

AUS Repository

Evaluating Reading Comprehension Testing in Kuwait

Item Type	Thesis
Authors	Barman, Meshari A.
Download date	2026-03-11 03:39:59
Link to Item	http://hdl.handle.net/11073/8320

EVALUATING READING COMPREHENSION TESTING IN KUWAIT

by

Meshari A. Barman

A Thesis Presented to the Faculty of the

American University of Sharjah

College of Arts and Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts in Teaching English to Speakers of Other Language

Sharjah, United Arab Emirates

May 2016

Copyright © 2016 Meshari Barman. All rights reserved

We, the undersigned, approve the Master's thesis of Meshari A. Barman

Thesis Title: EVALUATING READING COMPREHENSION TESTING IN KUWAIT

Signature

Date of signature
(dd/mm/yy)

Dr. Betty Lanteigne
Associate Professor
Thesis Advisor

Dr. Khawlah Ahmed
Associate Professor

Dr. Gary Linebaugh
Assistant Professor

Dr. James Griffin
CAS Graduate Programs Director

Dr. Mahmoud Anabtawi
Dean of the College of Arts and Sciences

Dr. Khaled Assaleh
Interim Vice Provost, Graduate Studies and Research

Acknowledgement

First of all, I would like to express my gratitude to my advisor Dr. Betty Lanteigne for the continuous support of my MA study, for her patience, motivation, and immense knowledge. I was fortunate to be taught by Dr. Betty two courses, one of which was the assessment course. In addition, I was fortunate to have her as an advisor for my thesis. Dr. Betty's guidance helped me throughout the research and writing of this thesis.

I would also like to thank the members of my thesis committee: Dr. Khawlah Ahmed, and Dr. Gary Linebaugh for their insightful comments and encouragement.

Last but not least, I would like to thank my wife Altaf Barman, Ms. Huda Al-Ammar the senior supervisor in Al-Jahra Educational Area, and the English language supervision board in Al-Jahra Educational Area for their support and encouragement during my MA TESOL study.

Abstract

In order to improve Kuwait eighth grade intermediate stage reading comprehension tests, it is important to obtain concrete information about the usefulness of these tests. This study is an evaluation of the current eighth grade reading comprehension first instruction period test in Kuwait in terms of test usefulness. The current reading comprehension test specifications were reviewed and evaluated, based on Davidson and Lynch's (1993) model of test specifications. In this research 16 eighth grade English language teachers, 25 English language heads of department, and 6 English language supervisors were surveyed to evaluate the implementation of the test specifications, and to identify the Kuwait English language teachers', department heads', and supervisors' practices in constructing, validating and administering their eighth grade reading comprehension tests. In addition, the study analyzed eight schools' first period reading comprehension tests to evaluate their usefulness in terms of test reliability, construct validity and content validity. The study shows that there is a need to improve the eighth grade reading test specifications in terms of providing text readability level, text type, instructions, scoring criteria, score interpretation criteria, validity, and reliability, in order to help test writers develop more useful reading tests.

Keywords : *test specifications, test development, assessing reading, language assessment, Kuwait*

Table of Contents

Acknowledgement	8
Abstract	5
List of Figures	9
List of Tables	10
CHAPTER ONE: INTRODUCTION.....	12
Statement of the Problem.....	12
Purpose Statement.....	13
Significance of the Research.....	14
CHAPTER TWO: REVIEW OF LITERATURE.....	15
Reading Comprehension.....	15
Test Development.....	16
Design Stage.....	16
Operational Stage	17
Test Administration Stage.....	17
Test Specifications	18
Choosing Suitable Texts	21
Test Usefulness	24
Reliability	24
Validity	26
Authenticity.....	29
Test Interactiveness.....	29
Test Impact.....	30
Test Practicality.....	31
CHAPTER THREE: METHODOLOGY	32
Context.....	32
Data Collection	32

Participants	33
Principles of Analysis.....	34
Item Analysis.....	34
CHAPTER FOUR: FINDINGS	36
The Eighth Grade Reading Test Specifications Analysis	36
Specification Number.....	37
Title of Specification.....	37
Related Specifications	37
General Description.....	37
Prompt Attributes	37
Response Attributes.....	38
Sample Item.....	39
Specification Supplement.....	39
Discussion of the Eighth Grade Reading Test Specifications Analysis.....	39
Analysis of the Eight Schools’ Tests	40
Item Analysis.....	40
Test Analysis General Discussion.....	60
Analyzing Five Problematic Items	61
Test Texts’ Readability Level Analysis	65
Readability Level of the Reading Text in the Student Book and the Workbook.....	67
ELSS, HODs, Teachers Questionnaire Analysis.....	68
The Test Specifications Yielding Tests Related to the Specified Syllabus	68
Test Construction Procedures Used	71
Procedures Used to Ensure Test Reliability	75
Eighth Grade Learner’s Level of Reading	80
CHAPTER FIVE: CONCLUSIONS	82
Research Question One.....	82

Research Question Two83

Implications and Suggestions for Improvement87

Suggestions for Further Research87

References.....89

List of Figures

Figure 1: The reading comprehension part of the 9th and 8th grade test specifications.....	36
--	----

List of Tables

Table 1. School GLB IF and ID of the Reading Test Items	41
Table 2. Item #1 DE.....	42
Table 3. Item #2 DE.....	42
The following table (Table 4) is a display for item #3 DE. Table 4. Item #3 DE.....	43
Table 5. School WLB IF and ID of Reading Test Items	44
Table 6. Item #1 DE.....	44
Table 7. Item #2 DE.....	45
Table 8. Item #3 DE.....	45
Table 9. School GSB IF and ID of the Reading Test Items.....	46
Table 10. Item #1 DE.....	47
Table 11. Item #2 DE.....	47
Table 12. Item #3 DE.....	48
Table 13. School WSB IF and ID of the Reading Test Items.....	49
Table 14. Item #1 DE.....	49
Table 15. Item #2 DE.....	50
Table 16. Item #3 DE.....	50
Table 17. School GLG IF and ID of the Reading Test Items	51
Table 18. Item #1 DE.....	52
Table 19. Item #2 DE.....	52
Table 20. Item #3 DE.....	53
Table 21. School WLG IF and ID of the Reading Test Items	54
Table 22. Item #1 DE.....	54
Table 23. Item #2 DE.....	55
Table 24. Item #3 DE.....	55
Table 25. School GSG IF and ID of the Reading Test Items	56
Table 26. Item #1 DE.....	56
Table 27. Item #2 DE.....	57
Table 28. Item #3 DE.....	57

Table 29. School WSG IF and ID of the Reading Test Items.....	58
Table 30. Item #1 DE.....	59
Table 31. Item #2 DE.....	59
Table 32: Item #3 DE.....	59
Table 33. The Schools' IF, ID, Type of Text, Flesch-Kincaid, and Texts'	
Word Count	60
Table 34: The Reading Texts in the Eighth Grade Student Book and the	
Workbook	67
Table 35. The Relationship Between the Test and the Syllabus.....	69
Table 36. Test Construction Procedures	71
Table 37. Validity Assurance Procedures.....	72
Table 38. Test Environment.....	76
Table 39. Scoring Reliability	77
Table 40. Statistical Item Analysis	79
Table 41. Learners' Reading Level	80

CHAPTER ONE: INTRODUCTION

In Kuwait, English is taught as a foreign language. Students start learning English in public schools right from the first grade in the primary stage, and continue through grade twelve in the secondary stage. Most of the students are Kuwaitis, and all the learners' mother tongue is Arabic. The eighth grade students are 13-14 years old and have studied English for seven years, having five-six English classes per week each academic year. English is one subject among a number of other subjects which the students have to learn at school.

Learners have to pass all school subjects, including English, in order to move to grade nine. The eighth grade general English exams are conducted once at the end of each of the four periods of an academic year. (The four periods are times of instruction that are seven weeks each.) The general English exams consist of six parts (vocabulary, grammar, language functions, set-books questions, writing, and reading comprehension). However, many students finish their secondary stage with reading comprehension problems. How did these learners go through the primary, intermediate, and secondary grade levels, without anyone identifying their reading comprehension weakness?

Statement of the Problem

One answer to this question may be found in the reading comprehension assessment in Kuwait public schools. Assessing learners' English language reading comprehension ability in Kuwait public schools is done using pen-and-paper tests. In all grades, at the end of each of the four seven-weeks teaching periods in the scholastic year, learners are given one reading text and are asked to read and answer a number of varied questions about the text.

I have recently noticed potential problems in the eighth grade, intermediate stage, first period reading comprehension test in Kuwait public schools. Examples of these problems are variance in the readability levels of the tests' reading passages, lack of clear and detailed item specifications, and lack of test administration instructions.

Test specifications are one of the core elements of the test item writing process (Bachman & Palmer, 1996; Saville, 2012). One reason for the importance of test specifications is that they play the role of a generative blueprint from which many

equivalent test tasks or items can be produced (Davidson & Lynch, 2002). However, there are not many studies that empirically evaluate whether and how test specifications contribute to the quality of a test (Zandi, Kaivanpanah, & Alavi, 2014). Similarly, the extent to which qualitative evaluation of test tasks or items can be helpful in detecting problems with them is still an open question for research (Zandi et al., 2014).

This research investigated how useful are the eighth grade intermediate stage first period reading comprehension tests in Kuwait, how effective the implementation of the test specifications is.

Purpose Statement

The purpose of this research is to find out how useful the eighth grade reading comprehension test specifications are. Fulcher and Davidson (2007) say that test specifications are a generative explanatory document for the creation of test items/ tasks. Test specifications provide the test writers with the nuts and bolts of how to phrase the test items/ tasks, how to structure the test layout, and how to locate the passages. More importantly, the specifications document tells us the rationale behind the test. Test specifications are often called blueprints, and from this test blueprint many equivalent tests can be developed. In Kuwait public schools, the English Language Supervision Unit (ELSU) has a document entitled “Distribution of Marks” (hereafter referred to as test specifications), which functions as the test specifications document.

A related issue is how useful the current eighth grade reading comprehension tests are, particularly in terms of representing the eighth grade reading curriculum content.

Thus, this research seeks to answer the following questions:

- 1- How effective are the current test specifications for the eighth grade first period reading comprehension test?
- 2- How effective is the implementation of the test specifications in the development of the current eighth grade first period reading comprehension tests in Al-Jahra?

Significance of the Research

This research may provide useful information to the ELSU in Kuwait about the effectiveness of the current test specifications in producing tests that are useful. In addition, the research may also contribute to improving the usefulness of the eighth grade reading comprehension tests in Kuwait public schools. This improvement can be accomplished by highlighting problems in the current reading comprehension testing, if any, and then suggesting some solutions. Improving the usefulness of the reading comprehension testing may also lead to improvement in reading comprehension instruction.

This thesis consists of four chapters. This first chapter introduces the study. The second chapter discusses pertinent issues about evaluation of Kuwait's eighth grade testing of reading comprehension: the theoretical framework of reading comprehension, what is involved in test development (particularly test specifications), choosing suitable texts, and test usefulness. The third chapter discusses the context in which the study was conducted. In addition, it discusses the methods by which the data was collected for the study, and how the data was processed and analyzed. Furthermore, the chapter provides a description of the research participants. The fourth chapter contains analysis and discussion of the current reading comprehension test specifications, based on Davidson and Lynch's (1993) model of test specifications. In addition, the chapter presents analysis of eight schools' administered tests, in terms of test usefulness. The test analysis is conducted in the form of test item analyses (distractor efficiency, item facility, and item discrimination). The fifth chapter presents the conclusions made from this research, implications, and suggestions for improvement and further research.

CHAPTER TWO: REVIEW OF LITERATURE

This chapter discusses issues pertinent to evaluation of Kuwait's eighth grade testing of reading comprehension: the theoretical framework of reading comprehension, what is involved in test development (particularly test specifications), choosing suitable texts, and test usefulness.

Reading Comprehension

Reading comprehension testing is a challenge for all English as second/ foreign language teachers. Simply put, teachers need to measure the learners' understanding of a given text.

It is important for test designers and test writers to know what the theoretical framework is that underlines reading comprehension, a multidimensional and complex construct. Carlson, Seipel and McMaster (2014) say, "reading comprehension theories help identify constructs that work during the process of comprehension and specify the relationships among them so that researchers can better operationalize the dimensions to be assessed" (p. 40). There are numerous theories and a considerable amount of research about the nature of reading and the assessment of reading. Alderson (2000) says that if reading itself is a skill, it must be possible to break it down into different levels of component sub-skills. Many researchers have attempted to divide reading skills into component sub-skills (see Munby, 1978; Haynes & Carr, 1990; Hughes, 2003; Alderson, 2000; Clarke, Truelove, & Hulme, 2013). For example, Davis (1968) defines eight reading sub-skills, which are the same sub-skills tested in the eighth grade reading comprehension test in Kuwait:

- 1- Recalling lexical item meaning
- 2- Drawing inferences about the meaning of a lexical item in context.
- 3- Finding answers to questions answered in an explicitly stated manner in paraphrase
- 4- Connecting ideas in the content
- 5- Drawing inference from the reading content

- 6- Identifying the writer's technique
- 7- Recognizing the writer's purposes, tone, attitude and mood
- 8- Following the structure of a reading passage

Alderson (2000) says that such lists offer a theoretically justified means of devising test items or tasks, and of isolating reading skills to be tested. They also suggest the possibility of diagnosing a candidate's reading problems, with a view to identifying remediation. They are very powerful frameworks for test construction.

Test Development

According to Bachman and Palmer (1996) in their classic model, test development is a process that consists of three stages: design, operationalization, and administration. Hughes's (2003) model of test development describes the same process but in 10 steps. These two models are described together here.

Design Stage

Bachman and Palmer (1996) say that in the design stage, the test designers describe in detail the component of the test design that assures that performance on the test tasks will correspond as closely as possible to language use, and that the test scores will be useful for the intended purpose. They indicate that the product of this stage is a design statement document that includes a description of the purpose(s) of the test, the target language domain and task type, the test takers for whom the test is intended, and the construct(s) to be measured. In addition, the document includes a plan for evaluating the qualities of usefulness of required and available resources and a plan for their allocation and management (Bachman & Palmer, 1996).

In this stage Hughes (2003) recommends:

- Making a full and clear statement of the testing 'problem'. Hughes says that this stage provides answers to questions such as What kind of test is it to be? What is the precise purpose of the test? And what abilities are to be tested?

Operational Stage

The next stage of test development is the operational stage. Bachman and Palmer (1996) say that this stage involves development of test specifications for the types of tasks to be included in the test, and a blueprint that describes how the test item writers will be instructed to form actual tests. According to Bachman and Palmer, this stage also involves developing and writing the actual test tasks, writing instructions, and the procedures for scoring the test. In the operational stage, the focus is the test scoring procedure and item development.

In this stage Hughes (2003) recommends:

- Writing complete test specifications for the test. Hughes says that in this stage a set of specifications is developed that includes information on content, test structure, timing, medium/channel, techniques to be used, criteria for levels of performance, and scoring procedures. (Test specifications are discussed in more detail below.)
- Writing and scrutinizing items. Once specifications are completed the writing of items can begin.

Test Administration Stage

According to Bachman and Palmer (1996), the third stage of test development is test administration. This stage involves giving the test to a group of individual, collecting information, and analyzing the information for two purposes: assessing the usefulness of the test, and making the inference or decisions for which the test is intended.

Bachman and Palmer (1996) explain that administration typically takes place in two phases: the *try out* and *operational test use*. Try out involves administering the test for the purpose of collecting information about the usefulness of the test itself, and for the improvement of the test and testing procedures. On the other hand, they say that operational test use involves administering the test in order to accomplish the specified purpose of the test, which is assessing the learners' ability. Information about test usefulness is also collected in the operational test use phase. (To clarify, in the try out, the test is administered as a trial/pre-test, while in operational test use the test is administered to assess the learners.)

In this stage Hughes (2003) recommends:

- Informal trialing of the test on a group of native speakers and then rejecting or modifying problematic ones as necessary.
- Trialing the test on a group of non-native speakers similar to those for whom the test is intended.
- Analyzing the results of the trial and making any necessary changes.
Hughes suggests two kind of analysis that should be carried out. The first is a statistical analysis that can reveal qualities such as test reliability as a whole and of individual items, for example how difficult the test items are, and how they discriminate between candidates with different levels of language proficiency. The second type of analysis is qualitative. Hughes says that responses should be examined for test item misinterpretation or unintended but correct responses.
- Calibrating scales. According to Hughes, in this stage, samples of performance should be collected which cover a full range of the rating scale. Then a team of experts looks at them and assigns each of them to a point of relevant rating scales. The assigned samples provide reference point for all future uses of the scale, and for training material.
- Validating. In this stage an evidence-based argument is developed for the accuracy of test content, consistency of test administration, and interpretation of scores
- Writing handbooks for test takers, test users and staff. Hughes says that each handbook may have rather different content, depending on the targeted audience.
- Training any necessary staff. According to Hughes, this may include training raters, computer operators, and proctors.

Test Specifications

Test specifications are a document which sets out what a test is designed to measure and how it will be tested. It is primarily used by test developers/item writers. According to Walters (2010), following test specifications is in contrast to the usual

practice of simply writing test items “from scratch,” where the items developed by item writers may or may not test the same intended skill. Test specifications can be thought of as a “recipe” that allows items writers to create test tasks efficiently and to realize specific testing goals consistently over time. The test specifications document may also be used by test evaluators and test users (Davies et al., 1999). According to Alderson, Clapham, and Wall (1995)

a test’s specifications provide the official statement about what the test tests and how it tests it. The specifications are the blueprint to be followed by test and item writer, and they are also essential in the establishment of the test’s construct validity. (p. 9)

The test specifications are a detailed document, sometimes confidential to the examining body. Its development is a central part of the test construction and evaluation process (Alderson et al., 1995). Davidson and Lynch (2002) say that

the purpose of a well written specification is to result in a document that if given to a group of similarly trained teachers working in a similarly constituted teaching context, will produce a set of test tasks that are similar in content and measurement characteristic. (p.15)

Therefore, those who use test specifications to construct a test need to have a clear description of whom the test is aimed at, what the purpose of the test is, what content is to be covered, what testing methods are to be used, how many papers or sections there are, how long the test takes, and so on (Alderson et al., 1995). The test specifications should be consulted when items and tests are developed and reviewed. Davidson and Lynch (2002) say that this central role of specifications works because the ultimate goal of specifications is clarity. Therefore, the test specifications need to be clearly written so that they can be referred to easily (Alderson et al., 1995).

Alderson et al. (1995) contend that test specifications should include answers to the following questions:

- What is the purpose of the test? Is it a placement test? Is it a progress test? Is it an achievement test? Is it a proficiency test? Or, is it diagnostic test?

- What sort of students will be taking the test, their age, sex, stage of learning/ level of proficiency, first language, country of origin, nature of education, reason for taking the test?
- How many papers/sections should the test have? How much time should be specified? Four separate two-hour papers? Two 45-minute sections? Reading tested separately from language structure? Writing and listening integrated into one paper? and so on.
- What should be the sources of text, the topics, and the degree of authenticity? How long or difficult should they be? What language function should be embodied in the texts? Definition? Persuasion? Summarizing, etc.? How complex should the language be?
- What language skills should be tested? Are micro skills specified, and should items be designed to test micro skills individually or in integrated tasks?
- What language constituents should be tested? Is there a list of grammatical structures to be included? Is the vocabulary specified in some way, as in frequency lists, etc.? Are notions and functions such as speech act features specified?
- What kind of tasks are required to be assessed as discrete point? Objectively assessable? Integrative? Simulated “authentic”?
- How many tasks/items are required for each section? What is the relative weight for each item/task?
- What test methods are to be used? Gap filling? Multiple choice? Matching? Short answer? Transformation? Picture description? Essay? Structured writing?
- What rubrics are to be used? Will examples be required to help test takers know what is expected? Should the criteria by which test takers will be assessed be included in the rubric?

