “huge” in line (7) is closest in meaning to:
a. large
b. dry
c. little
d. hot

26) Sahara Desert is located in:
a. Asia
b. Africa
c. Australia
d. none of the above

27) Which of the following would be the best title for this passage?
a. Desert
b. Rain
c. Weather
d. Earth

28) The second largest desert in the world is:
a. Sahara Desert
b. Gobi Desert
c. Arabian Desert
d. Australian Desert

29) The pronoun “They” in line (2) refers to:
a. years
b. rainfall
c. plants
d. deserts

30) The word “survive” in line (15) could be replaced by:
a. sleep
b. eat
c. stay alive
d. move around

# Answer Sheet

## B

| | | | | |
|---|---|---|---|---|
| **16** | a | b | c | d |
| **17** | a | b | c | d |
| **18** | a | b | c | d |
| **19** | a | b | c | d |
| **20** | a | b | c | d |
| **21** | a | b | c | d |
| **22** | a | b | c | d |
| **23** | a | b | c | d |
| **24** | a | b | c | d |
| **25** | a | b | c | d |
| **26** | a | b | c | d |
| **27** | a | b | c | d |
| **28** | a | b | c | d |
| **29** | a | b | c | d |
| **30** | a | b | c | d |

## A)

From space, Earth looks blue because 71 % of its surface is covered by water. About 97 % of the planet’s water is in the sea and is salty. The remaining water is in the rivers, lakes and glaciers.

Sea water contains common salt and other minerals. On average, the sea is 3.5 % salt. However, the Dead Sea is 25 % salt. High salt content gives water great buoyancy, and so it is very easy for swimmers to float in the Dead Sea.

Oceans are huge areas of sea water. There are four oceans - the Pacific, the Atlantic, the Indian and the Arctic. The Pacific is the largest and the deepest. It is more than twice as big as the second largest ocean, the Atlantic. The Pacific is wide enough to fit all the continents, and deep enough to swallow Mount Everest, the world highest mountain.

The longest river in the world is the Nile in Africa. It runes 6695 kilometres from its source in Lake Victoria, Burundi, to the Mediterranean Sea. Rivers are a vital resource for humans, providing food and water for drinking and irrigation. The Nile is such a long river that it is even visible from space!

### الأسئلة:

**1) أين تقع بحيرة فيكتوريا؟**

**2) أكتب عنوانا مناسبا للمقطع الرابع من هذه القطعة**

**3) ما هي نسبة اليابسة في الكرة الأرضية؟**

**4) إلى ماذا يعود الضمير "its" في السطر 12 ؟**

**5) اكتب مرادفا لكلمة "visible" في السطر 15**

6) من أين يبدأ أطول نـهر في العالم؟

---

7) أكتب عنوانا مناسبا لـهذه القطعة

---

8) ما هي نسبة الماء الصالح للشرب من إجمالي الماء الموجود على سطح الأرض؟

---

9) إلى ماذا يعود الضمير "its" في السطر 1 ؟

---

10) اكتب كلمة قريبة في معناها من كلمة "buoyancy" في السطر 5

---

11) لماذا من الممكن رؤية نـهر النيل من الفضاء؟

---

12) عن ماذا يتحدث المقطع الثالث من هذه القطعة؟

---

13) كم تبلغ مساحة المحيط الأطلسي مقارنة بالمحيط الـهادي؟

---

14) إلى ماذا يعود الضمير "it" في السطر 8 ؟

______________________________________________

15) اكتب  كلمة قريبة في معناها من كلمة "vital" في السطر  13

______________________________________________

B)

Deserts are regions where the annual rainfall is less than 250 millimetres, but some deserts have no rain for several years. They are the driest places on Earth and are often very windy. Few plants can survive in these conditions.

The three largest deserts in the world are the Sahara, the Arabian and the Gobi. The Sahara covers about one third of Africa. It is about 8.6 million square kilometres. The Arabian Desert and Gobi in Asia measure 2.3 and 1.2 million square kilometres respectively. Australia has several deserts that cover a huge area of land.