One well known model of test specifications is the Davidson and Lynch (1993) model of test specifications. This test specifications model will be used in this research to analyze the Kuwait reading comprehension tests. In this model the construct, format, items, and scoring are described. The specifications show the skills to be tested, number and type of items or tasks per skill, weighting for each item, and any special materials

needed. In Davidson and Lynch's model, item specifications include the following:

- *Specification number*: an index number.
- *Title of specification*: a short title that characterizes the specification.
- *Related specifications if any*: list numbers, titles of specifications, or both, related to the particular item; for example, in a reading test, separate detailed specifications would be given for the text readability and for each item type.
- *General description*: a short general statement of the skill(s) to be tested and the purpose for testing it/them.
- *Prompt attributes*: a clear, complete and detailed description of what items/tasks the student will encounter.
- *Response attributes*: a clear, a complete and detailed description of what will constitute success and what will constitute failure in the student's answer. This description should include the criteria for evaluating or rating the student's response. Two types of responses may be elicited: selected response (true/false, matching, multiple choice) or constructed response (prolonged essay, short answer, fill in the blank) for which a complete and detailed description is needed of the type of response the test taker will perform.
- *Sample item*: a sample item/task that illustrates and reflects the specification.
- *Specification supplement*: a detailed explanation of any additional information required to construct items for a given specification. Vocabulary specifications might provide a list of words and their sources (word frequency lists, dictionaries, etc.); grammar specifications might list the precise grammar forms to be tested.

Choosing Suitable Texts

Describing the target language use, which is a part of the design stage in the test development process, involves describing task types in terms of distinctive characteristics (Bachman & Palmer, 1996). This description includes the characteristics of reading texts in a reading comprehension test. A test writer's choice of reading texts may affect test validity. Some tests are constructed with built-in lack of validity, which may occur when items do not match the objectives or the content of the instruction. Also, it may happen when the test wrongly requires knowledge of structures and vocabulary to which the

students were never exposed (Henning, 1987). Therefore, choosing a reading text that is suitable for the learners' level is crucial. According to Hughes (2003), texts that test takers are expected to be able to deal with can be specified along a number of parameters: readability or difficulty type, form, graphic features, topic, style, intended readership, length, range of vocabulary and grammatical structure.

The test writer needs to be aware of the suitable length (number of words) that the text in the test contains in order not to negatively affect the learners' comprehension of the reading passage. Two factors can affect text comprehension. These two factors are complexity of structure and sentence length. When these two factors both occur in a text, text comprehension will be reduced (Pearson & Camperell, 1981, as cited in Jalilehvand, 2012).

Providing test takers with texts that are well matched to their reading abilities has been a challenge for educators. McNamara, Crossley, and Greenfield (2008) contend that accurately predicting the difficulty of reading texts for language learners is extremely important for teachers, publishers, writers, and others to ensure that texts' readability level matches prospective readers' proficiency.

The creation and use of readability formulas have been a solution to this problem. The majority of these formulae are based on two broad aspects of comprehension difficulty. The first one is semantic or lexical features, and the second one is syntactic or sentence complexity (Chall & Dale, 1995). According to Chall and Dale (1995), formulas that depend on these two variables are popular because they are easy to simplify in a text. For instance, a text written for low level readers generally contains more frequent, everyday lexical items and shorter sentences. Therefore, measuring the sentence length and word frequency of a text provides a basis for understanding how readable the text is.

An example of a traditional readability formula is the Flesch-Kincaid grade level (Kincaid, Fishburne, Rogers, & Chissom, 1975), which is based on Flesch reading ease (Flesch, 1948). According to Zamanian and Heydari (2012), the popularity of the Flesch reading ease formula has made it a leading authority on readability. In Flesch reading ease, the readability rating is based on the average number of syllables per word and the

number of words per sentence. The Flesch reading ease formula rates texts on a scale of 100 points; the higher the score, the easier it is to understand the reading text.

According to Hippensteel (2015), a text readability and the grade level that suits its level can be determined using Flesch–Kincaid readability test, where:

$$\text{Readability grade level} = 0.39 (\text{total words}/\text{total sentences}) + 11.8 (\text{total syllables}/\text{total words}) - 15.59$$

Hippensteel (2015) says that on this scale, a text entirely composed of monosyllabic, single-word sentences would have a low potential readability grade level of –3.40, and most reading materials intended for college undergraduates would preferably fall between 12 and 17. Hippensteel (2015) also says that there are simple computational methods in which the passage to be analyzed is cut directly from the text and pasted it into Microsoft Word program, which has a function to analyze and calculate Flesch Reading Ease and Flesch-Kincaid Grade Level.

He says that this process breaks down when technical or scientific texts, and even textbooks, are analyzed. For example, with the sentences “This is a well written text” and “This is a well written text (Hippensteel, 2015).” Microsoft Word assigns readability grade levels of 2.4 and 10.3 respectively. For scientific articles with multiple citations (as in the second example), this readability grade level inflation can be significant.

As mentioned, in the Flesch readability formula the score ranges from 0 to 100, with 100 corresponding to the lowest reading difficulty and 0 corresponding to the highest reading difficulty (Zamanian & Heydari, 2012). For example, if the readability score of a text ranges from 90-100, this means that the text is very easy to read, and thus 5th grade learners would be able to comprehend the text. On the other hand, if the readability score ranges between 0-30, this means that the text is very difficult to comprehend, and thus a college graduate may be able to comprehend it (Zamanian & Heydari, 2012).

Although the Flesch score is based on native English language speakers, it still is useful for giving an indication of reading difficulty of a text that is given to nonnative English language speakers. Greenfield (cited in Zamanian & Heydari, 2012) used a corpus of a number of academic texts that had been evaluated for textual difficulty for L1

speakers using a cloze test. Greenfield used this corpus, and compared L2 learners' performance to L1 readers' performance under a constant passage set. He found that traditional L1 readability formulas such as the Flesch-Kincaid Grade Level and Flesch Reading Ease formula had strong correlation to L2 cloze test performance. This study shows that the Flesch-Kincaid Grade Level and Flesch Reading Ease formula can be useful for giving an indication of reading difficulty of a text that is given to nonnative English language learners.

Test Usefulness

The quality of language tests, such as the Kuwait eighth grade English exam, is affected by a wide variety of factors. One of the main considerations is the usefulness of the test. What makes a test useful? According to Bachman and Palmer (1996), test usefulness includes six test qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality.

Reliability

According to Bachman and Palmer (1996), reliability is defined as consistency of measurement. They say that a reliable test score will be consistent across different testing situations. Thus, reliability can be considered to be a function of the consistency of scores from one set of tests and test tasks to another. However, according to Alderson et al, (1995), if the same test is given to the same learners more than once, the students' scores may vary from one attempt to another. They say that some of these variations in the learners' scores might be caused by systematic differences such as the learners' improvement in the targeted skill/skills. Other variations in the learners' scores might be due to random "errors," which are unsystematic changes caused by, for example, distracting noise in the examination place or students' lack of concentration. Alderson et al. (1995) say,

The aim in testing is to produce a test which measures systematic rather than unsystematic changes, and the higher the proportion of systematic variation in the test score, the more reliable the test is. A perfectly reliable test would measure only systematic changes. (p. 87)

When developing a language test, it is important to identify potential sources of errors in the measurement of language ability in order to reduce their effect and have a more reliable test. Examples of these sources of errors related to the test takers are poor health, lack of interest, fatigue, and lack of test wiseness (Bachman, 1990). Therefore, Alderson et al. (1995) contend that by reducing the causes of unsystematic variation to a minimum, test developers may make their tests more reliable. Three particularly crucial aspects of reliability are scoring, administration, and the actual test itself. However, Bachman and Palmer (1996) say that it is not possible to eliminate inconsistencies entirely and test designers can only try to minimize the effects of any potential sources of inconsistency that are under their control.

Scoring reliability (rater reliability). Rater error, bias, and subjectivity may affect test scoring. When discussing rater reliability, two terms will appear: *inter-rater reliability* and *intra-rater reliability* (Alderson et al., 1995). Inter-rater reliability is achieved when two or more scorers yield consistent scores of the same test. In other words, inter-rater reliability is the level of consensus between more than one independent rater in their judgments of test takers' performance (Davies et al., 1999). Intra-rater reliability, in contrast, is the degree to which a particular rater is consistent in using a proficiency scale (Davies et al., 1999). An examiner is judged to have high intra-rater reliability if s/he gives the same marks to the same test items on two different occasions (Alderson et al., 1995). However, there is evidence that raters sometimes apply the contents of the scale in quite different ways (Lumley, 2002). Some raters appear to differ in the emphases they give to the various components of the scale descriptors. Thus, careful specification of scoring rubrics can increase both inter- and intra-rater reliability (Lumley, 2002).

Test administration reliability. In addition to challenges with rater reliability, low reliability may result from the way a test is administered. Bachman (1990) indicates that random factors may affect scores, including temporary conditions such as test takers' mental alertness or emotional state, change in the test environment from one day to the next, or idiosyncratic differences in the way different test administrators carry out their responsibilities. He adds that the testing environment may affect the reliability of the test, and one aspect of the testing environment is the physical condition of the test

environment like the presence of noise in the environment or humidity. Bachman also says, “test performance is also affected by the characteristics of the test method used to elicit test performance” (p.111). For example, in a silent reading comprehension test, some test takers may perform differently if the reading text were to be read to them by someone (a change during test administration), instead of them reading it. The change in method of testing would change the construct being evaluated. Instead of testing a reading comprehension sub-skill, the altered test would be testing listening comprehension sub-skills.

Test reliability. The nature of the test itself may cause measurement errors. Differences in the clarity of the test instructions, the time of the test administration, and the extent of test administrator interaction with examinees are potential errors in the sources of measurement (Henning, 1987). In addition, test items can be a source of error and thus should be carefully designed to be more reliable. For example, in multiple-choice questions (MCQs) the items need to be equally difficult, items need to be well distributed, and the distractors need to be well designed. Furthermore, a test’s difficulty level may have an influence on the test’s reliability. When tests are overly difficult or easy for a given group of students, it becomes more difficult to measure learners’ abilities (Henning, 1987).

Validity

The most important criterion of an effective test is validity, the second quality of test usefulness. At the most basic level a measurement, such as a test, is valid if it does what it is intended to do, which is typically to act as an indicator of an abstract concept which it claims to measure (Davies et al., 1999). Henning (1987) defines validity as the appropriateness of a given test or any of its components as a measure of what is purposed to be measured.

Any test may be valid for some purposes, but not for others. Validity is not an all or nothing matter (Alderson et al., 1995). This characteristic means that users will have to use their own or other expert judgment when deciding, on the basis of evidence, on the relative validity of a test (Alderson et al., 1995). When designing a test, the *intention* is to measure something real. Therefore, validity enquiry should be concerned with finding out

whether a test measures what is intended (Fulcher & Davidson, 2007). Subsequently, test scores' wrong interpretation may affect test validity. According to Bachman and Palmar (1996), the method used to make decisions about individuals are a crucial part of the measurement process. The process of scoring and interpreting scores plays an important part in insuring that the test scores are valid.

There are varied ways of finding out how valid a test is, using different types of evidence as appropriate for the testing context.

Construct validity. Most testers have identified three main evidence of validity: *empirical, construct* and *content validity* (Alderson et al., 1995). Of concern in this research are construct and content evidence of validity because they are the most pertinent for tests such as the Kuwait English tests, which do not have in place a system of empirical validation research.

The purpose of construct validation is to provide evidence that the theoretical constructs underlying the language skills being measured are themselves valid (Henning, 1987). Brown and Abeywickrama (2010) say that “a construct is any theory, hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions.” (p. 33). They explain that a construct can be linguistic or psychological. Proficiency, communicative competence, and fluency are examples of linguistic constructs. Self-esteem and motivation are psychological constructs. Davies et al. (1999) say:

Construct validity of a language test is an indication of how representative it is of an underlying theory of language learning. Construct validation involves an investigation of the qualities that a test measures, thus providing a basis for the rationale of a test. (p. 33)

Alderson et al. (1995) say that every test has a theory behind it, and every theory contains constructs or psychological concepts (which are its principal components) and these components' relationship. For example, some theories of reading state that there are many different sub-skills/constructs involved in reading, like skimming, scanning, etc., which could be included in a test of reading. Alderson et al. (1995) say that construct validation involves assessing the quality of a test's measurement of the construct(s).

Construct validity pertains to the appropriateness and meaningfulness of the interpretations that we make on basis of test scores (Bachman & Palmer, 1996). Construct validation answers questions like these: What do the test scores mean? What do they tell us about the test takers' ability? Does the test in fact measure the targeted ability/abilities or not? Therefore, for validation purposes, test specifications need to establish the theoretical framework and spell out relationships among its constructs. Thus, one way for test developers and test writers to assess construct validity is to see to what extent the test is based upon the pertinent underlying theory.

Content Validity. This aspect of validity is “a conceptual or non statistical validity based on a systematic analysis of the test content to determine whether it includes an adequate sample of the target domain to be measured” (Davies et al., 1999, p. 34). Mousavi (2009) says that if a test samples the subject matter about which conclusions are to be drawn, and if it requires the candidate to perform the behavior that is being measured, it can claim content validity. Content validation depends on logical analysis of the test content to investigate whether the test contains a representative sample of the relevant language skills (Alderson et al., 1995). In order to ensure content validity of a test, it is necessary to seek the advice of content experts. In addition, it is important to develop clear and detailed specifications for test items in different domains representative of the objectives of instruction (Henning, 1987). A common way is for experts to analyze the content of the test and compare it with the a formal teaching curriculum or syllabus, or it may be a domain specification (Alderson et al., 1995).

Therefore, one aspect of validity in a test is that the test has an aim. According to Siddiek (2010), test aimlessness is a main cause of ineffectiveness in assessment. Siddiek (2010) asks, “how can we as teachers construct a test or examination, or set an assignment, unless we have been able to learn which skills we wish the student to acquire?” (p. 135). He says that if we want our test to be valid, it must measure what we have set as objectives in our lessons. Siddiek says that our task then in preparing a test is to write test items which cover all the material taught and which measure the course objectives. In the case of a reading achievement test, content validity refers to the degree to which a test measures the reading program objectives.

Alderson et al. (1995) say that, in principle, a test cannot be valid unless it is reliable. If a test does not measure a targeted skill consistently, then it cannot always be measuring it accurately. On the other hand, it is possible for a test to be reliable but invalid. For example, a test may not be measuring what it is supposed to, although it consistently gives the same result. Therefore, although reliability is needed for validity, it alone is not sufficient. According to Bachman (1990), the concerns of validity and reliability lead to two complementary objectives in designing and developing tests: 1) minimizing the effects of measurement error, and 2) maximizing the effect of the language ability we want to measure.

Authenticity

Authenticity is the third quality of usefulness. According to Bachman and Palmer (1996), authenticity is “the degree of correspondence of the characteristics of a given language test task to the features of a target language task” (p. 23). Brown and Abeywickrama (2010) say that when a claim is made for authenticity in a test task, it is said that this task is likely to be enacted in the real world. According to Messic (1996), “in the case of language testing, the assessment should include authentic and direct samples of the communicative behaviors of listening, speaking, reading, and writing of the language being learned.” (p.4). Brown and Abeywickrama (2010) say that an authentic test contains as natural language as possible, has items/tasks that are contextualized and not isolated, includes meaningful and interesting topics, provides some thematic organization to items/tasks, such as through a story, and offers tasks that replicate real-world tasks.

Test Interactiveness

Test interactiveness is the fourth quality of test usefulness. Bachman and Palmer (1996) define interactiveness as the extent and type of involvement of the test taker's in accomplishing a test task. Current scholarship in the field of applied linguistics shows that communicative language consists of interactions between aspects of strategic competence and language knowledge (see Bachman & Palmer, 1996; and Douglas, 2000). According to Weigle (2002), language knowledge (i.e., grammatical knowledge, knowledge of rhetorical or conversational organization), and strategic competence (i.e., evaluating the

correctness or appropriateness of the response, deciding how and whether to respond to the communicative situation) are relevant characteristics for language testing. She also says that

interactiveness is important in language testing because these characteristics are all engaged in actual language use. Thus an assessment task that only involves language knowledge but not the other characteristics may give us some idea of how much a test taker knows about the language, but not about how well he or she can use the language. (p. 53)

For example, in a non-interactive reading task, the task would require test takers to identify and underline all the past tense verbs in a text. In this task, test takers need to use their knowledge of English grammar, but they do not need know anything about the topic of the passage, and their need to metacognitive strategies is limited. Such a task may not engage test takers' interest as much as a more interactive task would.

On the other hand, an interactive reading task would involve both linguistic competence and strategic competence. According to Luo (2015), “the interaction between the test taker and the task can be described as how a test task engages the test-taker’s language knowledge, metacognitive strategies” (p. 20). Highly interactive test tasks require test takers to demonstrate their strategic competence, in addition to their linguistic knowledge (Weigle, 2002).

Test Impact

Test impact is the fifth quality of test usefulness. According to Bachman and Palmer (1996), impact refers to consequential validity. Brown and Abeywickrama (2010) say that “consequential validity encompasses all the consequences of a test, including such considerations as its accuracy in measuring intended criteria, its effect on the preparation of test-takers, and the social consequences of a test’s interpretation and use” (p. 34). Weigle (2002) defines impact as “the effect that the tests have on individuals (particularly test takers and teachers) and on larger systems, from a particular education system to the society at large” (p. 54). Bachman and Palmer (1996) say that test developers and users must always consider the social and educational values and goals that inform test use, which may vary from one culture to another. For example, one

culture may emphasize and give more value to individual achievement, while another culture may value group work more. Bachman and Palmer (1996) say that a test developer needs to think carefully of what might happen as a result of using the test for a particular purpose.

Bachman and Palmer (1996) say that “*washback*” is an aspect of test impact. The most common view of washback in language teaching is that it is “the effect of testing on teaching and learning” (Hughes, 2003, p. 1). Messic (1996) suggests the term “washback validity”, reflecting the role of washback in the validity of a test. A test's validity should be reflected in a degree of positive influence on teaching (Morow, 1986, cited in Messic, 1996). In order to have maximum positive washback, Messic says, the difference between the learning tasks and test tasks should be minimal.

Test Practicality

The last but not the least test quality that needs to be considered in test usefulness is practicality. Practicality is primarily related to the ways in which the test will be implemented and developed. Bachman and Palmer (1996) define practicality as “the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities” (p. 36) In other words, if the required resources for implementing, developing, or using the test exceed the resources available, the test will be impractical. According to Bachman and Palmer (1996), these resources are human resources, material resources, and time. Examples of human resources are test writers, raters or scorers, test administrators, and clerical support. Examples of material resources are place (room for test development and test administration), equipment, computers, tape and video recorders. Examples of time are the time specified for the test development, test administration, writing and scoring (Bachman & Palmer, 1996).

CHAPTER THREE: METHODOLOGY

This chapter discusses the context in which the study was conducted, the methods by which the data was collected, a description of the research participants, and how the data was processed and analyzed.

Context

The study was conducted in Al-Jahra area, in Kuwait. Al-Jahra area contains 32 intermediate stage public schools, 16 for boys and 16 for girls. The schools are varied in terms of number of students and school administration quality.

The quality of school administration was chosen based on the advice of the manager of the intermediate stage (level), who is the administrative supervisor of the intermediate stage schools in Al-Jahra educational area. Some elements which were taken into consideration by the supervisor when deciding upon school administration quality, include the number of complaints from learners, teachers and parents, and good learner scores compared to other schools.

I chose eight different intermediate stage schools from Al-Jahra as a representative sample for Kuwait intermediate stage schools: four girls' schools and four boys' schools. As for the girls' schools, school number one was chosen from the girls' schools that have a large number of learners. School number two was chosen from the schools that have a small number of learners. School number three was chosen from the schools that have low quality school administration. Finally, school number four was chosen from the schools that have high quality school administration. The same procedures were followed with the four boys' schools.

Data Collection

Four sources of material were analyzed in this research: 1) the Kuwait eighth grade reading comprehension exam specifications, which is a document on the English language supervision website 2) administered tests from the eight schools (described above), 3) surveys of English language supervisors, heads of English language departments and English language teachers in Al-Jahra, and 4) the eighth grade reading texts in the student book and the workbook.

First, the current reading comprehension test specifications document was analyzed, which is a unified document for the intermediate schools in Kuwait (see Appendix A).

Second, the eight schools' administered tests were analyzed for test usefulness, as well as distractor efficiency, item facility and item discrimination, to evaluate the effectiveness of the tests' items.

Third, English language teachers, HODs and ELS were surveyed, using pen-and-paper questionnaires for ELS and HODs (see Appendix B) and teachers (see Appendix C). These questionnaire responses were analyzed to see what guidance the current test specifications provide for them, possible problems, and what improvements these test specifications need. The questionnaire took 10-15 minutes to be completed.

Fourth, the first period reading lessons in the eighth grade student book, workbook, and the reading texts in the tests were analyzed using the Flesch-Kincaid readability scale to determine the level of the reading texts and to see to what extent the test reflected what was taught. Furthermore, the tests' reading texts' readability levels were evaluated to see to what extent the tests followed the test specifications. Then, the test content was compared with the first period reading lessons in the eighth grade student book and workbook to see to what extent the test content related to what was being taught.

Participants

The research involved 25 English language heads of departments (HODs), 6 English language supervisors (ELs), and 16 eighth grade English language teachers.

The 25 HODs were all the English language HODs in the area. They ranged in terms of experience as HOD from 1 year to 15 years. Of the HODs, 14 HODs were females and 11 were males. The HODs were chosen to do the questionnaire because they were responsible for explaining the test specifications to the teachers. In addition, they were responsible for reviewing the exams before submitting the tests' draft to the ELS. The HODs were responsible for all the procedures related to test writing, administration, scoring, analyzing scores, and reporting the exam results. Thus, they would have knowledge of how the tests were developed.

The six participating ELSs were all working in Al-Jahra area. They ranged in terms of experience as ELS from one year to ten years. Of the chosen ELS, three were females and three were males. They were chosen to do the questionnaire because they were the intermediate stage schools' ELS and they were responsible for explaining the test specifications to the HODs and teachers. In addition, they were responsible for reviewing and approving the tests before they were administered. They, too, were knowledgeable about the tests in terms of the specifications, test content and administration.

The 16 eighth grade teachers were chosen based on their role in the eighth grade teaching and test development in their schools. The teachers were chosen from the targeted sample of schools. Half of the teachers were females from the girls' schools, and the other half were males from the boys' schools.

Principles of Analysis

The test specifications, eighth grade textbooks, questionnaires, and tests were analyzed. The eighth grade reading comprehension test specifications were analyzed qualitatively, through comparison with the Davidson and Lynch (1993) model of test specifications. The survey data were analyzed quantitatively, comparing the perspectives of ELSs, HODs, and teachers from large/small schools and high/low quality administration schools.

The tests were analyzed quantitatively and qualitatively. First, the readability levels were analyzed to see if they reflect the reading level of the reading texts in the textbook. Second, test item analysis results were displayed in tables. (See discussion of item analysis below.)

Item Analysis

To make sure that a test is working well, there are some procedures test designers need to be familiar with, which include item analysis. There are different kinds of item analysis, and two in particular are relevant for classroom-based assessment: item facility (IF) and item discrimination (ID) (Bailey, 1998). IF is an index of the easiness of an individual item was for the candidates. It is the percentage of how many test takers answered the item correctly. In other words, it gives item writers an idea of how easy the

item is for the trial sample of test takers (Alderson et al., 1995). ID is the extent to which an item in a test differentiates between low and high-performing students on a test. If an item is working well, test developers should expect more of the top scoring/ high-performing students to answer it correctly than the low-performing students. If the high-performing students answer the item wrongly while the low-performing students answer it correctly, there is a problem with the item, and it needs investigating (Alderson et al., 1995).

These simple measures can give the item writers and test designers some idea about the effectiveness of an item. According to Alderson et al. (1995), if the IF of an item is 0.0, this means that 0% of the students have answered it correctly, which indicates that the item is very difficult. On the other hand, if the IF is 1.0, this means that 100% of the learners have answered it correctly, which indicates that the item is very easy. According to Alderson et al. (1995) in both cases the item will not provide useful information about the students' ability. Brown and Abeywickrama (2010) say that appropriate items have IF scores that range between 0.15 and 0.85. Alderson et al. (1995) say that items that have IF of 0.5 provide the widest scope for variation among students. Items writers are satisfied with ID of 0.4 or above.

In addition, distractor efficiency (DE) is an important measure of multiple-choice questions (MCQs). Each distractor (wrong answer) should have a degree of effectiveness and discrimination. The degree of effectiveness is measured by the number of students choosing a specific distractor. If no test taker chooses a particular distractor, this means that the distractor is not participating effectively in making the question challenging. On the other hand, when a distractor attracts numerous test takers (has a high DE score), this might show that the item is a badly posed question (Goodrich, 1977). Although a high DE does not necessarily indicate a bad question, it raises questions about the effectiveness of the test item.

As mentioned above, four sources of material were analyzed in this research: the Kuwait eighth grade reading comprehension exam specifications, the questionnaire responses, administered tests from eight schools, and the student books. The results of this analysis are presented in Chapter Four.

CHAPTER FOUR: FINDINGS

This chapter presents analysis and discussion of the current eighth grade reading comprehension test specifications based on Davidson and Lynch’s (1993) model of test specifications. In addition, the chapter contains analysis of test usefulness of eight schools’ administered tests. The test analysis is conducted in the form of test items analyses (distractor efficiency, item facility, and item discrimination). The texts in the eighth grade textbook and work are analyzed, as are the HOD, supervisor, and teacher questionnaire responses.

The Eighth Grade Reading Test Specifications Analysis

The eighth grade reading comprehension test specifications are part of the eighth and ninth grade general English language test specifications document (see Appendix A). This specifications document contains test item specifications for all the eighth and ninth grade exams. The different parts in the eighth and ninth grade tests include vocabulary, structure, language functions, writing, and reading. The eighth and ninth grade general English language test specifications show that the first and third period exams have fewer test items in all the parts of the test, than do the second and the fourth period tests. Also, the test specifications show that the first and third period tests should be done in no more than one hour and a half, and the second and fourth period tests should be done in no more than two hours. In addition, the weight of the items differs; in the first and third period tests some items have fewer marks than in the second and fourth period tests. Figure 1 shows the reading comprehension part of the test specifications.

No	Branch	Types of questions	First & Third Periods			Second & Fourth Periods & 2 nd Session		
			Item	Mark	Total	Item	Mark	Total
VI	Reading Comprehension	<u>Unseen (passage /e-mail/ letter / short story /dialogue)</u>						
		<u>Grade (8) (160 – 180 words)</u>	3	1	3	4	2	8
		<u>Grade (9) (200 – 220 words)</u>	3	1	3	3	2	6
		A- Multiple choice (a, b , c & d) (reference words / word meanings / main idea / title ...etc.) B- Productive questions (Questions should include inference / prediction / guessing / anticipation.. etc.)	6		6	7		14

Figure 1. The reading comprehension part of the ninth and eighth grade test specifications

It is noted that some newly appointed teachers/HODs might find difficulty understanding this test specification document because of its format and the lack of clear organization of the tables. They may need additional explanation to be able to understand its content, in particular, what the numbers for item, mark, and total indicate.

Specification Number

As seen in Figure 1, the test specifications contain an index number for the parts of the general test specifications, including the eighth and ninth grade reading comprehension sections. In addition, the items of each part of the test are given letters.

Title of Specification

As can be seen in Figure 1, this test specifications document contains a specification title to outline skills across the specifications, such as writing, grammar, language functions, and vocabulary.

Related Specifications

The test specifications also contain related specifications for each part of the test. For example, it can be seen in Figure 1 that in the reading test specifications, the number of items, and grade distribution for each item are provided.

General Description

There are suggestions of what types of reading skills are to be tested in the eighth grade reading test. For example, multiple choice questions (MCQs) are assigned to test word meaning/reference words. However, there is no general statement of the reading skill to be tested. For example, there are no statements to describe the purpose of each suggested item, or the reason for assessing a particular skill. Each item should assess a particular reading construct, such as finding inexplicitly stated information in the text. Thus, the test specifications do not indicate the purpose for testing. Also, the specifications do not provide a general sense of the mandate or the contextual and motivational constraints in this particular test setting.

Prompt Attributes

As indicated in the reading comprehension specifications, there are suggestions of

two main types of questions, MCQs and short answer questions, and for each main type there are suggestions of what to test, such as main idea, word meaning, best title, or word reference for the MCQs. However, the expression "...etc." at the end of the suggested items may make the test item writers not sure of what other elements could be tested. In addition, the number of choices is specified, indicating that the MCQs should have four choices (a, b, c, and d).

In relation to the reading text, the specifications indicate that 1) it should be unseen, which means that it should be the first time the learners read a particular text, 2) the text's number of words should be 160 -180, and 3) the type of text should be a passage, email, letter, short story, or dialogue. However, there is no indication of when (first, second, third or fourth period) to use each type of reading. This lack of clarity may make the test writers provide different text types for the same period test in different schools and thus produce extremely different tests. The purpose of well written test specifications is to provide indications as to how to produce a set of test tasks that are similar in content and measurement characteristics. However, this reading comprehension test specifications may not help in producing a set of test tasks that are similar in type of reading text.

In addition, there is no indication, of any kind, as to the readability level of the texts to be included in the tests. This lack of specification may make the test writers use reading texts with different readability levels. Thus, the test specifications may not be helpful in producing a set of test tasks that are similar in content and form.

Response Attributes

In contrast to the specifications for the MCQs, it can be seen that for the short answer questions there is no clear description of how the learners are going to provide answers in response to the prompt attribute (e.g. words, phrases, sentences), or what would constitute a failure or success. In particular, there are no criteria for evaluating or rating the responses, especially the short answer questions. For example, what would be done if a learner provided a correct answer but with some spelling mistakes, or what if a learner provided an incomplete answer? The test specifications give no indication of how to assess such responses.

Sample Item

There are no illustrative sample items that reflect the specification of each test item type. This lack may result in the test writers developing widely divergent test items.

Specification Supplement

The eighth grade reading test specifications provide information related to the reading texts. It is stated that the reading texts should be unseen, which means that the learners should not have encountered the texts before. Further information included is the type of text, such as a dialogue, a letter, etc. In addition, the specifications provide the range for the number of words to be included in the texts. However, there is no explanation of any other additional information needed to construct the tests. For example, there is no text readability level recommended for the test reading text, no rubrics, no font size. Nothing indicates that an answer key should be provided, and there is no list of reading texts or suggestions of sources from which the reading test passage may be chosen.

Discussion of the Eighth Grade Reading Test Specifications Analysis

As seen in the eighth and ninth grade joint general test specification document and the eighth grade reading comprehension test specifications part, the general test specifications document includes the language domains, such as reading an email/a letter, which are representative communicative tasks for work or community. In addition, language skills to be tested, such as reading, writing, vocabulary, structure, and language functions, are provided. The test specifications also indicate the test takers for whom the test is intended: eighth grade learners. However, the specifications do not include a description of an eighth grade learner in terms of range of language proficiency level based on any framework.

The reading test specifications include task types such as MCQs and short answer questions. In addition, they include examples of the construct(s) to be measured, such as word reference, word meaning, skimming (main idea or best title), and scanning (inference or prediction). However, the construct validity of the test may be affected by the specifications suggesting some examples of the construct then following these

suggestions with “etc.,” and not providing the test writers with all the possible constructs that the test writers should/may test.

One main issue that construct validation addresses is whether the test in fact measures the targeted ability/abilities or not. However, it is seen that this test specifications document may yield tests that assess different abilities, and thus it is unknown what abilities are targeted. By extension it would also be unknown to what extent the tests measure these ill-defined abilities.

Also, there are other factors that the test specifications document lacks. It does not include a clear description of the purpose(s) of the test. In addition, the document does not include a plan for evaluating the qualities of usefulness of required and available resources or a plan for their allocation and management. In fact, it does not indicate what these elements are. This lack of detailed description of a component of the test may make it difficult to ensure that performance on the test tasks corresponds as closely as possible to real world language use, and that the test scores are useful for the intended purpose.

Analysis of the Eight Schools’ Tests

Eight schools’ administered tests are analyzed for test usefulness. This test analysis is conducted in the form of test item analyses (distractor efficiency, item facility, and item discrimination) to evaluate the effectiveness of the tests’ items. As mentioned above, that appropriate items have IF scores that range between 0.15 and 0.85 and items writers are satisfied with ID of 0.4 or above. In addition, the reading texts’ readability levels were examined and evaluated to determine to what extent each test followed the test specifications. The following eight sections are the results of each of the eight schools’ test item analysis.

Item Analysis

GLB school test item analysis. The first reading comprehension test is from a boys’ school. It is coded in this paper as GLB. According to the intermediate stage manager, this school has a good quality administration. It has a large number of students in all grades (832 students).

The reading comprehension test in school GLB consists of six items: three MCQs and three short answer questions. The Flesch-Kincaid readability level of the reading passage in this school's test is 5.8 (see Appendix D). The number of words in the passage is 194 words. The reading is a factual reading text about sleeping, information which could be in the learners' background knowledge. MCQ items #1, #2, and #3 were about the main idea of a paragraph in the text, word meaning and word reference, respectively. The short answer items #4, #5, and #6 were about specific information in the passage. (See Appendix E for the scores of the students in each item in the reading test and each student's total test score.)

ID and IF give indication of how useful the school GLB test items were. The following table (Table 1) shows school GLB reading test item facility (IF) and item discrimination (ID).

Table 1. School GLB IF and ID of the Reading Test Items

Item No.*	1	2	3	4	5	6
IF	0.33	0.39	0.70	0.46	0.66	0.21
ID	0.12	0.5	- 0.12	0.5	0.5	0.37

*Note: Items 1-3 are MCQ and items 4-6 are short answers.

It can be seen in Table 1 that the IF ranges between 0.21 and 0.70, which is a broad range of difficulty, and indicates the items may provide information about the learners' reading ability.

It can be seen that the ID of item #3 is -0.12, which means that more low-performing than high-performing students answered this item correctly. The statistic indicates that there is a problem with this item requiring further analysis (discussed below). In addition, the low ID scores of items #1 (0.12) and #6 (0.37) show that they provide minimal information about the learners' reading abilities. Alderson et al. (1995) say that item writers are usually satisfied with ID of 0.4 and above. The ID of items #2, #4, and #5 (0.5) indicate that these items discriminated well between high- and low-performing learners.

Only three items in the test have good ID. This result shows that the test was not useful in terms of providing information about the learners' reading ability due to poorly

distinguishing high-performing and low-performing students. In addition, the negative ID of item #3, even though it has a relatively high IF, raises questions about the effectiveness of this item. Why did more low-performing students answer it correctly than high-performing students? Item #1 also raises questions, although its low ID was still positive, indicating that a few more high-performing students answered it correctly than did low-performing students.

Looking more closely at these MCQ items, the distractors of items #1, #2, and #3 are investigated. The following table (Table 2) is a display for item #1 DE.

Table 2. Item #1 DE

	8 High-Performing Students	8 Low-Performing Students
a	4	3
b	1	1
c*	3	2
d	0	2

Note that (*) indicates the correct answer.

In item #1's DE we can see that while distractors "b" and "d" functioned effectively, attracting low-performing students, distractor "a" attracted four of the high-performing students and three of the low-performing students, This distractor attracting high-performing students explains why the IF was so low (0.12) for this item. This result may show that there is a problem with this distractor, such as poor wording of the distractor or poor text. Overall, item #1 is a difficult item with one inefficient distractor. The following table (Table 3) is a display for item #2 DE.

Table 3. Item #2 DE

	8 High-Performing Students	8 Low-Performing Students
a*	5	1
b	0	1
c	3	2
d	0	3

Note that (*) indicates the correct answer.

In item #2 DE we can see that the key answer "a" attracted more of the high-performing students. Distractors "b" and "d" attracted low-performing students, which shows that these distractors are functioning well. Although distractor "c" attracted more high-performing students than low ones, this item functions well overall.

Table 4 is a display for item #3 DE.

Table 4. Item #3 DE

	8 High-Performing Students	8 Low-Performing Students
a*	5	6
b	2	1
c	0	0
d	1	1

Note that (*) indicates the correct answer.

In item #3 DE we can see that the correct answer "a" attracted more of the low-performing students than high-performing students, where as the distractor "b" attracted more of the high students than low students. In addition, distractors "c" and "d" did not attract more low-performing than high-performing students, which may show that there is a problem with the efficiency of the three distractors. In addition, the answer key "a" attracted more low-performing students than high ones, which is the biggest problem with this item. The low efficiency of the distractors "b," "c," and "d" may explain why the ID of this item is -0.12. Overall, item #3 is a moderately easy item with inefficient distractors that fails to distinguish between high- and low-performing students.

WLB school test item analysis. The second test is also from a boys' school, coded as WLB. According to the intermediate stage manager, this school has a low quality administration. It has a large number of students in all grades (543 students).

The reading comprehension test in school WLB consists of six items: three MCQs and three short answer questions. The Flesch-Kincaid readability level of the reading passage in this school test is 4.1. (See Appendix F.) The number of words in the passage is 168 words. The reading topic is about the moon, which could be in the learners' background knowledge. The MCQs were about the best title of the passage, word meaning, and word reference. The short answer questions were about specific information in the passage. See Appendix G for the scores of the students in each item in

the reading test, each student's total test score, and the scores of the highest seven students and the scores of the lowest seven students.

Table 5 shows school WLB reading test IF and ID.

Table 5. School WLB IF and ID of Reading Test Items

Item	1	2	3	4	5	6
IF	0.95	0.86	0.44	0.42	0.62	0.90
ID	0.14	0.14	0.43	0.57	0.71	0.14

It can be seen in Table 5 that items #1, #2, and #6 IF scores are 0.95, 0.86, and 0.90, respectively, which shows that they are easy items. These high IF results provide little information about the learners' reading abilities. Coinciding with the high IF, it can be seen that the same items' ID is low; all have ID of 0.14, which means that these items may need reconsideration and may affect the test's usefulness. In general, extremely easy items do not distinguish learners with low reading skill from learners with high reading skill because almost all answers are answered correctly.

However, IF for items #2, #3, #4, and #5 may provide information about the learners since they range between 0.42 and 0.86. The same items' ID, except item #2, ranges between 0.42 and 0.71 and thus may provide useful information about the learners. Having a few easy questions in a test is acceptable, but in this case, three of the six reading comprehension questions were very easy, with corresponding low ID. Thus, this assessment of the learners' reading comprehension may not be very effective.

Looking more closely at these MCQ items, the distractors of items #1, #2, and #3 are investigated. The following table (Table 6) is a display for item #1 DE.

Table 6. Item #1 DE

	7 High-Performing Students	7 Low-Performing Students
a	0	0
b	0	1
c	0	0
d*	7	6

Note that (*) indicates the correct answer.

In item #1 DE it can be seen that distractors "a" and "c" did not attract any of the low-performing students. The low efficiency of the distractors "a" and "c" partly explains why the IF of this item is very high – 0.95. It also explains why the ID is low – 0.14. Overall, item #1 is an easy item with two inefficient distractors.

Table 7 is a display for item #2 DE.

Table 7. Item #2 DE

	7 High-Performing Students	7 Low-Performing Students
a*	6	5
b	1	1
c	0	1
d	0	0

Note that (*) indicates the correct answer.

In item #2 DE we can see that the distractors "b" and "d" did not attract more low-performing than high-performing students. This result may show that there is a problem with these distractors. The low efficiency of the distractors, particularly of distractor "d" which no one chose, may indicate why the IF of the question is relatively high – 0.86 – and ID is low – 0.14. Overall, item #2 is an easy item with one inefficient distractor.

Table 8 is a display for item #3 DE.

Table 8. Item #3 DE

	7 High-Performing Students	7 Low-Performing Students
a	0	1
b	0	1
c	0	2
d*	5	2

Note: Two from the high-performing and one from the low-performing learners did not answer item #3.

In item #3 DE we can see that the all the distractors attracted low-performing students. The efficiency of the distractors explains why this item has IF of 0.44 and ID of

0.43, indicating the item functions effectively and provides useful information about the learners.

GSB school test item analysis. The third test is also from a boys' school. It is coded as GSB in this paper. According to the intermediate stage manager, this school has a good quality administration and a small number of students in all grades (322 students).

The reading comprehension test in school GSB consists of six items: three MCQs and three short answer questions. The Flesch-Kincaid readability level of the reading passage in this school test is 7.5. (See Appendix H.) The number of words in the passage is 188 words. The reading topic is about the importance of sources for electricity, which could be in the learners' background knowledge. The MCQs were about the best title of the passage, word meaning, and word reference. The short answer questions were about specific information in the passage. (See Appendix I for the scores of the students in each item in the reading test, and each student's total test score.)