Daytime temperature can rise to a scoring 50˚C in some deserts. Once the sun sets, however, temperatures can drop dramatically because there are few clouds over deserts to keep the day’s heat in. The difference between day and night temperatures in the Western Sahara can be more than 45˚C, where as the Gobi has a more temperate climate.

Many desert animals, such as foxes, rabbits and coyotes, are nocturnal. They sleep during the scorching hot day and come out to find food at night when it is cooler. Some desert animals are able to survive with little or no water – camels can go for many days without drinking.

**الأسئلة:**

**16- لماذا تنخفض درجة حرارة الصحراء في المساء؟**

a ) لا يوجد مطر
b) الظلام شديد
c) لا يوجد سحاب
d ) الرياح شديدة

**17- ماهو العنوان الأفضل للمقطع الثالث من هذه القطعة؟**

a ) صحراء "قوبي"
b) "صحارى الغربية"
c) درجة حرارة المساء
d ) درجة حرارة الصحراء

**18- تبلغ درجة حرارة "صحارى الغربية" في المساء حوالي:**

a ) 50° درجة مئوية
b) 5° درجة مئوية
c) 45° درجة مئوية
d ) 30° درجة مئوية

**19- يعود الضمير "They" في السطر 13 إلى:**

a ) الثعالب
b) الأرانب
c) ذئاب صغيرة
d ) جميع ما سبق

**20- كلمة "annual" في السطر 1 قريبة في معناها من كلمة:**

a ) amount
b) heavy
c) yearly
d ) unusual

**21- مساحة صحراء "قوبي" هي:**

a ) 8.6 مليون كم$^2$
b) 1.2 مليون كم$^2$
c) 2.3 مليون كم$^2$
d ) 3.5 مليون كم$^2$

**22- عن ماذا يتحدث المقطع الرابع من هذه القطعة؟**

a ) الجمال
b) الثعالب
c) ذئاب صغيرة
d ) حيوانات الصحراء

**23- تبلغ مساحة أفريقيا حوالي:**

a ) 25.8 مليون كم$^2$
b) 8.6 مليون كم$^2$
c) 2.3 مليون كم$^2$
d ) 1.2 مليون كم$^2$

**24- يعود الضمير "It" في السطر 5 إلى:**

a ) قارة أفريقيا
b) صحراء صحارى
c) الصحراء العربية
d ) صحراء قوبي

**25- كلمة "huge" في السطر7 قريبة في معناها من كلمة:**

a ) large
b) dry
c) little
d ) hot

**26- تقع صحراء "صحارى" في:**

a ) آسيا
b) أفريقيا
c) استراليا
d ) الخيارات أعلاه غير صحيحة

**27- ماهو العنوان الأفضل لهذه القطعة من العناوين التالية؟**

a ) الصحراء
b) المطر
c) الطقس
d ) الأرض

**28- ثاني أكبر صحراء في العالم هي:**

a ) صحراء "صحارى"
b) صحراء "قوبي"
c) الصحراء العربية
d ) الصحراء الاسترالية

**29- يعود الضمير "They" في السطر 2 إلى:**

a ) السنوات
b) سقوط الأمطار
c) النباتات
d ) الصحاري

**30- يمكن استبدال كلمة "survive" في السطر 15 بكلمة:**

sleep ( a
eat (b
stay alive (c
move around ( d

# ورقة إجابة القطعة الثانية

B

| | | | | |
|---|---|---|---|---|
| **16** | a | b | c | d |
| **17** | a | b | c | d |
| **18** | a | b | c | d |
| **19** | a | b | c | d |
| **20** | a | b | c | d |
| **21** | a | b | c | d |
| **22** | a | b | c | d |
| **23** | a | b | c | d |
| **24** | a | b | c | d |
| **25** | a | b | c | d |
| **26** | a | b | c | d |
| **27** | a | b | c | d |
| **28** | a | b | c | d |
| **29** | a | b | c | d |
| **30** | a | b | c | d |