The following table (Table 9) shows school GSB reading test IF and ID.

Table 9. School GSB IF and ID of the Reading Test Items

Item	1	2	3	4	5	6
IF	1	0.95	0.21	0.78	0.34	0.47
ID	0.0	0.12	0.25	0.37	0.62	0.75

We can see in Table 9 that items #1 and #2 IF is 1.0 and 0.95, respectively, which indicates that these items are very easy for these students. This IF indicates the items may provide little information about the learners' reading abilities. Coinciding with the high IF, it can be seen that the same items' ID is low. Item #1 ID is 0.0, and item #2 ID is 0.12, which means that these items may need reconsideration and their ineffectiveness may affect the test's usefulness.

In terms of ID, item #5 is acceptable with ID of 0.62, which means that it successfully discriminates between low- and high-performing students. In contrast, item #3 ID is low, 0.25, which may show that it does not provide much information about the learners.

Looking more closely at these MCQ items, the distractors of items #1, #2 and #3 are investigated. The following table (Table 10) is a display for item #1 DE.

Table 10. Item #1 DE

	8 High-Performing Students	8 Low-Performing Students
a	0	0
b*	8	8
c	0	0
d	0	0

Note that (*) indicates the correct answer.

In item #1 DE we can see that the distractors "a," "c," and "d" did not attract any of the low-performing students, which shows that these distractors are not efficient enough. Making the distractors more attractive to low-performing students would increase the effectiveness of this item. The low DE in the three distractors explains the extremely high IF of 1.0 and low ID of 0.0. An item with IF of 1.0 indicates that all students answered it correctly, so the item makes no distinction between high reading skill learners and low reading skill learners. Overall, item #1 is a very easy item with inefficient distractors.

Table 11 is a display for item #2 DE.

Table 11. Item #2 DE

	8 High-Performing Students	8 Low-Performing Students
a	0	0
b	0	1
c*	8	7
d	0	0

Note that (*) indicates the correct answer.

In item #2 DE we can see that the distractors "a" and "d" did not attract any of the low-performing students, which shows that these distractors are not efficient enough. The low DE in the two distractors explains the high IF and low ID in this item. Overall, item #2 is an easy item with two inefficient distractors.

The following table (Table 12) is a display for item #3 DE.

Table 12. Item #3 DE

	8 High-Performing Students	8 Low-Performing Students
a*	3	1
b	0	1
c	0	1
d	5	5

Note that (*) indicates the correct answer.

In item #3 DE we can see that distractor "d" attracted most of the low- and high-performing students. It attracted the same number of both, which shows that this distractor is not efficient enough, and this raises the question of lack of clarity in the answer key (correct answer) and/or distractor "d." Perhaps "d" is a correct answer as well. This distractor may be the cause of the low IF and ID in this item. Overall, item #3 is a difficult item with one extremely inefficient distractor.

WSB school test item analysis. The fourth test is also from a boys' school, coded as WSB. According to the intermediate stage manager, this school has a weak quality administration. It has a small number of students in all grades (310 students).

The reading comprehension test in school WSB consists of six items: three MCQs and three short answer questions. The Flesch-Kincaid readability level of the reading passage in this school test is 5.2 (see Appendix J). The number of words in the passage is 249 words. The reading was a factual text about the Mayan Indians, which could not be within the learners' background knowledge. The MCQ items #1, #2, and #3 were about the best title for the text, word meaning, and word reference. The short answer question items #4, #5, and #6 were about specific information in the passage. (See Appendix K for the scores of the students in each item in the reading test, and each student's total test's score.)

Table 13 shows school WSB reading test IF and ID.

Table 13. School WSB IF and ID of the Reading Test Items

Item	1	2	3	4	5	6
IF	1.0	1.0	1.0	0.19	0.23	0.33
ID	0.0	0.0	0.0	0.14	0.57	0.57

It can be seen in Table 13 that items #1, #2, and #3 have IF of 1.0 and ID of 0.0, indicating these are very items that will not provide useful information about the students' reading ability. The IF and ID of these items show that they need reconsideration because they indicate that the test items were not suitable in terms of level of difficulty (too easy). The IF for #5 and #6 is acceptable, indicating these items may provide information about the learners. IF for #4 shows that it is a rather difficult item with a low ID, indicating it was difficult for most students.

ID for both #5 and #6 is 0.57, which shows that they discriminate successfully between high- and low-performing learners. However, item #4 ID is very low, 0.14, which indicates that this item may need reconsideration in order to improve the usefulness of the test.

It can be seen from the IF and ID of the MCQs and short answer questions of this test that the items provide minimum information about the learners' reading ability.

Looking more closely at these MCQ items, the distractors of items #1, #2, and #3 are investigated.

Table 14 is a display for item #1 DE.

Table 14. Item #1 DE

	7 High-Performing Students	7 Low-Performing Students
a	0	0
b*	7	7
c	0	0
d	0	0

Note that (*) indicates the correct answer.

In item #1 DE it can be seen that distractors "a," "c," and "d" did not attract any of the low-performing students, which shows that there is a problem with these distractors. The low efficiency of the distractors explains why the IF of the questions is very high, 1.0, and the ID is very low, 0.0. Overall, item #1 is a very easy item with inefficient distractors.

Table 15 is a display for item #2 DE.

Table 15. Item #2 DE

	7 High-Performing Students	7 Low-Performing Students
a	0	0
b	0	0
c	0	0
d*	7	7

Note that (*) indicates the correct answer.

In item #2 DE it can be seen that distractors "a," "b," and "c" did not attract any of the low-performing students, which shows that there is a problem with these distractors. As above, the low efficiency of the distractors may indicate why the IF of the questions is very high, 1.0, and the ID is very low, 0.0. Overall, item #2 is a very easy item with inefficient distractors.

Table 16 is a display for item #3 DE.

Table 16. Item #3 DE

	7 High-Performing Students	7 Low-Performing Students
a	0	0
b	0	0
c *	7	7
d	0	0

Note that (*) indicates the correct answer.

In item #3 DE it can be seen that distractors "a," "b," and "d" did not attract any of the low-performing students, which shows that there is a problem with these distractors. The low efficiency of the distractors explains why the IF of the questions is

very high, 1.0, and the ID is very low, 0.0. Overall, item #3 is a very easy item with inefficient distractors.

GLG school test item analysis. The fifth test is from a girls’ school, coded as GLG. It has a good quality administration according to the intermediate stage manager and has a large number of students in all grades (739 students).

The reading comprehension test in school GLG consists of six items: three MCQs and three short answer questions. The Flesch-Kincaid readability level of the reading passage in this school test is 3.4 (see Appendix L). The number of words in the passage is 161 words. The reading topic is a short narrative text, which could be within the learners’ background knowledge. The MCQ items #1, #2, and #3 were about the best title for the text, word meaning, and word reference. The short answer questions items #4, #5, and #6 were about specific information in the passage. (See Appendix M for the scores of the students in each item in the reading test each student’s total test score.)

Table 17 shows school GLG reading test IF and ID.

Table 17. School GLG IF and ID of the Reading Test Items

Item	1	2	3	4	5	6
IF	0.04	0.62	0.54	0.2	0.08	0.0
ID	-0.12	0.37	0.25	0.37	0.25	0.0

It can be seen in Table 17 that item #1 has IF of 0.04, which shows that it is a very difficult item, and the ID – -0.12 – indicates more low-performing students got this item right than high-performing students. Item #5 has IF of 0.08, which shows that it is a difficult item, which is probably the reason for the low ID. Both high- and low-performing learners may not be able to answer a very difficult item, and thus the item would not discriminate well between low- and high-performing learners. Item #6 has IF of 0.0, which shows that it is extremely difficult item, one which no one got right, thus making it impossible to distinguish high-performing students from low-performing (ID of 0.0).

Although the IF of items #2, #3, and #4 is acceptable, the ID of these items is low. These items may not provide useful information about the students’ reading ability

because they do not distinguish effectively between low- and high- performing students. The IF and ID of these items show that, in order to improve the usefulness of the test, they need reconsideration.

Looking more closely at these MCQ items, the distractors of items #1, #2, and #3 are investigated. (Table 18) is a display for item #1 DE.

Table 18. Item #1 DE

	8 High-Performing Students	8 Low-Performing Students
a*	0	1
b	1	4
c	6	2
d	1	1

Note that (*) indicates the correct answer.

In item #1 DE it can be seen that distractor "c" attracted more high-performing than low-scoring learners, which may indicate that the distractor is not efficient. It could also mean that the question is not clear, or the key is not clear. In addition, distractor "d" attracted the same number of low- and high-performing students. The lack of distractor efficiency for "c" and "d" contributed to the low IF, 0.04, and ID, -0.125.

Table 19 is a display for item #2 DE.

Table 19. Item #2 DE

	8 High-Performing Students	8 Low-Performing Students
a	1	2
b	0	1
c*	7	4
d	0	1

Note that (*) indicates the correct answer.

In item #2 DE it can be seen that all the distractors attracted low-performing students, which indicates that these distractors are efficient. However, four of the low-performing students managed to answer the question correctly, which is reflected in the moderately high IF and somewhat low ID of 0.37.

Table 20 is a display for item #3 DE.

Table 20. Item #3 DE

	8 High-Performing Students	8 Low-Performing Students
a	2	1
b*	5	3
c	1	3
d	0	0

Note: One from the low-performing learners did not answer this question.

In item #3 DE we can see that the distractor "d" did not attract any students, which may show that there is a problem with the efficiency of this distractor. In addition, distractor "a" attracted more high-performing learners than low-performing ones. This low efficiency in "a" and "d" may have caused the low ID for this item, 0.25. Overall, item #3 is a moderately easy item with two inefficient distractors.

WLG school test item analysis. The sixth test is from a girls' school, coded as WLG. According to the intermediate stage manager, this school has a weak quality administration. It has a large number of students in all grades (619 students).

The reading comprehension test in school WLB consists of six items: three MCQs and three short answer questions. The Flesch-Kincaid readability level of the reading passage in this school test is 5.2 (see Appendix N). The number of words in the passage is 127 words. The reading text is a factual type of text about why some people go to the jungle, which could be within the learners' background knowledge. The MCQ items #1, #2, and #3 were about the best title for the text, word meaning, and word reference. The short answer items #4, #5, and #6 were about specific information in the passage. (See Appendix O for the scores of the students in each item in the reading test, and each student's total test's score.)

Table 21 shows school WLG reading test IF and ID.

Table 21. School WLG IF and ID of the Reading Test Items

Item No.	1	2	3	4	5	6
IF	0.78	0.52	0.39	0.60	0.26	0.82
ID	0.37	0.12	0.75	0.37	0.62	0.25

It can be seen in Table 21 that the items' IF is acceptable. The ID of items #1, #2, #4, and #6 is low. This result indicates that items #1, #2, #4, and #6 may not provide useful information because of the low ID.

In general the test items were varied; some were more difficult, and some were easy to answer. However, all the items, except items #3 and #5, did not discriminate successfully between low- and high-performing learners. Thus, in order to improve the usefulness of this test, the low ID items need reconsideration.

Looking more closely at these MCQ items, the distractors of items #1, #2, and #3 are investigated. The following table (Table 22) is a display for item #1 DE.

Table 22. Item #1 DE

	8 High-Performing Students	8 Low-Performing Students
a	0	1
b	0	2
c	0	0
d*	8	5

Note that (*) indicates the correct answer.

In item #1 DE it can be seen that distractor "c" did not attract any of the low-performing students. This result shows that there is a problem with this distractor. The other distractors functioned efficiently. Overall, item #1 is a moderately easy item with one inefficient distractor.

Table 23 is a display for item #2 DE.

Table 23. Item #2 DE

	8 High-Performing Students	8 Low-Performing Students
a*	5	4
b	0	1
c	2	2
d	1	1

Note that (*) indicates the correct answer.

In item #2 DE we can see that distractors "c" and "d" attracted the same number of high- and low-performing students. This shows that there is a problem with these distractors. This low efficiency of the distractors "c" and "d," combined with the moderately high IF, explains the low ID for the item 0.12. Overall, item #2 is a moderately easy item with two inefficient distractors.

Table 24 is a display for item #3 DE.

Table 24. Item #3 DE

	8 High-Performing Students	8 Low-Performing Students
a	0	3
b*	6	0
c	2	4
d	0	1

Note that (*) indicates the correct answer.

In item #3 DE we can see that all the distractors attracted low-performing students, which indicates that they are efficient. This result, combined with the moderately high IF, explains the high ID of 0.75.

GSG school test item analysis. The seventh test is also from a girls' school, coded as GSG. It has a good quality administration and has a small number of students in all grades (405 students).

The reading comprehension test in school GSG consists of six items: three MCQs and three short answer questions. The Flesch-Kincaid readability level of the reading passage in this school test is 3.7. (See Appendix P.) The number of words in the passage

is 148 words. The reading is a short narrative about how the narrator's car gets stolen, which could be within the learners' background knowledge. MCQ items #1, #2, and #3 were about the best title for the text, word meaning, and word reference. The short answer items #4, #5, and #6 were about specific information in the passage. (See Appendix Q for the scores of the students in each item in the reading test, and each student's total test score.)

Table 25 shows school GSG reading test IF and ID.

Table 25. School GSG IF and ID of the Reading Test Items

Item No.	1	2	3	4	5	6
IF	0.09	0.71	0.09	0.04	0.23	0.33
ID	0.28	0.28	0.28	0.14	0.0	0.42

It can be seen in Table 25 that items #1, #3, and #4 IF is low. It can also be seen that the ID of all the items is low. This result means that the test items may not provide useful information about the learners.

In general, except item #2, all the test items were extremely difficult items, which only a few, if any, students got right. By their very nature, extremely difficult items do not effectively discriminate. It is acceptable to have a few difficult items and a few easy items in a test, but too many of either makes the test overall less effective.

Looking more closely at these MCQ items, the distractors of items #1, #2, and #3 are investigated. The following table (Table 26) is a display for item #1 DE.

Table 26. Item #1 DE

	7 High-Performing Students	7 Low-Performing Students
a	3	2
b*	2	0
c	1	2
d	1	3

Note that (*) indicates the correct answer.

In item #1 DE it can be seen that the distractor "a" attracted more high-performing students than low-performing ones. This may show that there is a problem with this

distractor's efficiency and/or with the key. However, distractors "c" and "d" attracted more low-performing students, which shows that they were efficient. All the distractors attracted high-performing students, which explains why the IF of the questions is very low – 0.09. Overall, item #1 is a difficult item with one inefficient distractor.

Table 27 is a display for item #2 DE.

Table 27. Item #2 DE

	7 High-Performing Students	7 Low-Performing Students
a	0	1
b	0	1
c	0	0
d*	7	5

Note that (*) indicates the correct answer.

In item #2 DE we can see that distractor "c" did not attract any of the low-performing students. This result shows that there is a problem with this distractor's efficiency. The low efficiency of distractor "c" partly explains why the IF of the question is high – 0.71. However, it was simply an easy question. Perhaps the answer was too obvious from the reading text. Extremely easy test items by their very nature do not discriminate well. Overall, item #2 is an easy item with one inefficient distractor.

Table 28 is a display for item #3 DE.

Table 28. Item #3 DE

	7 High-Performing	7 Low-Performing Students
a*	2	0
b	4	6
c	1	0
d	0	1

Note that (*) indicates the correct answer.

In item #3 DE it can be seen that the distractor "c" attracted one high-performing student and did not attract any of the low-performing students, which may show that there is a problem with the efficiency of this distractor. Distractor "b" attracted more of the high- and low-performing students than did the correct answer, which raises the question

as to why. Was the answer “a” not clear? Was the question confusing? Was the question wording unclear? The effect of distractors “b” and “c” may explain the low IF, 0.09, and ID, 0.28. Overall, item #3 is a difficult item with one inefficient distractor.

WSG school test item analysis. The eighth test is also from a girls’ school, coded as WSG. It has a low quality administration and a small number of students in all grades (287 students).

The reading comprehension test in school WSG consists of six items: three MCQs and three short answer questions. The Flesch-Kincaid readability level of the reading passage in this school test is 10.3. (See Appendix R.) The number of words in the passage is 168 words. The reading is a short factual text about the importance of dictionaries, which could be within the learners’ background knowledge. The MCQ items #1, #2, and #3 were about the best title for the text, word meaning, and word reference. The short answer question items #4, #5, and #6 were about specific information in the passage. (See Appendix S for the scores of the students in each item in the reading test, and each student’s total test score.) Table 29 shows school WSG reading test IF and ID.

Table 29. School WSG IF and ID of the Reading Test Items

Item No.	1	2	3	4	5	6
IF	0.68	0.89	0.47	0.42	0.47	0.47
ID	0.16	0.33	0.16	0.33	0.5	0.0

It can be seen in Table 29 that all the tests’ items IF is acceptable, except item #2’s. However, the ID of items #1, #2, #3, #4, and #6 is low, which indicates that the items do not distinguish low- and high-performing students for some reason. This means that the test items may not provide useful information about the learners. These items need revision and reconsideration in order to improve the usefulness of this test.

Looking more closely at these MCQ items, the distractors of items #1, #2, and #3 are investigated. Table 18 is a display for item #1 DE.

Table 30. Item #1 DE

	6 High-Performing Students	6 Low-Performing Students
a	1	0
b*	5	4
c	0	1
d	0	1

Note that (*) indicates the correct answer.

In item #1 DE it can be seen that distractor "a" attracted one high-performing student and no low-performing students. This may show that there is a problem with this distractor's efficiency. However, distractors "c" and "d" attracted more low-performing students, which shows that they were efficient. Overall, item #1 is a moderately easy item with one inefficient distractor.

Table 31 is a display for item #2 DE.

Table 31. Item #2 DE

	6 High-Performing	6 Low-Performing Students
a	0	0
b	0	2
c*	6	4
d	0	0

Note that (*) indicates the correct answer.

In item #2 DE it can be seen that distractors "a" and "d" did not attract any of the low-performing students. This shows that there is a problem with these distractors' efficiency, resulting in a very easy item with two ineffective distractors.

The following table (Table 32) is a display for item #3 DE.

Table 32: Item #3 DE

	6 High-Performing Students	6 Low-Performing Students
a*	4	3
b	1	1
c	1	1
d	0	1

Note that (*) indicates the correct answer.

In item #3 DE we can see that the distractors "b" and "c" attracted the same number of high- and low-performing students, which may show that there is a problem with the efficiency of these distractors. Overall, the item is moderately easy with two somewhat ineffective distractors.

Test Analysis General Discussion

Table 33 shows the eight schools' school IF, ID, type of text, Flesch-Kincaid level, and the texts' word count.

Table 33. The Schools' IF, ID, Type of Text, Flesch-Kincaid, and Texts' Word Count

School	Item	1	2	3	4	5	6	Text Type	F.K.	Words
GLB	IF	0.33	0.39	0.70	0.46	0.66	0.21	Factual	5.8	194
	ID	0.12	0.5	-0.12	0.5	0.5	0.37			
WLB	IF	0.95	0.86	0.44	0.42	0.62	0.90	Factual	4.1	168
	ID	0.14	0.14	0.43	0.57	0.71	0.14			
GSB	IF	1.0	0.95	0.21	0.78	0.34	0.47	Factual	7.5	188
	ID	0.0	0.12	0.25	0.37	0.62	0.75			
WSB	IF	1.0	1.0	1.0	0.19	0.23	0.33	Factual	5.2	249
	ID	0.0	0.0	0.0	0.14	0.57	0.57			
GLG	IF	0.04	0.6	0.54	0.2	0.08	0.0	Narrative (story)	3.4	161
	ID	-0.12	0.37	0.25	0.37	0.25	0.0			
WLG	IF	0.78	0.52	0.39	0.60	0.26	0.82	Factual	5.2	127
	ID	0.37	0.12	0.75	0.37	0.62	0.25			
GSG	IF	0.09	0.71	0.09	0.04	0.23	0.33	Narrative (story)	3.7	148
	ID	0.28	0.28	0.28	0.14	0.0	0.42			
WSG	IF	0.68	0.89	0.47	0.42	0.47	0.47	Factual	10.3	168
	ID	0.16	0.33	0.16	0.33	0.5	0.0			

The MCQ questions in all the eight tests varied in terms of IF and ID. All of school WSB's MCQs were extremely easy. The other schools had one or two easy questions, with GLG and GSC having one or two very difficult questions. In addition, all of school WSB's MCQs did not discriminate between low- and high-performing students.

The other schools had one or two items that did not discriminate between low- and high-performing students, with GLB and GLG each having negative ID in one of their test items. This result might be because of the low DE in many items in the tests. Low DE might show that the item is a badly posed question. The lack of effective DE in many items in most of the eight tests may be a result of the lack of clear descriptions of how the distractors should be developed when writing MCQs or poor item writing training.

In addition, the short answer questions also varied in terms of IF and ID. The tests of schools WSB, GLG, and GSG somewhat, had very low IF for all three short answer questions. All the schools have one or two low ID items, with GLG, GSG, and WSG each having one item that has ID of 0.0, which also might be a result of lack of clarity in the test specifications about how to develop such items.

It can be seen in Table 33 that there was no clear distinction between schools with good administration and schools with weak administration. In addition, there was no distinction between girls' schools and boys' schools.

Analyzing Five Problematic Items

For the purposes of illustration, five test items are looked at more closely. These test items are item #3 in school GLB's test, items #1 and #6 in school GLG's test, item #4 in school GSG's test, and item #6 in school WSG's test.

1- Item #3 in school GLB's test is an MCQ. The question is thus:

The word they in second paragraph refers to:

- | | |
|-------------------|-----------|
| a) Tea and coffee | c) meals |
| b) most people | d) babies |

As can be seen in the test item analysis section, item #3 has ID of -0.12, which means that more of the low-performing students than high-performing ones answered this item correctly. This indicates that there is a problem in this item. The right answer for this question is "a) Tea and coffee." It is noticed that several factors may have contributed to why this item might have failed to discriminate between learners. The first factor is that there is a mistake in the prompt. The pronoun "they" is not in the second paragraph, as mentioned in the prompt (see Appendix D for the passage); it is in the first paragraph. This

mistake may have made some high-performing learners be unsure of the answer. The second factor is that the answer is written starting with a capital letter “Tea and coffee, while all the other distractors are written without an initial capital letter. This capitalization difference may have made some low-performing students choose this answer only because it was different from the other distractors.

In order to have more efficient items, I suggest revising the items thoroughly, and having a trusted teacher pilot before the test administration. In addition, the key answer (correct answer) should be capitalized in the same way as the distractors in order not to contain factors that attract the learners other than being the correct answer for the question. Furthermore, the characteristics of the key answer and the distractors should be described clearly in the test specifications.

2- Item #1 in school GLG’s test is an MCQ. The question is thus:

The best title of this passage could be -----

- a- The wise old man
- b- The lazy farmers
- c- Selling the farm
- d- Earning money

As can be seen in the test item analysis section, item #1 has IF of 0.04, which shows that it is a very difficult item, and more low-performing students got this item right than did high-performing students, which resulted in an ID of -0.12. This negative ID indicates that there is a problem with this item.

Looking at the reasons why this item failed to discriminate between learners, several factors are noticed which may have contributed to this result. The first factor is that the words in the correct answer, “The wise old man,” are not used in the text at all (see Appendix J for the passage), while the words in the three distractors are used in the text. This feature may have made the high-performing learners not choose it as an answer. In addition, distractor “c- Selling the farm” could be a reasonable answer based on the reading text content. The effect of these factors may explain why six high-performing learners chose “c”.

In order to have a more efficient item, the test should be revised thoroughly before administration to make sure that there are no reasonable answers in the distractors. In

addition, the characteristics of the key answer and the distractors should be described clearly in the test specifications. Furthermore, item writing training should be provided.

3- Item #6 in school GLG's test is a short answer question. The question is thus:

How many fields did they offer to the poor?

As seen in the test item analysis, item #6 has IF of 0.0, which shows that it is an extremely difficult item (no one answered it correctly). Therefore it did not discriminate between the high- and low-performing students at all. This result indicates that there might be a problem with this item. Looking at why this item failed to discriminate between learners, one factor may have contributed to the extremely low IF, 0.0. This factor is the reading sub-skills that the question targets, which are connecting ideas in the content and drawing inference from the reading content. The correct answer for this question is "They offered three fields." The answer is not explicitly stated in the text (see Appendix J for the text). It is mentioned in the first paragraph that the farmer's farm contains three fields, and it is mentioned in the last paragraph that the sons offered their father's farm to the poor. Therefore, the learners needed to connect these ideas together and draw inference to answer the question, which made the question difficult because it required a high level of thinking. The learners may not have mastered such skills, and it shows that they may have not been exposed to such kinds of questions.

Test writers should make sure that the learners have tackled the targeted sub-skill during the instruction period before writing an item that targets that sub-skill in a test. In addition, each test item's targeted sub-skill needs to be clearly addressed and explained in the test specifications.

4- Item #4 in school GSG's test is a short answer question. The question is thus:

What was wrong with the writer's car?

As seen in the test item analysis, item #6 has IF of 0.04, which shows that it is a difficult item. It has an ID of 0.14, which shows that it did not discriminate well between the high- and low-performing learners. This low IF and ID indicate that there might be a problem with this item. One factor is noticed that may have contributed to why this item failed to discriminate between learners well. This factor is the wording of the item. In the

text there are many problems with the narrator's car (see Appendix N for the text). It was stolen by someone, it was an old car, and there was a problem with the wheels. However, the scorers were targeting one answer only, which is that there was a problem with the wheels. Many students said that it was stolen, which the scorer did not accept as a correct answer. This wording of the question made many reasonable answers possible, yet only one was accepted by the scorer, resulting in the item having an ID of 0.14 and IF of 0.04.

Changing the question to another question targeting a more specific answer would address this issue, for example, "Why was the writer on his way to the garage?" Adding more detailed description and examples of the short answer items to the test specifications would make it less likely that the test writers would write such items.

5- Item #6 in school WSG's test is a short answer question. The question is thus:

What do some books have (in) its end?

As seen in the test item analysis, item #6 has IF of 0.47, which shows that it is not a very difficult item. However, it did not discriminate between the learners at all. This lack of item discrimination indicates that there might be a problem with this item. Looking at the reasons why this item failed to discriminate between learners, it can be noticed that one factor may have contributed to this problem. This factor is the way this item is scored. Five out of six high-performing learners and three of six low-performing students answered this question correctly. However, the item scorers deducted 0.25 mark from the item mark if there were any spelling or punctuation mistakes in the answer or if the learners provided more than the required answer. Two of the high-performing learners lost 0.25 mark from this item's mark because one of them did not start his answer with a capital letter, and the other wrote more than the required answer. Therefore, and because we based our ID and IF analysis on the items' full mark only, an ID of 0.0 for this item was the result.

The grammatical mistake in the item "its" to refer to books instead of "their" may also have contributed to making the item not discriminate between high- and low-performing learners.

In order to have a more efficient item, I suggest that clear rubrics should be used to mark the learners' answers, so that there will be consistency in scoring across all schools,

and the learners should be taught these rubrics in order not to make such mistakes. Furthermore, I suggest that the characteristics of the key answer should be described clearly in the test specifications.

Test Texts' Readability Level Analysis

The Flesch-Kincaid score gives indication of reading difficulty of a text. The eight schools' reading texts' Flesch-Kincaid scores were varied. The Flesch-Kincaid readability levels of the reading passages are GLB school, 5.8; WLB school, 4.1; GSB school, 7.5; WSB school, 5.2; GLG school, 3.4; WLG school, 5.2; GSG school, 3.7; and WSG school, 10.3. (See Table 33.) The readability differences between the reading texts in the tests are considerable, which indicates that the difficulty levels of the reading texts in the eighth grade first period tests differ from one school to another. This difference very likely is because the test specifications do not provide detailed specification about the readability level of the reading text. McNamara, Crossley, and Greenfield (2008) contend that accurately predicting the difficulty of reading texts for language learners is highly important for teachers, publishers, writers, and others to ensure that texts' readability level matches prospective readers' proficiency. Therefore, it is important to have a standardized readability level that matches the reading proficiency of the readers to help the test writers choose reading texts that are suitable to the learners' level, and as a result, have a more useful test.

Although the number of the words of the reading text is provided in the test specifications, that specified number (160-180 words), was not followed by some schools. GLB's number of words in the passage is 194 words; WSB's number of words in the passage is 249; and WLG's number of words in the passage is 127 words. This difference of the number of words in the texts may indicate that the actual implementation of the test specifications is not as adequate as it should be, which may also result in more differences between reading tests in schools. According to Hughes (2003), texts that candidates are expected to be able to deal with can be specified along a number of parameters, including the length of the text. The test writer needs to be aware of the suitable length (number of words) that the text in the test is to contain in order not to negatively affect the learners' comprehension of the reading passage. One of the

factors that can affect text comprehension is text length (Pearson & Camperell, as cited in Jalilehvand, 2012). Therefore, lack of consistency in text length affected consistency of measurement of students' reading comprehension across all schools.

It can also be seen that all the schools used the same types of questions that were explicitly stated in the test specifications. The MCQs were about the main idea of a paragraph in the text, best title of the text, word meaning, and word reference. The short answer questions were about specific information in the passage. Using only the types of questions that are clearly stated in the test specifications may indicate that the test writers may not know what sub-skills the "etc." in the test specifications refers to exactly. They may have preferred to stay on the safe side and not to assess any other reading constructs. They may not have been sure what exactly the other possible constructs are that the test writers could assess in this test.

In addition, it can be seen that the reading text topics are varied. Some texts are narrative, others are factual. This variation of the text topics also indicates that the teachers may not know what genre of reading they should provide in the first period reading test. It also raises questions such as these: Are they testing what they were teaching? What genre of the readings did the teachers teach? If they taught all genres of reading, then why are they testing one and not all? The test specifications could provide answers to these questions. However, there is nothing mentioned about this issue in the test specifications.

Furthermore, we can see that schools WSG, GSG, GLG, and WLB used fractions in scoring the short answer questions. Other schools did not; they considered the answers either right (giving full mark) or wrong (giving zero mark) for each item, which shows that the scoring is different from one school to another in terms of scoring short answer questions, which could be a result of the lack of scoring rubrics in the test specifications.

Davidson and Lynch (2002) say,

the purpose of a well written specification is to result in a document that if given to a group of similarly trained teachers working in a similarly constituted teaching context, will produce a set of test tasks that are similar in content and measurement characteristic. (p.15)

However, we can see from the above analysis that these tests have many differences in terms of content. For example, the content of the reading text in school GLB is a factual reading text, and the content of the reading text in school GSG is a short narrative text. Therefore, a more detailed test specifications document is required to help test writers produce similar tests in relation to the type of reading text.

Readability Level of the Reading Text in the Student Book and the Workbook

Eight of the first period reading lessons in the eighth grade student's book and workbook were analyzed using the Flesch-Kincaid readability scale to determine the level of the reading texts and to see to what extent the test reflects what is being taught. The following table (Table 34) shows the types of the first period reading texts, their word counts, and their readability levels.

Table 34: The Reading Texts in the Eighth Grade Student Book and the Workbook

No.	Reading Title	Text Type	F.K.	Words
1	Start Doing Athletics	Factual	4.7	282
2	Keeping Fit	Factual	6.3	260
3	The Olympic Games	Factual	7.6	280
4	How Water Lilies Began	Narrative	3.9	231
5	Abdullah's Diary	Narrative	3.8	147
6	Pearl Diving	Factual	7.4	318
7	The Nazca Lines	Dialogue	4.9	270
8	Deserts	Factual	6.0	231
9	Eman Al-Kout	Factual	9.1	95
10	A Star in Kuwait	Factual	8.9	109

It can be seen in Table 34 that the reading texts vary in terms of type (actual, narrative, and dialogue). This variation of the reading texts type shows that the learners have the chance to tackle reading passages in class that are similar in terms of type to the suggested types of reading in the test specifications, which increases the content validity of the test.

The texts are varied in terms of readability level. As seen, the texts' levels range from 3.8 to 9.1, which is a huge range of variation. In addition, with regard to readability, the reading texts have not been arranged in a logical order, from lower to higher level of readability as the learners' reading skill progresses during a certain period and during all the periods. It is strange to see this reading level variation in an EFL/ESL student book. Not having one readability level suitable for eighth grade learners may explain why there is no readability level suggested in the test specifications. Not having one certain readability level might be a result of the lack of a framework that describes the reading level of the learners of this stage.

ELSS, HODs, Teachers Questionnaire Analysis

Information about the learners and the test can be provided to the test writers, using different methods like circulars, training programs, meetings, books, or websites. To find out whether or not the information not provided in the test specifications were provided to teachers using any other method, three questionnaires were used (one for HODs, one for supervisors, and one for teachers).

A total of 16 teachers, 25 HODs, and 6 supervisors participated in this survey. The three questionnaires contained questions to determine the following:

1. the test specifications yielding tests related to the specified syllabus
2. test construction procedures
3. validation procedures used to insure test validity
4. procedures used to insure test reliability
5. eighth grade learner level of reading

Each of these issues is discussed separately.

The Test Specifications Yielding Tests Related to the Specified Syllabus

Teachers, HODs and supervisors were asked three questions to find out what they think about how the first period reading test is related to the syllabus. The following table (Table 35) shows the three questions and their answers.

Table 35. The Relationship Between the Test and the Syllabus

#	Item	Participants	Never	Rarely	Sometimes	Often
5	The content of the 8 th grade reading comprehension test is related to what is taught in class.	Teachers	2	7	5	2
		HODs	9	9	4	3
		Supers	0	3	1	2
6	The 8 th grade reading comprehension test follows the assigned 8 th grade syllabus.	Teachers	4	4	5	3
		HODs	10	4	6	5
		Supers	2	1	1	2
7	Are you provided with a description of the content of the 8 th grade reading comprehension test with respect to:		No			Yes
Structure		Teachers	16			0
		HODs	20			5
		Supers	6			0
Vocabulary		Teachers	11			5
		HODs	14			11
		Supers	5			1
Timing		Teachers	5			11
		HODs	15			10
		Supers	5			1
Language functions		Teachers	16			0
		HODs	23			2
		Supers	6			0
Topics		Teachers	13			3
		HODs	19			6
		Supers	5			1

It can be seen in Table 35 that 9 teachers (56%), 18 HODs (72%) and 3 supervisors (50%) think that the content of the eighth grade reading comprehension test is never/ rarely related to what is taught in class. In addition, 8 teachers (50%) and 3 supervisors (50%), and 14 HODs (56%) said that the test never/ rarely follows the assigned syllabus. This large number (more than half) of the participants who think that the test is not related to what is taught/the syllabus, may result from the lack of clear explanation in the specifications for the suggested items.

It was seen above, in the test specifications analysis, that there are suggestions of what types of reading skills are to be tested. For example, MCQs are assigned to test word meaning/ reference words. However, there is no general statement of the skill to be tested. For example, there are no statements to describe the purpose of each suggested item, the reason for assessing the particular skill. Thus, the test specifications do not indicate the purpose for testing. In addition, there is no indication in the test specification of when to test a certain sub-skill (first, second, third, or fourth period). Therefore, it is very likely due to the lack of general statement of the reading skill to be tested, and why, that half of the participants think that there is never/ rarely a relation between the test and what is taught/the syllabus.

In addition, 16 teachers (100%), 20 HODs (80%) and 6 supervisors (100%) said that they are not provided with a description of the grammar to be assessed in the tests. Also, 11 teachers (68%), 23 HODs (92%) and 5 supervisors (83%) said that they are not provided with the language function. Although the test does not assess structure, vocabulary, speed-reading or language functions, it may be important to specify them in the test specifications in order for item writers not to confuse the learners with language that they are not familiar with. This lack of clarity may make test writers in different schools write different tests, which may affect the difficulty level of the reading text and the questions. It may also affect the tests' construct validity. Construct validity is involved in what the test scores mean, what they tell us about the test takers' ability, and whether the test in fact measures the targeted ability/abilities or not.

It can be seen that 13 teachers (81%), 19 HODs (76%), and 5 supervisors (83%) said that they are not provided with reading topics to be used in the test. This lack of specified reading topics may also make the test writers use reading texts that have different genres of writing and different topics, resulting in very different tests being produced.

The above findings show that the teachers, HODs, and supervisors indicate that the provided test specifications may not yield tests that are related to the specified syllabus. Thus content validity may be affected.

Test Construction Procedures Used

Since item writer qualification and training are crucial to test construction, the participants were asked about any information or guidance given test writers, in addition to the test specifications. Two teachers, and three HODs answered this question. Their answers were that only the number of words is provided.

They were also asked about the criteria used to appoint the eighth grade reading test writers. The following table (Table 36) shows the question and their answers.

Table 36. Test Construction Procedures

Item	Participants	Response/s
12 What criteria are used in the appointment of the 8 th grade reading comprehension item/test writers?		
None	Teachers	0
	HODs	4
	Supers	3
Should be 8 th grade teacher	Teachers	5
	HODs	17
	Supers	2
Should have specific teaching experience	Teachers	8
	HODs	12
	Supers	1
Should have item writing training	Teachers	3
	HODs	1
	Supers	0

In Table 36 participant responses were varied. For example, although 17 HODs (68%) responded that the test item writers should be eighth grade teachers, only 5 out of the 16 teachers (31%) and 2 out of the 6 supervisors (33%) said so. In addition, 8 teachers (50%) and 12 HODs (48%) said that the test writer should have specific teaching experience. However, only one supervisor (16%) said so. This variation of understanding the requirements for test writers may have resulted from the lack of clear instructions regarding this matter.

It can be seen from these results that the use of the test specifications is very likely not supported with additional documents/ training in order to provide more information to test writers, which may result in them producing different tests.

Validation Procedures Used to Insure Test Validity

The participants were asked questions that may help find indication of any validity assurance procedures/information provided to the test writers. The following table (Table 37) shows the questions and their responses.

Table 37. Validity Assurance Procedures

#	Item	Participants	Never	Rarely	Sometimes	Often
1	When I review*/administer **the 8 th grade reading comprehension test, I find that the reading text is at the students'	Teachers	0	1	8	7
		HODs	4	12	9	4
		Supers	0	0	4	2
2	I have a clear description of the 8 th grade learners reading level.	Teachers	8	2	4	2
		HODs	17	0	5	3
		Supers	4	0	0	2
	Item		No	Yes		
6	Are you given a description of the difficulty of the 8 th grade reading comprehension test?	Teachers	15	1		
		HODs	14	9		
		Supers	6	0		
9	Are you given a description of how to interpret 8 th grade students' reading comprehension ability based on the test results?	Teachers	12	4		
		HODs	22	2		
		Supers	6	0		
14	Are you given a description of the 8 th grade reading text difficulty (readability) level?	Teachers	15	1		
		HODs	19	6		
		Supers	6	0		
15	Are you provided with a description of reading sub-skills to be tested in end of 1 st period reading comprehension test?	Teachers	15	1		
		HODs	21	4		
		Supers	5	1		
16	Are special validity studies conducted on your 8 th grade reading comprehension test, to make sure that the test is measuring what it should measure, before administration?	Teacher	3	13		
		HODs	18	7		
		Supers	6	0		

* Used in the HODs and supervisors questionnaire. ** Used in the teachers questionnaire.

It can be seen in Table 37 that out of the 16 teachers, 25 HODs, and 6 supervisors who participated in the questionnaire, 16 HODs (64%) said that they rarely or never find that the test reading text in the test is at the learners' level. However, 15 teachers (93%) and 6 supervisors (100%) responded that they sometimes or often find that the reading text is at the students' level. This contradiction may have resulted from a lack of clear understanding of the actual reading level of the learners. It also could be a result of lack of use of a common framework and a lack of clear description of the learners' level of reading in the test specifications. In addition, this contradiction may be a result of a possible confusion from inconsistency in the textbook texts' readability level.

Confirming this observation, 10 teachers (62%), 17 HODs (68%), and 6 supervisors (100%) said that they have never had a clear description of the eighth grade reading level. The lack of clear description may cause difficulties in choosing reading texts for testing the learners, and HODs and supervisors may depend on their common sense instead of a solid, clear description. In addition, 15 teachers (93%), 14 HODs (56%), and 6 supervisors (100%) said that they are never provided with a description of the difficulty of the reading test text and items. The lack of description of the difficulty of the reading text and items may result in having tests that are varied in terms of difficulty levels. In addition, 15 teachers (93%), 19 HODs (76%) and 6 supervisors (100%) said that they are not provided with the required readability level for the test, which also may result in having different texts with different readability levels.

Thus, the teacher, HOD, and supervisor responses show that the test specifications are very likely not supported by any other source of information regarding the readability level of the reading text and the difficulty level of the text and the test items, which may result in having different reading tests in different schools.

Furthermore, 12 teachers (75%), 22 HODs (88%), and 6 supervisors (100%) said that they are not provided with a description of how to interpret the students' reading comprehension ability based on the test's results. The lack of a description of how to interpret the students' reading comprehension ability based on the test's results may make the marks that the learners achieve in the test be meaningless. A teacher may not be able

to determine what exactly the learners' reading problems are. This lack of description may make the test users not able to draw conclusions about the skill that is being measured; thus, the test scores may become less valid in terms of score interpretation, which also lessens positive washback.

Moreover, 15 teachers (93%), 19 HODs (76%) and the 6 supervisors (100%) said that they are not given a description of the eighth grade reading text readability level, and there was considerable variation in the textbooks' readability level. The item writers may depend only on their common sense and/or experience to judge the difficulty level of a reading passage. This variation in the textbook's readability level may explain why the readability levels in the tests are varied. Item writers may follow the inconsistent examples in the textbooks, which again may make them write different tests in terms of difficulty for students in the same stage.

Finally, 15 teachers (93%), 21 HODs (84%), and 5 supervisors (83%) said that they are not provided with a description of reading sub-skills to be tested in the first period. Lack of a description of the reading sub-skills to be tested may affect the construct validity of the test, because the HODs may include test items that test a reading construct that was not taught. However, analysis of the test specifications (see above) reveals that there are some specified items that address many reading sub-skills, such as recalling lexical item meaning, drawing inferences about the meaning of a lexical item in context, finding answers to questions answered in an explicitly stated manner in paraphrase, and drawing inference from the reading content. These contradictory findings between the test specifications and participant responses, may be caused by the lack of general description in the test specifications explaining why each item is suggested and what it tests.

Also, there are no indications in the test specifications of when (what period) to test these items. Thus, test content validity may be affected because the test may not target the specific skills that were taught during the instruction period.

Table 37 indicates that 18 HODs (72%) and 6 (100%) supervisors said that no special validity studies are conducted on the test to make sure that the test is measuring what it should measure, before administration. Yet some test items may be wrongly

worded or require more than the targeted sub-skill, and these issues can be detected when conducting test validity studies. Not doing a validity study may result in using some invalid test items, and as a result, the test may be less useful.

However, 13 teachers (81%) answered that there are special validity studies conducted on the test. The 13 teacher responses in comparison with the HODs' and the supervisors' responses may indicate that the 13 teachers who answered that there are special validity studies may not understand the question, or they may have another validity study in mind that the HODs and the supervisors may not know of.

In order to ensure content validity of a test, it is necessary to seek the advice of content experts. In addition, it is important to develop clear and detailed specifications for test items in different domains representative of the objectives of instruction (Henning, 1987). Item writers may pretest the test by asking some learners from other schools to do the test then analyzing the results. Not doing validity studies may make it not possible to ensure the eighth grade reading test content validity.

Procedures Used to Ensure Test Reliability

Because the environment in which a test is administered can affect its reliability, the 16 teachers, 25 HODs , and 6 supervisors were asked questions about procedures that may ensure test reliability, such as the instructions that are given to proctors. The following table (Table 38) shows the questions and their responses.

Table 38. Test Environment

	Item	Participant	Responses
17	What special instructions are given to proctors?	Teachers	2
		HODs	3
		Supers	1
	The test time limit	Teachers	13
		HODs	17
		Supers	5
	How to deal with disruptive behavior	Teachers	11
		HODs	15
		Supers	3
	Identify the student with a photo ID	Teachers	0
		HODs	5
		Supers	0

It can be seen in Table 38 that 13 teachers (81%), 17 HODs (68%), and 5 supervisors (83%) said that proctors are provided with instructions about test time limit. In addition, 11 teachers (68%), 15 HODs (60%), and 3 supervisors (50%) said that test proctors are given instructions regarding how to deal with disruptive behavior. These instructions may help in providing a reliable test environment. Only five supervisors mentioned requesting photo ID.

Rater reliability was also investigated since low inter-rater reliability and low intra- rater reliability may cause the tests to be less valid. Therefore, the participants were asked about the procedures that may affect test scoring reliability, including teaching experience and scoring experience. The following table (Table 39) shows the questions and the participants' answers.

Table 39. Scoring Reliability

Item		Participants	Responses	
18	What criteria are used in the appointment of the teachers who grade the test?			
	None	Teachers	2	
		HODs	5	
		Supers	4	
	Should be 8 th grade teacher	Teachers	10	
		HODs	13	
		Supers	0	
	Teaching experience	Teachers	2	
		HODs	12	
		Supers	1	
	Test scoring training	Teachers	3	
		HODs	7	
		Supers	1	
Item			No	Yes
10	Are you given detailed rubrics to mark answers in the 8 th grade reading comprehension test?	Teachers	12	4
		HODs	18	7
		Supers	6	0
19	Are any parts of the test graded by a group working together?	Teachers	3	13
		HODs	14	11
		Supers	1	5
20	Are any parts of the test marked by individuals only?	Teachers	13	3
		HODs	7	18
		Supers	0	6
21	What procedures, if any, are implemented to insure that the rater/raters are grading the test item fairly and equally during grading?			
	None	Teachers	0	
		HODs	3	
		Supers	0	
	Scorers are allowed to give marks individually before discussing with colleagues	Teachers	6	
		HODs	13	
		Supers	3	
	Live demonstration of the best scoring practice	Teachers	8	
		HODs	7	
		Supers	0	

It can be seen in Table 39 that 12 teachers (75%), 18 HODs (72%), and 6 supervisors (100%) said that they do not have detailed rubrics to mark the tests' answers. Lack of detailed rubrics may affect the test reliability and especially inter-rater and intra-

rater reliability. The lack of scoring rubrics explains why in the test items analysis, some schools were found using fractions to mark the short answer questions and some not.

The participants were asked about the criteria used in the appointment of the teachers who grade the test. In response 10 teachers (62%) and 13 HODs (52%) said that the raters should be eighth grade teachers. However, none of the supervisors said that they should be eighth grade teachers. In addition, 12 HODs (48%) said that raters should have teaching experience. However, only 2 teachers (12%) and 1 supervisor (16%) said so. These differences may have resulted from the lack of clear scoring instructions (related to selecting item scorers) in the test specifications, which may affect test scoring reliability.

It can also be seen that only 3 teachers (18%), 7 HODs (28%), and 1 supervisor (16%) answered that scoring training is provided. This lack of training may affect test scoring reliability and eventually affect the test's usefulness.

The participants were asked if there are any other procedures that insure rater reliability. In answer to that question, 9 teachers (56%), 17 HODs (68%) and 6 supervisors (100) said that a model answer is provided while scoring the test.

Different answers were found when the participants were asked about the way test items are scored individually/in groups. On one hand, 13 teachers (81%) and 5 (83%) supervisors said that there are parts of the test that are graded by a group of teachers working together. On the other hand, 14 HODs (56%) said that there are not. Also, 18 HODs (72%) and 6 supervisors (100%) said that there are parts of the test that are marked by individuals only. However, 13 teachers (81%) said that there are no parts of the test that are graded by individuals.

More contradiction showed in responses about the procedures, if any, that are implemented to insure that the rater/raters are grading the test items fairly and equally during grading. It can be seen in Table 39 that 13 HODs (52%) and 3 supervisors (50%) said that scorers are allowed to give marks individually before discussing with colleagues. However, only 6 teachers (37%) said so.

These different answers may have resulted from the lack of clear explanation of scoring procedures in the test specifications or lack of understanding of/paying attention

to the scoring procedures in test specifications. Different scoring methods may affect scoring reliability and eventually affect the tests' usefulness.

In addition, two questions were asked about statistical test reliability studies. The following table (Table 40) shows the questions and the participants' answers

Table 40. Statistical Item Analysis

Item		Participants Responses	
22	Which 8 th grade reading comprehension test items/questions are statistically analyzed after the test has been calculated?		
None		Teachers	11
		HODs	14
		Supers	5
MCQs		Teachers	3
		HODs	5
		Supers	1
Short answer questions		Teachers	2
		HODs	8
		Supers	0
23	How are the test item results analyzed?		
None are analyzed		Teachers	10
		HODs	17
		Supers	4
Item facility		Teachers	2
		HODs	3
		Supers	0
Item discrimination		Teachers	3
		HODs	5
		Supers	1
Distractor efficiency		Teachers	1
		HODs	2
		Supers	1

It can be seen from Table 40 that, when asked about item analysis after the test has been calculated, most answered that none is analyzed. Lack of item analysis may

show that there is a lack of follow up reliability and validity studies to assure the usefulness of the test.

Eighth Grade Learner’s Level of Reading

Seen above, there is no clear description of an eighth grade learners’ level of reading mentioned in the test specifications. Therefore, the teachers were asked about what they think an eighth grader level of reading might be. The question was “Which of the following describes an average eighth grade learner’s level of reading?” The provided levels were taken from the Common European Framework of Reference for Languages (*CEFR*) (2001). The following table (Table 41) shows the participants’ answers.

Table 41. Learners' Reading Level

	Items	Participants	Responses
B2	<i>Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms.</i>	Teachers	0
		HODs	3
		Supers	0
B1	<i>Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.</i>	Teachers	4
		HODs	7
		Supers	4
A2	<i>Can understand short, simple texts on familiar matters of a concrete type, which consist of high frequency everyday or job-related language.</i>	Teachers	11
		HODs	19
		Supers	4
A1	<i>Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.</i>	Teachers	10
		HODs	7
		Supers	1
A1	<i>Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.</i>	Teachers	12
		HODs	12
		Supers	0

It can be seen from Table 41 that the answers about eighth grade students' level are varied. Varied answers might be natural because the learners differ in terms of reading level; some are better than others. However, we can see that the teachers and HODs answers were concentrated around levels A1 and A2. In contrast, the supervisors' answers were concentrated around levels A1+ and B1. This difference between teachers and HODs on one hand and the supervisors on the other hand may be because the teachers and the HODs are in daily contact with the learners, but the supervisors are not. These different opinions about the learners' reading levels may result in the test writers choosing different text levels in the test. If the test specifications do not provide the test writers with a clear description of the reading text level of reading, differing tests will very likely result.

CHAPTER FIVE: CONCLUSIONS

The purpose of this research was to find out how useful the eighth grade reading comprehension tests specifications are, particularly in terms of developing the current eighth grade reading comprehension test. An additional issue was what improvements, if any, the current eighth grade reading comprehension test specifications require to help test item writers produce more useful tests. This chapter addresses the answers for the two research questions in order.

Research Question One

1. How effective are the current test specifications for the eighth grade first period reading comprehension test?

The purpose of well written test specifications is to provide indication as to how to produce a set of test tasks that are similar in content and measurement characteristics. After analyzing the test specifications and eight tests, it is evident that the eighth grade first period reading comprehension test specifications may not help in producing a set of test tasks or items that are similar in form and content across all schools. This unfortunate state of affairs is because there was no clearly specified reading construct and when to test each construct, no indication of when to use any of the specified types of readings, no specified readability level, no scoring rubrics, and no clear relation between what was taught in class and what is tested.

It can be seen from the eighth grade reading comprehension test specifications analysis that the targeted reading constructs are not clearly specified, both in terms of what skills to include and when (first, second, third or fourth period) to test each construct.

It was also found from the test specifications analysis that there are many types of reading suggested. However, there is no indication of when (first, second, third or fourth period) to use each type of reading.

Furthermore, it was found from the eighth grade reading test specifications analysis, and the participants' responses to the questionnaire that the test writers are not

provided with a description of the grammar, language function, and vocabulary used in the reading tests. It may be important to specify them in the test specifications in order not to confuse the learners with language that they are not familiar with. Failure to provide this information may also affect the construct validity of the tests.

There are other factors that the test specifications document lacks. It does not include a clear description of the purpose(s) of the test. In addition, the document does not include a plan for evaluating the qualities of usefulness. This lack of detailed description of how to evaluate the qualities of test usefulness may make the item writers produce tests that are less useful.

Overall, the test specifications for the Kuwait reading comprehension test were incomplete and lacked clarity. Therefore they are not effective enough to help similarly trained teachers working in a similarly constituted teaching context, produce a set of test tasks that are similar in content and measurement characteristic.

Research Question Two

2. How effective is the implementation of the test specifications in the development of the current eighth grade first period reading comprehension tests in Al-Jahra?

Because the provided test specifications do not help item writers write useful tests, the implementation of the test specifications in the development of the current eighth grade first period reading comprehension tests in Al-Jahra has many problems. These problems are related to test usefulness (reliability, validity, test impact, authenticity, interactiveness, and practicality).

Poorly written test items, items that have many reasonable answers, or items that are ambiguous may cause test unreliability. The test item analysis shows that the tests have many problems that affect their reliability in assessing the learners' reading ability. The MCQs questions in all eight tests varied and have many problems in terms of IF and ID. Many of the MCQs were IF and ID problems were resulted from the low DE in many items in the tests. Low DE might show that the item is a badly posed question. The lack of effective DE in many items in most of the eight tests may be a result of the lack of

clear descriptions in the test specifications of how should the distractors be developed when writing MCQs or lack of item writing training. In addition, the short answer questions also varied in terms of IF and ID. This also might be a result of lack of clarity of how to develop such items in the test specifications and training.

Test reliability is also affected by the accuracy and consistency of scores of the same test. The test analysis results show that the schools used different methods in scoring the short answer questions. This finding is consolidated by the finding in the questionnaires results analysis and the test specifications analysis, which show that there is a lack of detailed rubrics and a lack of clear instructions and training (related to selecting item scorers) that may affect the test reliability and especially inter-rater and intra-rater reliability and eventually affect the test's usefulness.

In the eight schools test analysis, it was found that all the tests tested only the reading sub-skills that were clearly stated in the reading test specifications. This lack of specific test construct may affect the test construct validity. Narrowing the construct in the test specifications may narrow the curriculum to only the stated construct to be assessed. In addition, test analysis results showed that the tests have many differences in terms of content. Thus, test content validity may be affected because the test may not target the specific skills that were taught during the instruction period.

In a related matter, the variation of the reading texts in the tests also indicates that the teachers may not know what genre of reading they should provide in the first period reading test. It also raises questions like are they testing what they were teaching? What genre of the readings did the teachers teach? If they taught all genres of reading, then why are they testing one and not all? The test specifications could provide answers to these questions. However, there is nothing mentioned about it in the test specifications.

In addition, the test specifications do not contain rubrics for scoring the test items. The questionnaire results also show that the test item scorers are not provided with any rubrics. The lack of rubrics may make item scorers score items differently, which was found in the schools' tests analysis. This lack of scoring rubrics negatively affects the test reliability.

Furthermore, the test specifications and the questionnaires results show that the teachers are not provided with a description of how to interpret the students' reading comprehension ability based on the test's results. The lack of a description of how to interpret the students' reading comprehension ability based on the test's results, may make the marks that the learners achieve in the test be meaningless. A teacher may not be able to determine what exactly the learners' reading problems are. This lack of description may make the test users not able to draw conclusions about the skill that is being measured; thus, the test scores may become less valid.

In terms of readability level, the eight schools' reading test texts' Flesch-Kincaid scores were varied, and this variation in the test texts very likely was because the test specifications do not provide detailed specification about the readability level of the reading text. This finding is supported by the questionnaire results which indicate a lack of clear understanding of the actual reading level of the learners. In addition, lack of reading level understanding may be a result of confusion from inconsistency in the textbook texts' readability level. It is important to have a standardized readability level that matches the reading proficiency of the readers. Doing so may help the test writers choose reading texts that are suitable to the learners' level, and as a result, have a more useful test.

It is also seen from the eight schools tests analysis that the test writers provided different types of texts: some short stories others factual passages. In addition, the survey results show that the participants are not provided with any specified topics for the reading tests. The lack of clarity of the text type in the test specifications made the test writers provide different text types for the same period test in different schools and thus produce extremely different tests.

Although the number of the words of the reading text is provided in the test specifications, the specified number (160-180 words) was not followed by some schools. Lack of consistency in text length could affect measurement of students' reading comprehension across all schools. This shows that the implementation of the test specification is not as effective as it should be. Test items writers may not be adhering to the test specifications.

In terms of test impact, based on the questionnaire results, which show that there is no relationship between what is taught in class and what is tested, and the results of the tests analysis that show that the texts do not exactly reflect the texts in the eighth grade student's book and workbook. Thus it can be seen that the test may have no/ little positive washback effect on teaching and learning reading. In addition, as mentioned above, the lack of a description of how to interpret the test's results, may make the test users not able to draw conclusions about the skill that is being measured. Thus no/little test social impact is achieved.

In terms of test authenticity, it can be seen that the eighth grade reading test cannot claim authenticity in a test task which is because the tasks (the MCQs and short answer questions) are not likely to be enacted in the real world. In addition, the test specifications lack any instructions related to test authenticity such as authentic texts.

In terms of interactiveness, it can be said that in order to answer the questions, the test takers need to use their linguistic knowledge in addition to their strategic competence. Therefore, the tests are interactive in terms of interacting with the test (particularly the reading text) but not with people.

Nothing that would negatively affect test practicality was noticed in the test specifications analysis and the eighth grade schools tests. On the contrary, the test analysis shows that the tests are one page test that contains only three MCQs and three short answer questions, which indicates that the tests do not require resources for implementing, developing, or using the test that exceed the resources available.

To conclude, this research found that the current test specifications for the eighth grade reading comprehension tests are not as effective as they should be. That is, if this test specification document was given to a group of similarly trained teachers working in a similarly constituted teaching context, they would not produce a set of test tasks that are similar in content and measurement characteristics. In addition, implementation of the test specifications (as evidenced in the eight schools' tests) was inconsistent in following what directions actually were provided in the specifications.

Implications and Suggestions for Improvement

It is seen that the test specifications contain the specification number, title of specification and related specifications. However, in order to produce more useful reading comprehension tests for the eighth grade, there is a need for a more detailed general description that describes the purpose of each suggested item and the reason for assessing the particular skill. In addition, a more detailed description of the prompt attributes is required that clearly specifies all the targeted sub-skills to be tested, instead of including "...etc.". The test specifications should indicate when to test a certain sub-skill (first, second, third or fourth period) and when to use each type of reading. The test specifications should also indicate the readability level of the text, using an appropriate readability assessment tool. In addition, the reading text readability level in the book needs to be unified using a readability level that suits eighth grade students' reading ability. The test specifications should also indicate the type reading for each instructional period, and the type of reading should be related to what has been taught in class. Moreover, the test specifications should contain more detailed sample items that reflect the specification of each test item type.

In addition to information about what is included in test items, the test specifications document should contain more detailed description of response attributes about of how the learners are going to provide answers in response to the prompt, or what would constitute a failure or success. Also, it should contain criteria for evaluating or rating the responses, especially the short answer questions. For example, it should indicate what would be done if a learner provided a correct answer but with some spelling mistakes, or what if a learner provided an incomplete answer.

Having very detailed test specifications alone would not automatically guarantee a useful test. Test writers need to be trained on test developing and item writing and test scorers need to be trained on test items scoring. Also the test should be pretested, if feasible, and analyzed in terms of ID, IF, and DE before actual administration.

Suggestions for Further Research

Further studies are required in the area of reading comprehension testing in Kuwait. There is a need for a study to investigate the effect of having different

readability levels for the reading texts in the students' books on teaching and on learners' reading comprehension. Another study is required to investigate and compare between 1st, 2nd, 3rd and 4th period eighth grade reading comprehension tests to see if there are any changes (in terms of item difficulty, targeted sub-skills, and text readability level) occurring as the learners progress in the course. In addition, a study is required to investigate the affect of test specifications on the eighth grade reading curriculum. Another study is required to investigate the eighth grade learners reading level in Kuwait public schools. One more study is required to investigate the social impact of the eighth grade reading test in Kuwait.

References

- Alderson, J. C. (2000). *Assessing reading*. Cambridge, New York: Cambridge University Press.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: New York: Cambridge University Press.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Reading, Mass: Addison-Wesley Pub. Co.
- Bachman, L. F., & Palmar, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
- Bailey, K. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. Boston: Heinle & Heinle.
- Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.
- Carlson, S., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences, 32*, 40-53. doi:10.1016/j.lindif.2014.03.003
- Chall, J. S., & Dale, E. (1995). *Readability revisited, the new Dale-Chall readability formula*. Cambridge, MA: Bookline Books.
- Clarke, P. J., Truelove, E., & Hulme, C. (2013). *Developing reading comprehension* (1). Somerset, GB: Wiley-Blackwell. Retrieved from <http://www.ebrary.com>
- Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (2001). Cambridge, UK; New York;: Cambridge University Press.
- Davidson, F., & Lynch, B. K. (1993). Criterion-referenced language test development. In A. Huhta, K. Sajavaara, & S. Takala (Eds.), *Language testing: New openings* (pp. 73–89). Jyväskylä, Finland: Institute for Educational Research, University of Jyväskylä.
- Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven: Yale University Press.
- Davies, A., Brown, A., Elder, C., Hill, C., Lumley, T., McNamara, T. (1999). *Dictionary of language testing*. Cambridge; New York: Cambridge University Press.
- Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly, 3*(4), 499-545.
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York; London: Routledge.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221-

233.

- Goodrich, H. C. (1977). Distractor efficiency in foreign language testing. *TESOL Quarterly*, 11(1), 69-78.
- Greenfield, G. (1999). Classic readability formulas in an EFL context: Are they valid for Japanese speakers? Unpublished doctoral dissertation, Temple University, Philadelphia, PA, United States. (University Microfilms No. 99-38670).
- Haynes, M., & Carr, T. (1990) Writing system background and second language reading: A component skills analysis of English reading by native speaker-readers of Chinese. In T. Carr & B. Levy (Eds.), *Reading and its development: Component skills approaches* (pp.375-421). San Diego: Academic Press.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, UK; New York: Cambridge University Press.
- Hippensteel, S. P. (2015). Assessing the readability of geoscience textbooks, laboratory manuals, and supplemental materials. *Journal of College Science Teaching*, 44(6), 24.
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Boston, Mass: Heine & Heine Publishers.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, UK;New York;: Cambridge University Press.
- Jalilehvand, M. (2012). *The effects of text length and picture on reading comprehension of Iranian EFL students*. *Asian Social Science*, 8(3), 329. doi:10.5539/ass.v8n3p329
- Kincaid, P., Fishburne, P., Rogers, L., & Chissom, S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Counand Flesch Reading Ease enlisted Research Branch 8-75 Formula) for Navy Personnel, Report. Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246-276. doi:10.1191/0265532202lt230oa
- Luo, K. (2015). Validity considerations in designing a writing test. *Studies in Literature and Language*, 10(5), 19-21. doi:10.3968/6957
- McNamara, D. S., Crossley, S. A., & Greenfield, J. (2008). *Assessing text readability using cognitively based indices*. *TESOL Quarterly*, 42(3), 475.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-256.
- Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in Language Testing* (pp. 1-3). London: NFER/Nelson.
- Mousavi, S. A. (2009). *An encyclopedic dictionary of language testing*. Tehran, Iran: Rahnama Press.

- Munby, W. (1978). *Communicative syllabus design*. Cambridge: CUP.
- Pearson, P. D. & Camperell, K. (1981). Comprehension of text structures. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews*. Newark, Delaware: International Reading Association.
- Saville, N. (2012). *Quality management in test production and administration*. In G, Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 395-412). London: Routledge.
- Siddiek, A. G. (2010). The impact of test content validity on language teaching and learning. *Asian Social Science*, 6(12), 133. doi:10.5539/ass.v6n12p133
- Walters, F. S. (2010). Cultivating assessment literacy: Standards evaluation through language-test specification reverse engineering. *Language Assessment Quarterly*, 7(4), 317-342. doi:10.1080/15434303.2010.516042
- Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
- Zamanian, M., & Heydari, P. (2012). Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1), 43.
- Zandi, H., Kaivanpanah, S., & Alavi, S. M. (2014). The effect of test specifications review on improving the quality of a test. *Iranian Journal of Language Teaching Research*, 2(1), 1-14.

Appendix A: The Eighth and Ninth Grade General English Language Test Specifications Document

State of Kuwait Ministry of Education ELT General Supervision School Year 2013 – 2014		دولة الكويت وزارة التربية التوجيه الفني العام للغة الإنجليزية العام الدراسي: 2013 - 2014
--	---	---

Distribution of marks & types of questions Target English (Grades 8 & 9)

Time Allowed: 1st & 3rd periods : (Two periods – One hour and a half)
 2nd & 4th periods: (Two hours)

No	Branch	Types of questions	First & Third Periods			Second & Fourth Periods & 2 nd Session		
			Item	Mark	Total	Item	Mark	Total
I	Vocabulary	A) Multiple choice (a , b , c & d) B) Gap filling	4	1	4	4	1	4
			2	1	2	4	1	4
			6		6	8		8
II	Grammar	A) Multiple choice (a , b,c & d) (Question tags / pronouns / articles prepositions / conjunctions /tenses...etc) B) Transformation(Do as required) (negation /questions / plurals /...etc)	4	½	2	4	½	2
			2	1	2	3	1	3
			6		4	7		5
III	Language Functions	Communicative situations	4	1	4	4	1½	6
			4		4	4		6
IV	Set Book	A)Productive questions of general nature (2 nd & 4 th periods : 3 out of 4) B) Literature Time (2 nd & 4 th periods : 1 out of 2)	2	1½	3	3	1	3
			1	1	1	1	2	2
			3		4	4		5

No	Branch	Types of questions	First & Third Periods			Second & Fourth Periods & 2 nd Session		
			Item	Mark	Total	Item	Mark	Total
V	Writing	Writing a report / story / e-mail of two paragraphs with the help of two main ideas with 5 words each Grade 8 (1 st & 3 rd periods (8 sentences) (2 nd & 4 th periods (10 sentences) Grade 9 (1 st & 3 rd periods (10 sentences) (2 nd & 4 th periods (12 sentences)	1 Mark to be considered for using pre-writing techniques (brainstorming, mind mapping, graphic organizers)			2 Marks to be considered for using pre-writing techniques (brainstorming, mind mapping, graphic organizers)		
			6			12		
VI	Reading Comprehension	Unseen (passage /e-mail/ letter / short story /dialogue) Grade (8) (160 – 180 words) Grade (9) (200 – 220 words) A- Multiple choice (a, b , c & d) (reference words / word meanings / main idea / title ...etc.) B- Productive questions (Questions should include inference / prediction / guessing / anticipation..etc.)	3	1	3	4	2	8
			3	1	3	3	2	6
			6		6	7		14
Total			25		30	30		50

N.B: * Enough practice should be given before testing.

1st & 3rd periods : 30 + 10 (Daily assessment) = 40 Marks

2nd & 4th periods : 50 + 10 (Daily assessment) = 60 Marks

End of year mark : 40 + 60 + 40 + 60 = 200 ÷ 2 = 100 marks


 وزارة
 التوجيه الفني العام للغة الإنجليزية

ELT Supervisor General
 Nouria Al-Sedra
Nouria Al-Sedra
 21 / 10 / 2013

Appendix B: Heads of Departments and English Language Supervisors Questionnaire

Heads of Departments and English Language Supervisors Questionnaire

Your position: **HOD** () **Supervisor** ()

Gender: **Male** () **Female** ()

How long have you worked in the above capacity?

How long have you worked in the field of English language teaching?

#	Item	Never	Rarely	Sometimes	Often
1	When I review the 8 th grade reading comprehension test, I find that the reading text is at than the students' level.				
2	I have a clear description of the 8 th grade learners reading level.				
3	I request a pre-test of the 8 th grade reading comprehension test before using it.				
4	I find marking discrepancies between teachers scoring the same item.				
5	The content of the 8 th grade reading comprehension test is related to what is taught in class.				
6	The 8 th grade reading comprehension_test follows the assigned 8 th grade syllabus.				

7- Are you given a description of the difficulty of the 8th grade reading comprehension test? **Yes** () **No** ()

8- Are you provided with a description of the content of the 8th grade reading comprehension test with respect to:

Structure **Yes** () **No** ()

Language functions **Yes** () **No** ()

Vocabulary **Yes** () **No** ()

Topics **Yes** () **No** ()

Timing **Yes** () **No** ()

Other (please specify)

9- Are you given a description of how to interpret 8th grade students' reading comprehension ability based on the test results? **Yes** () **No** ()

If yes, please explain.
.....

10- Are you given detailed rubrics to mark answers in the 8th grade reading comprehension test? **Yes** () **No** ()

11- What additional information or guidance for the 8th grade reading test distribution of marks is provided?
.....

12- What criteria are used in the appointment of the 8th grade reading comprehension item/test writers? (Check all that apply.) **None** () **Should be 8th grade teacher** () **Should have specified teaching experience** () **Should have had item writing training** () **Other (please specify)**

13- Are 8th grade reading comprehension items/test questions pre-tested? **Yes** () **No** ()

If yes, what test item scores are calculated and analyzed about the results? (Check all that apply.) **None** () **Item facility** () **Item discrimination** () **Distractor efficiency** () **Other (please specify):**

14- Are you given a description of the 8th grade reading text difficulty (readability) level? **Yes** () **No** ()

If yes, what is the reading text readability level for the 8th grade reading test?

15- Are you provided with a description of reading sub-skills to be tested in end of 1st period reading comprehension test? **Yes** () **No** ()

If yes, please indicate some of the reading sub-skills specified for the 8th grade reading test.
.....
.....

16- Are special validity studies conducted on your 8th grade reading comprehension test, to make sure that the test is measuring what it should measure, before administration? **Yes** () **No** ()

17- What special instructions are given to proctors? (Check all that apply.) **None** () **The test time limit** () **How to deal with disruptive behavior** () **Identify the student with a photo ID** () **Other (please specify):**

18- What criteria are used in the appointment of the teachers who grade the test? (Check all that apply.) **None** () **Should be 8th grade teacher** () **Teaching experience** () **Test scoring training** () **Other (please specify):**

19- Are any parts of the test graded by a group working together? **Yes** () **No** ()

20- Are any parts of the test graded by individuals only? **Yes () No ()**
 21- What procedures, if any, are implemented to insure that the rater/raters are grading the test item fairly and equally during grading? (Check all that apply.)
None () Scorers are allowed to give marks individually before discussing with colleagues ()
Live demonstration of the best scoring practice () Other (please specify).....

22- Which 8th grade reading comprehension test items/questions are statistically analyzed after the test has been calculated? **None () MCQs () Short answer questions ()**

23- How are the test item results analyzed ? **None are analyzed () Item facility ()**
Item discrimination () Distractor efficiency () Other (please specify):.....

24-What happens to the results of the test items analysis?

25- Which of the following describes an average 8th grade learner’s level of reading?
Please put (✓) beside the description of an average 8th grade learner

B 2	<i>Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms.</i>	
B 1	<i>Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.</i>	
A 2	<i>Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language.</i>	
A 2	<i>Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.</i>	
A 1	<i>Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.</i>	

**This reading framework is taken from the Common European Framework of Reference for Languages: Learning, teaching, assessment (2001). Cambridge, UK; New York; Cambridge University Press.*

Thank you for your valuable insights

Appendix C: Teachers' Questionnaire

Teachers' Questionnaire

Gender: Male () Female ()

How long have you been an English language teacher?

#	Item	Never	Rarely	Sometimes	Often
1	When I administer the 8 th grade reading comprehension test, I find that the reading text is at my students' level.				
2	I have a clear description of the 8 th grade learners reading level.				
3	I pre-test the 8 th grade reading comprehension test before using the test.				
4	I find marking discrepancies between teachers scoring the same item.				
5	The content of the 8 th grade reading comprehension test is related to what I teach in class.				
6	The 8 th grade reading comprehension test follows the assigned 8 th grade syllabus.				

7- Are you given a description of the difficulty of the 8th grade reading comprehension test? **Yes () No ()**

8- Are you given a description of the content of the 8th grade reading comprehension test with respect to:

Structure	Yes () No ()	Language functions	Yes () No ()
Vocabulary	Yes () No ()	Topics	Yes () No ()
Timing	Yes () No ()	Other (please specify)

9- Are you provided with a description of how to interpret 8th grade students' reading comprehension ability based on the test results? **Yes () No ()**

If yes, please explain.

.....

10- Are you given detailed rubrics to mark answers in the 8th grade reading comprehension test? **Yes () No ()**

11- What additional information or guidance to the 8th grade reading test distribution of marks, if any, is provided?

.....

12- What criteria are used in the appointment of the 8th grade reading comprehension item/test writers?
None () **Should be 8th grade teacher** () **Should have specified teaching experience** ()
Should have had item writing training () **Other (please specify)**

13- Are 8th grade reading comprehension items/test questions pre-tested? **Yes** () **No** ()

If yes, what test item scores are calculated and analyzed about the results? **None** () **Item facility** ()
Item discrimination () **Distractor efficiency** () **Other (please specify)**:.....

14- Are you given a description of the 8th grade reading text difficulty (readability) level? **Yes** () **No** ()

If yes, what is the reading text readability level for the 8th grade reading test?.....

15- Are you provided with a description of reading sub-skills to be tested in end of 1st period reading comprehension test? **Yes** () **No** ()

If yes, please indicate some of the reading subs-kills specified for the 8th grade reading test.

.....
.....

16- Are special validity studies conducted on your 8th grade reading comprehension test, to make sure that the test is measuring what it should measuring, before administration? **Yes** () **No** ()

17- What special instructions are given to test proctors? **None** ()

How to deal with disruptive behavior () **Identify the student with a photo ID** () **The time limits on the test** () **Other (please specify)**:

18- What criteria are used in the appointment of test graders? **None** () **Should be 8th grade teacher** ()
teaching experience () **Test scoring training** () **Other (please specify)**

19- Are any parts of the test marked by a group working together? **Yes** () **No** ()

20- Are any parts of the test marked by individuals only? **Yes** () **No** ()

21- What procedures, if any, are implemented to insure that the rater/raters are grading the test item fairly and equally during marking? **None** () **Live demonstration of the best scoring practice** () **Scorers are allowed to give their marks individually before they discuss them with colleagues** () **Other (please specify)**:.....

22- Which 8th grade reading comprehension test items/questions are statistically analyzed after the test has been calculated? **None** () **MCQs** () **Short answer questions** ()

23- How are the test items results analyzed ? **No analysis** () **Item facility** () **Item discrimination** () **Distractor efficiency** ()
Other (please specify):.....

24-What happens to the results of the test item analysis?

25- Which of the following describes an average 8th grade learner’s level of reading?
Please put (✓) beside the description of an average 8th grade learner

B 2	<i>Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms.</i>	
B 1	<i>Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.</i>	
A 2	<i>Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language.</i>	
	<i>Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.</i>	
A 1	<i>Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.</i>	

**This reading framework is taken from the Common European Framework of Reference for Languages: Learning, teaching, assessment (2001). Cambridge, UK; New York; Cambridge University Press.*

Thank you for your valuable insights

Appendix D: School GLB First Period Reading Comprehension Test

تابع/ اختبار اللغة الإنجليزية (الثامن) الفترة الدراسية الأولى 2016/2015 م الصفحة السادسة

VI-Reading

*** Read the following passage then answer the questions below:**

Usually people sleep between seven and eight hours a day although some people need less than this and some need more. One- year- old babies don't work, so they spend their time either sleeping or eating meals. Some adults on the other hand have trouble getting to sleep every night. Many people don't know why they are unable to sleep. Most people know that tea and coffee often make it difficult to go to sleep because **they** contain caffeine. Some medicines such as cold tablets also contain caffeine. Sleeping pills can help you fall asleep, but when you wake up the next morning, you'll feel tired.

You will sleep more easily if your bedroom is used only for sleeping. You shouldn't use your bedroom as an office, a TV room or an exercise room. You should also have a regular sleeping **routine** and not to go to bed until you're tired. Try to go to bed at the same time every night and get up at the same time every morning. Don't eat just before you go to bed, but try a warm glass of milk. If all this doesn't work, try counting sheep.

A) Choose the correct answer from a, b, c and d : (3 X 1 = 3)

- The main idea of the second paragraph is.....
 - the importance of sleeping
 - troubles people face during sleeping
 - rules for having a good night sleep
 - the bad effects of sleeping pills.
- The word: **routine** in the 2nd paragraph means :
 - something that happens every day
 - something that never happens.
 - something expected to happen
 - something that is likely to happen.
- The word **they** in second paragraph refers to:
 - Tea and coffee
 - most people
 - meals
 - babies

B)-Answer the following questions : (3 X 1 = 3)

4. What routine should a person follow before going to bed?

.....

5. How many hours of sleep does a person usually need?

.....

6. Why do you think sleeping is so important?

.....

Best wishes

Appendix E: School GLB First Period Reading Comprehension Test Results

Section	MCQs (1M)			Short Answer Questions (1M)			Reading Scores	Total Test Score
	Item No.	1	2	3	Item 4	Item 5		
High student 1	c*	a*	a*	1	1	1	6	29
High student 2	c*	a*	a*	1	1	0	5	26.5
High student 3	a	a*	d	1	1	1	4	23
High student 4	b	a*	a*	0	1	1	4	21.5
High student 5	a	a*	b	0	1	0	2	21.5
High student 6	a	c	a*	1	1	0	3	19.5
High student 7	c*	c	a*	1	0	0	3	18.5
High student 8	a	c	b	1	1	0	2	18.5
9	a	d	a*	1	1	1	4	17
10	c*	a*	a*	1	1	0	5	16.5
11	a	a*	a*	0	1	0	3	16
12	c*	a*	a*	0	1	0	4	15.5
13	a	b	b	0	0	0	0	15
14	a	d	a*	0	1	0	2	14.5
15	d	c	b	0	1	1	2	14
16	c*	d	a*	1	0	0	3	14
Low student 17	a	Missi ng	a*	0	1	0	2	12.5
Low student 18	b	a*	a*	1	0	0	3	12
Low student 19	c*	b	a*	0	1	0	3	11
Low student 20	a	d	a*	1	0	0	2	10
Low student 21	a	d	d	0	0	0	0	8.5
Low student 22	d	c	b	0	0	0	0	7.5
Low student 23	d	c	a*	0	1	0	2	8
Low student 24	c*	d	a*	0	0	0	2	8

(*) indicates the correct answers.

Appendix F: School WLB First Period Reading Comprehension Test

8th. Grade 1st. Period Exam2015-2016

II- Reading Comprehension (6 M)

Read the following passage then answer the questions below:

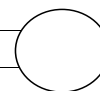


There is no doubt that science has told us so much about the moon that it is easy to know many things about *it*. The moon is not a friendly place. For mile after mile, there are many big mountains. Above, the sun and stars shine in a black sky. If you move away from the mountain shadows, it will mean moving from very low temperatures into great heat. These temperatures break rocks away from the surface of the mountains.

The Moon has been known since historic times. It is the second brightest object in the sky after the Sun. The moon is also a very silent place; you can't hear any sound there because sounds can only travel through air. From this distance, the earth is shining more than the stars. It looks like a big ball, colored blue, green and brown.

At last, if you want to *praise* someone, Are you going to tell him "you are as beautiful as the moon "? Think before you answer.

A: From a, b , c and d choose the correct answer :- (3x1=3)



20- The best title for the passage could be ----- .

a- A noisy place	b- Mountains	c- Shining stars	d- Facts about the moon
-------------------------	---------------------	-------------------------	--------------------------------

21- The pronoun (*it*) in the 1st paragraph refers to ----- .

a- moon	b- the earth	c- the sun	d- ball
----------------	---------------------	-------------------	----------------

22- The word "*praise*" in the 3rd paragraph means ----- .

a- ignores	b- to say bad words	c- to say hello	d- to say good words
-------------------	----------------------------	------------------------	-----------------------------

B. Answer the following questions: (3x1=3)



23 – Why is the moon a very silent place?

24 – What does moving away from the mountains shadows mean?

25 - Would you like to go to a journey to the moon? Why?

Appendix G: School WLB First Period Reading Comprehension Test Results

Section	MCQs (1M)			Short Answer Questions (1M)			Reading Scores	Total Test Score
Item No.	1	2	3	Item 4	Item 5	Item 6	6 Ms	50 Ms
High student 1	d*	a*	d*	1	1	1	6	27.5
High student 2	d*	a*	d*	1	1	1	6	26
High student 3	d*	a*	d*	1	1	1	6	24.5
High student 4	d*	a*	Mis sing	1	1	1	5	24.5
High student 5	d*	a*	Mis sing	1	1	1	5	24
High student 6	d*	b	d*	0	1	1	4	24
High student 7	d*	a*	d*	1	1	1	6	24
8	d*	a*	c	1	1	1	5	23
9	d*	a*	c	0	1	1	4	21.5
10	d*	a*	b	0	0	1	3	21
11	d*	a*	a	0	0	0.5	2.5	18
12	d*	a*	a	0	1	1	4	17.5
13	d*	a*	d*	0	0	1	4	17
14	d*	a*	c	0	1	1	4	16
Low student 15	d*	a*	c	1	1	1	5	15
Low student 16	b	a*	d*	0	0	1	3	15
Low student 17	d*	c	a	0	0	1	2	15
Low student 18	d*	a*	c	0	0	1	3	13.5
Low student 19	d*	a*	b	0	0	1	3	13.5
Low student 20	d*	a*	Mis sing	0	0	1	3	12.5
Low student 21	d*	b	d*	1	1	0	4	11.5

(*) indicates the correct answers.

Appendix H: School GSB First Period Reading Comprehension Test

الصفحة السادسة

الفترة الدراسية الاولى

الصف الثامن

تابع / اختبار اللغة الإنجليزية

VI. Reading Comprehension (6 marks)

6

*** Read the following passage then answer the questions below:**

Electricity is used everywhere in our lives. It lights up our homes, cooks our food, powers our electronic devices. Electricity from batteries keeps our cars running and makes our flashlights shine in the dark. Take a walk through your school or house and write down all the different devices and machines that use it. You'll be amazed at how many things we use every day that depend on electricity. Nearly, you will find that ninety per cent of the tools we use need it. But we need to know a little about electricity and how it is made.

There are many sources that Man can use to generate electricity. Wind power is used to **generate** it using big windmills. A good example for that is what happens in Holland, in Europe. Another source to generate electricity is using water power. Near waterfalls, we can build dams to store water and also big turbines to generate electricity. A good example for that is the High Dam in Egypt. **It** is used to store water and generate electricity. Electricity is one of the greatest and the most useful discoveries ever.

A. Choose the right answers from a, b, c and d: (3 x 1 = 3 marks)

20. The best title for the passage is

- a) Building dams b) Electricity c) Holland d) Windmills

21. The word **generate** in the second paragraph means

- a) break b) buy c) produce d) drive

22. The underlined "**It**" in the second paragraph refers to

- a) The High Dam b) water c) Europe d) Egypt

B. Answer the following questions: (3 x 1 = 3 marks)

23. What is electricity used for?

24. How do they generate electricity in Egypt?

25. Why do cars need electricity from the batteries?

Appendix I: School GSB First Period Reading Comprehension Test Results

Section	MCQs (1M)			Short Answer Questions (1M)			Reading Scores	Total Test Score
	Item No.	1	2	3	Item 4	Item 5		
High student 1	b*	c*	a*	1	1	1	6	29.5
High student 2	b*	c*	a*	1	1	1	6	29.5
High student 3	b*	c*	a*	0	0	0	3	23.5
High student 4	b*	c*	d	1	1	0	4	23.5
High student 5	b*	c*	d	1	1	1	5	22.5
High student 6	b*	c*	d	1	0	1	4	22.5
High student 7	b*	c*	d	1	1	1	5	20
High student 8	b*	c*	d	1	0	1	4	20
9	b*	c*	d	1	0	1	4	19.5
10	b*	c*	a*	1	0	1	5	17.5
11	b*	c*	d	1	1	1	5	17.5
12	b*	c*	d	1	1	1	5	17
13	b*	c*	d	1	0	0	3	17
14	b*	c*	d	1	0	1	4	16
15	b*	c*	d	1	1	0	4	15.5
Low student 16	b*	c*	d	1	0	0	3	13.5
Low student 17	b*	b	b	0	0	0	1	13
Low student 18	b*	c*	a*	1	0	0	4	12.5
Low student 19	b*	c*	d	0	0	0	2	12.5
Low student 20	b*	c*	d	1	0	0	3	12
Low student 21	b*	c*	d	1	0	0	3	11.5
Low student 22	b*	c*	c	0	0	0	2	11
Low student 23	b*	c*	d	0	0	0	3	9

(*) indicates the correct answers.

Appendix J: School WSB First Period Reading Comprehension Test

اختبار الفترة الدراسية الأولى / اللغة الإنجليزية / الصف الثامن / نوفمبر 2015 م / الصفحة السادسة

VI- [Reading Comprehension]

(6 m.)

Read the following passage, and then answer the questions below:

The Mayan Indians lived in Mexico for thousands of years before the Spanish arrived in the 1500s. The Maya were intelligent, rich people who did many things. They had farms, beautiful palaces, and cities with many buildings. The Mayan people knew a lot about nature and the world around them. This knowledge helped them to live a better life than most people of that time, because they could use it to make their lives more comfortable. Knowledge about tools and farming made their world easier and more productive.

In ancient Mexico there were many small clearings in the forest. In each, there was a village with fields of corn, beans, and other crops. To clear the land for farms, the Maya cut down trees with stone axes. They planted seeds by digging holes in the ground with pointed sticks. A farmer was able to **grow** crops that produced food for several people. But not every Maya had to be a farmer. Some were cloth makers or builders.

Although the cities that the Maya built were beautiful, and the people worked hard to build them, very few of the people lived in them. The other people lived in small villages in the forests. Their houses were much simpler than the houses in the cities. **They** lived in small huts with no windows. The walls were made of poles covered with dried mud, and the roof was made of grass or leaves. Most Maya lived a simple life close to nature.

A) Choose the correct answer from a, b, c and d : (3 x 1 = 3m)

1. The best title for the text could be
a) Knowledge b) The Mayan Indians c) the nature d) the farmer
2. The underlined word “ **grow** ” in the second paragraph means
a) know b) produce c) build d) plant
3. The underlined pronoun “ **They** ” in the third paragraph refers to
a) cities b) houses c) people d) forests

B) Answer the following questions: (3 x 1 = 3 m)

4- How did knowledge help the Mayan people live a better life?

.....

5- Why did the Mayan people cut trees?

.....

6- What effects did the nature have on the Maya?

.....

=====

[**BEST WISHES.**]

Appendix K: School WSB First Period Reading Comprehension Test Results

Section	MCQs (1M)			Short Answer Questions (1M)			Reading Scores	Total Test Score
	Item No.	1	2	3	Item 4	Item 5		
High student 1	b*	d*	c*	0	1	1	5	24.5
High student 2	b*	d*	c*	0	1	1	5	19.5
High student 3	b*	d*	c*	1	0	0	4	19.5
High student 4	b*	d*	c*	0	1	1	5	18
High student 5	b*	d*	c*	0	0	1	4	18
High student 6	b*	d*	c*	1	1	0	5	17.5
High student 7	b*	d*	c*	0	0	0	3	17
8	b*	d*	c*	0	0	0	3	17
9	b*	d*	c*	0	0	0	3	16.5
10	b*	d*	c*	0	0	1	4	16
11	b*	d*	c*	0	0	1	4	15
12	b*	d*	c*	1	1	0	5	14.5
13	b*	d*	c*	0	0	1	4	14.5
14	b*	d*	c*	0	0	0	3	14
Low student 15	b*	d*	c*	1	0	0	4	13.5
Low student 16	b*	d*	c*	0	0	0	3	13
Low student 17	b*	d*	c*	0	0	0	3	13
Low student 18	b*	d*	c*	0	0	0	3	13
Low student 19	b*	d*	c*	0	0	0	3	10
Low student 20	b*	d*	c*	0	0	0	3	10
Low student 21	b*	d*	c*	0	0	0	3	8

(*) indicates the correct answers.

Appendix L: School GLG First Period Reading Comprehension Test

العام الدراسي 2015 2016
اختبار الفترة الدراسية الأولى
المجال : اللغة الإنجليزية



دولة الكويت
وزارة التربية
منطقة الجهراء التعليمية
الزمن : ساعة ونصف

VI. Reading Comprehension (6 Marks)

Read the following passage , then answer the questions below : .

A poor farmer had three sons but the sons didn't want to work on the farm . It was a small farm with three fields . "The farm is too small for us ", they said to their father , "We must go to the town to earn our living .

"No , " he said . " I shall give all the land to the one who proves to be the best farmer .

Each son wanted the whole farm . They said to themselves , " I must do better than the rest . I must learn more about growing rice." Secretly, each son bought books on farming and read them at night .At the end of the second year , the amount of rice growing in the field was doubled . "You have a lot of money from the extra rice , " their father said ."buy one more field each year ."After many years , the sons became very rich . They could buy many farms . They offered their father's farm to the poor of their village .

A)-Choose the correct answers from a, b, c and d :- (3 x 1 = 3 Marks)

20 - The best title of this passage could be ----- .

a- The wise old man b- The lazy farmers c- Selling the farm d- Earning money

21 – The underlined word 'extra' in the second paragraph means

a- less than usual b- the same as usual c- more than usual d- not as usual

22- The underlined pronoun (They) in the 2nd paragraph refers to ----- .

a- many years b- the sons c- other farms d- the fields

B)-Answer the following questions : (3 x 2 =6 Marks)

23- What did each son want to have ?

.....

24- Why did they buy books on farming ?

.....

25- How many fields did they offer to the poor?

.....

Appendix M: School GLG First Period Reading Comprehension Test Results

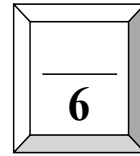
Section	MCQs (1M)			Short Answer Questions (1M)			Reading Scores	Total Test Score
	Item No.	1	2	3	Item 4	Item 5		
High student 1	c	c*	b*	1	1	0	4	26.75
High student 2	c	c*	b*	0	1	0	3	25.75
High student 3	c	a	c	0.75	0.5	0	1.25	25
High student 4	c	c*	b*	0	0.5	0	2.5	24.75
High student 5	c	c*	b*	1	0.5	0	3.5	24
High student 6	c	c*	a	1	0	0	2	22.5
High student 7	d	c*	b*	0	0	0	2	22.25
High student 8	b	c*	a	0	0.5	0	1.5	21.5
9	d	Mis	b*	0	0	0	1	21.25
10	b	c*	a	0	0	0	1	19.75
11	d	c*	a	1	0	0	1	18.25
12	c	d	b*	0	0	0	1	18.25
13	c	c*	b*	0	0	0	2	18
14	c	d	b*	0	0.5	0	1.5	17.5
15	d	b	d	1	0	0	1	17.5
16	d	c*	b*	0	0	0	2	17.25
17	c	d	a	0	0.5	0	0.5	17.25
Low student 18	b	c*	Mis	0	0	0	1	16.25
Low student 19	c	d	c	0	0.5	0	0.5	16.25
Low student 20	b	c*	b*	0	0	0	2	16
Low student 21	b	a	b*	0	0	0	1	16
Low student 22	c	b	a	0	0.5	0	0.5	15.5
Low student 23	b	a	c	0	0.5	0	0.5	15.25
Low student 24	a*	c*	b*	0	0	0	3	14.5
Low student 25	d	c*	c	0	0	0	1	14.5

(*) indicates the correct answers.

Appendix N: School WLG First Period Reading Comprehension Test

VI - READING COMPREHENSION (6 MARKS)

Read the following passage, then answer the questions:



Different people go to the jungle. Hunters go there to hunt animals and sell their skin or the valuable things they get from them. They also hunt animals and keep them alive to sell them to zoos. Some people go to the jungle because they are interested in collecting information about wild animals and birds. **They** also help some animals which are in danger of dying out. Some photographers are fond of taking photos of natural places and wild animals. Those people **risk** their lives. They may be attacked by a tiger, a lion or a bear. They may be bitten by snakes. They try to protect themselves with some ways. But still they face a lot of dangers. The beauty of nature attracts lovers of nature.

A) Choose the correct answer From a, b, c & d: (3x 1 = 3)

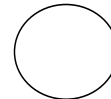
20- The best title of the passage could be.....

a- Attacking by bears

b- The great tiger

c- Taking photos

d- Wild animals in the jungle



21- The underlined word "**They**" in line (4) refers to.....

a- some people

b- hunters

c- birds

d-wild animals.

22-The word **risk** means

a- Be safe

b- danger

c- fond

d-build

B) Answer the following questions: (3x1=3)

23- Who enjoys going to the jungle?

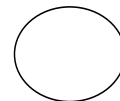
.....

24- What are the dangers which people can face in the jungle?

.....

25- Why do people go to jungles?

.....



Appendix O: School WLG First Period Reading Comprehension Test Results

Section	MCQs (1M)			Short Answer Questions (1M)			Reading Scores	Total Test Score
	Item No.	1	2	3	Item 4	Item 5	Item 6	6 Ms
High student 1	d*	a*	b*	1	1	1	6	30
High student 2	d*	a*	b*	1	1	1	6	29.75
High student 3	d*	a*	b*	1	1	1	6	28.5
High student 4	d*	a*	b*	1	0	1	5	27.75
High student 5	d*	a*	b*	1	1	1	6	27.5
High student 6	d*	d	c	0	0	1	2	25.5
High student 7	d*	c	b*	1	0	1	4	25
High student 8	d*	c	c	0	1	0	2	24.25
9	d*	b	a	1	0	1	3	24
10	d*	a*	c	1	0	1	4	23.5
11	d*	d	b*	0	0	1	3	23.5
12	a	a*	a	1	1	1	4	23.25
13	d*	a*	b*	1	0	1	5	22.25
14	b	d	a	1	0	1	2	22.25
15	d*	d	b*	0	0	1	3	21.5
Low student 16	d*	c	c	0	0	1	2	21
Low student 17	d*	c	c	0	0	1	2	20
Low student 18	d*	a*	a	1	0	0	3	19.75
Low student 19	d*	a*	c	1	0	1	4	19.25
Low student 20	b	a*	c	1	0	1	3	17.5
Low student 21	b	d	a	0	0	0	0	16.25
Low student 22	d*	a*	d	0	0	1	3	12
Low student 23	a	b	a	0	0	0	0	9.75

Appendix P: School GSG First Period Reading Comprehension Test

VI- Reading Comprehension (6 Marks)

Read the following story, then answer the questions below:-

My car was stolen last week . I left it in a side street .When I returned , it was gone .I was surprised , but I was **foolish** to leave it unlocked. I went to the police station. They asked me to describe the car . I mentioned that it was an old car but it was in a good condition expect there was one problem in the wheels . In fact, I was on my way to the garage to have the wheels adjusted when the car was stolen . I was certain that the thief would change **its** color.

The next day, the police told me they had found the car in the same street and said that someone had left a note: " If you want to kill yourself with this car, go ahead . I've got better things to do. " Anyway , the thief didn't laugh for long . He was soon arrested .

A)Choose the correct answer from a, b ,c and d : (3 x 1 = 3 MS)

1- The best title for this passage is

a- a modern car b- a funny theft c- a coloured car d-a wide street

2-The underlined pronoun "its" in line 6 refers to

a- the man b- the condition c- the street d-the car

3-The word "foolish" in line 2 means

a-stupid b- surprised c- lucky d-smart

B) Answer the following questions:- (3 x 1 = 3 MS)

4- What was wrong with the writer's car ?

.....

5-Where did the writer want to go before the car was stolen ?

.....

6-What happened at the end?

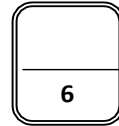
.....

Appendix Q: School GSG First Period Reading Comprehension Test Results

Section	MCQs (1M)			Short Answer Questions (1M)			Reading Scores	Total Test Score
Item No.	1	2	3	Item 4	Item 5	Item 6	6 Ms	30 Ms
High student 1	b*	d*	a*	0	0.75	1	4.75	25.5
High student 2	a	d*	a*	1	1	0.5	4.5	24
High student 3	d	d*	b	0	1	1	3	22
High student 4	b*	d*	b	0	0	0.5	2.5	21
High student 5	a	d*	c	0	0.75	1	2.75	20
High student 6	a	d*	b	0	0	1	2	18.5
High student 7	c	d*	b	0	0	0.5	1.5	18.5
8	c	b	b	0	1	1	2	17
9	a	b	b	0	0	0	0	16
10	a	d*	c	0	0	0	1	15.5
11	d	d*	b	0	0	0	1	14.5
12	a	b	b	0	0	0	0	14
13	d	d*	b	0	0	1	2	13.5
14	d	a	b	0.75	0	0	0.75	13.5
Low student 15	d	d*	b	0	0	0	1	12
Low student 16	a	d*	b	0	0	0	1	11.5
Low student 17	d	a	d	0	0	1	1	11
Low student 18	c	d*	b	0	0	0	1	11
Low student 19	d	b	b	0	1	0	1	11
Low student 20	c	d*	b	0	1	0	2	10
Low student 21	a	d*	b	0	0	0	1	9

Appendix R: School WSG First Period Reading Comprehension Test

IIV. Reading Comprehension (6 Marks)



Read the following passage then answer the questions below:

A dictionary is very important for all learners , it gives them the different meanings of a word .Most dictionaries help them to pronounce a word correctly by using pronunciation symbols. Moreover , they show the learners how to use the word by giving examples in sentences .You can always check your dictionary when you find a new word , However it isn't a good way of learning a language to think of using a dictionary all the time . It is better for you to try first to understand the idea of the passage and guess the meaning of the word . You can look at the spelling of the word and try to remember other similar words which may help you get the meaning .When you fail to get the right one, use the dictionary then and look it up.

Some books have a mini dictionary at its end . The computer dictionary is the latest dictionary . **It** has many advantages over older ones. Dictionaries are very helpful and **valuable** to learners.

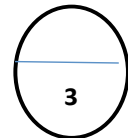
A) Choose the correct answer from a , b , c and d (3×1=3Ms) :

1. The best **title** for the passage is

- a) Mini dictionaries
- b) The importance of dictionaries
- c) Learning a foreign language
- d) Disadvantages of dictionaries

2.The pronoun "**It**" in the **second** paragraph **refers to**.....

- a) the spelling of the word
- b) the language
- c) the computer dictionary
- d) the meaning of the word

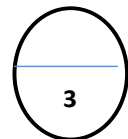


1. The word "**valuable**" in the **second** paragraph **means**

- a) useful
- b) not expensive
- c) un important
- d) similar

A) Answer the following questions (3×1=3Ms) :

4. How do most dictionaries help learners ?
.....



5.Why should we look at the spelling of the new word ?
.....

6.What do some books have in its end ?
.....



Appendix S: School WSG First Period Reading Comprehension Test Results

Section	MCQs (1M)			Short Answer Questions (1M)			Reading Scores	Total Test Score
Item No.	1	2	3	Item 4	Item 5	Item 6	6 Ms	30 Ms
High student 1	b*	c*	a*	1	0.5	1	5.5	28.5
High student 2	b*	c*	a*	0.75	1	1	5.75	25.5
High student 3	b*	c*	c	0.75	1	0	3.75	25
High student 4	b*	c*	a*	1	1	0.75	5.75	23.5
High student 5	a	c*	a*	0	1	0.75	3.75	23
High student 6	b*	c*	b	1	1	1	5	22.5
7	Missing	c*	a*	1	0	1	4	22.5
8	a	c*	d	1	1	0	3	21.5
9	a	c*	c	0.75	0.75	1	3.5	16.5
10	b*	c*	c	0.75	0	0	2.75	16.5
11	b*	c*	c	0.25	1	0	3.25	15
12	b*	c*	a*	1	0	1	5	14.5
13	b*	c*	c	1	0.75	0.75	4.5	12
Low student 14	b*	c*	a*	0.75	0	0	3.75	12
Low student 15	b*	c*	a*	1	0	1	5	11.5
Low student 16	b*	c*	a*	0	0	1	4	11
Low student 17	b*		c	0.75	1	1	4.5	10.5
Low student 18	c	b	d	0.5	0	0	0.5	4.5
Low student 19	d	b	b	0	1	0	1	11

Vita

Meshari Awad Barman was born on February 3, 1976, in Al-Jahra, Kuwait. He was educated in local public schools and graduated from Kuwait University, College of Education in 1999. He worked as English language teacher for five years and as a head of English language department for another five years. He is now an English language supervisor at the Ministry of Education in Kuwait. He is conducting his MA TESOL studies at the American University of Sharjah.