Association of Arab Universities Journal for Education and Psychology

Volume 21 | Issue 2 Article 1

2023

An evaluation of the English secondary school certificate test in Syria

ali suad hasan
faculty of education - damascus university, alisaudhasan@yahoo.com

An evaluation of the English secondary school certificate test in Syria

Dr. Ali Saud Hasan

Faculty of Education

Damascus University

Abstract

Language assessment is a continuous process that is used in almost every EFL classroom. Teachers constantly assess student performance when replying to certain questions, using a new vocabulary or structure in a sentence or providing a written work. Whether these forms of assessment are incidental or intended, they are certainly considered significant in assessing learners' performance.

The present research presents various types of language tests, illustrates what makes a good test including the reliability and validity of the test and examines how to evaluate tests. Specifically, this paper aims at evaluating one of the English language tests administered by the Ministry of Education in Syria. The test that is being evaluated is the 12th grade final exam issued by the Ministry of Education in 2016.

The most important findings of the study show that the test under review is practical in some aspects while not so in others. Test reliability is high as the items are clear, unambiguous, simple and not too long or require a lot of writing. In addition, the face validity of the test is high. However, the washback effect of the test is poor as students are not provided with the appropriate feedback.

1. Introduction:

Language assessment, which is an ongoing process, plays a significant role in the process of English language teaching and learning. Assessing students can be done by the following tools: homework, tests, final exams, interviews, oral reports and instructor observations. In addition, there are many reasons why classroom assessment, and tests as subset of it, are used and are of great importance. Through tests, teachers can gather information about the students' levels. They can also give and gain feedbacks, modify their teaching, and help students set future goals. Thus, assessment affects teachers, students, and the teaching process itself, and because of its vital role, assessment measures should be given consideration and continues evaluation.

This paper aims at evaluating one of the English language tests administered by the Ministry of Education in Syria. The test that is being evaluated is the 12th grade final exam issued by the Ministry of Education in 2016.

2. What is a test?

The definitions of 'test' vary according to different linguists. For example, for Bachman (1990) a test is "a measurement instrument designed to elicit a specific sample of an individual's behavior" (p. 20). For Brown (2004: 3) a test is 'a set of techniques, procedures, or items that requires performance on the part of the test-taker.' Furthermore, Carr (2011) states, "tests and other assessments are tools that are used for a particular reason" (P. 5). Tests are then used to evaluate a person's performance or aptitude in a given domain against a set of criteria. Testing is, therefore, "an important part of every teaching and learning experience " (Madsen, 1983, p. 3). Consequently, we should always bear in mind the purpose of the test as well as the potential test takers.

3. Test types

Every language test that exists is usually constructed with having a specific aim in mind. However, some test types are more flexible than others and can be used for a wider range of purposes. In educational contexts, language tests provide information for making a "wide variety of decisions" (Bachman, 1990, P. 70). Therefore, one way to categorize language tests is based on the type of decision to be made. Thus, according to Bachman (1990) we have:

- Selection, entrance, and readiness tests with regard to admission decisions.
- Placement and diagnostic tests with regard to identifying the appropriate instructional level or specific areas in which instruction is needed.
- Progress, achievement, or mastery tests with respect to how individuals should proceed through the program, or how well they are attaining the program's objectives (1990, P. 70).

Carr (2011) divides tests into two broad categories:

- Curriculum-Related Tests: tests that relate to a specific domain and are tightly tied to a teaching or learning curriculum. This category includes: admission tests, placement tests, diagnostic tests, progress tests and achievement tests that are employed to evaluate how well students have understood the materials of the course or accomplished its objectives.
- Other types of tests that are not related to a curriculum: "The most important of these are proficiency tests, which assess an examinees level of language ability, without respect to a particular curriculum." (Carr, 2011, p.6)

For Brown (2004), "An achievement test is related directly to classroom lessons, units, or even total curriculum. Achievement tests are (or should be) limited to particular material addressed in a curriculum within a particular time frame and are offered after a course has focused on the objectives in question." (p.47). On the other hand, for Carr (2011), "Achievement tests can be used for program exit or graduation decisions." (p.8)

In the light of the two previous quotes and taking into consideration test types that were explained earlier, we conclude that the Syrian 12th grade English final exam is considered an achievement test. It is seen as so because first, it is a curriculum-related test with items that are drawn from the content of instruction directly and test students with what they have covered during the year. Second, it is done within a particular time frame ,which is at the end of the year; so it is done after covering and focusing on the objectives in question. Third, it tests how well students have mastered the course content at a stage before graduating or exiting a program.

4. What makes a good test?

Henning (1987) provides a list of critical criteria that can be employed for the rating of the adequacy of any given test for any given purpose, most important of which are:

1- The purpose of the test: ***test validity***
   What to be tested determines the test's validity. Validity is related to the question, “is the content of the test consistent with the stated goal for which the test is being administrated?” (Henning, 1987, p.10)
2- Characteristics of the examinees: ***test difficulty***
   A test should never be designed without considering the skills and other qualities and abilities of the intended examinees. The test in general should be moderate: not too difficult nor too easy. In addition, a number of factors can determine tests’ appropriateness including the complexity of the reading passages and the familiarity of vocabulary.
3- Decision Accuracy: ***test reliability***
   All examinations are subject to errors. Even if some measurement error is unavoidable, it is nevertheless possible to quantify and significantly reduce its presence. “A reliable test is a test that has little measurement error and consistently rank-order the examinees in accordance with their comparative true abilities.” (Henning, 1987: 10)
4- Suitability of format and features: ***test applicability***
   Examinees should be familiar with the test format and the manner in which tasks are structured.
5- The developmental sample: ***test relevance***
   We have two kinds of relevance. Firstly, the test should be relevant to the test-takers for whom it was designed, that is it is crucial to take into account the test-takers' qualities. Secondly, relevance to the domain from which test items were drawn. By considering these aspects of relevance, we can assess a test's applicability to a certain sample and to a specific stated aim. (Henning, 987, 10-13).

For Bachman (1990) reliability and validity are "the essential qualities to the interpretation and use of measures of language abilities, and they are the primary qualities to be considered in using tests" (1990, p.24).

Reliability:

"It is a quality of test scores" (Bachman,1990, p. 24). In order for a score to be seen as perfectly reliable, it needs to be free from errors. However, many factors such as testing conditions, fatigue, and anxiety affect the test-takers' performance; students, thus, obtain scores that are varying from one test to another. Ultimately, "reliability has to do with the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context" (Bachman,1990, p. 24).

Validity:

It is the degree to which conclusions or decisions drawn from test results are meaningful, appropriate, and useful. "In order for a test score to be a meaningful indicator of a particular individual's ability, we must be sure it measures that ability and very little else " (Bachman, 1990, P. 25).

It should be noted that, according to Bachman, determining reliability and validity of a test is based on judgment and empirical research as neither of the aforementioned is a quality of tests themselves. Reliability is concerned with test scores, whereas validity is related to the uses of these scores. Furthermore, neither is absolute in the sense that, outside of the test itself, we can never achieve entirely error-free measurements (Bachman, 1990, p. 26). Carr (2011:19) stated, "No test is going to be perfect." The most important features are **usefulness** (Bachman and Palmer, 1996). Furthermore, Huerta- Macias,(2002) says, "An instrument is deemed to be *trustworthy* if it has *credibility* (i.e., truth value and *auditability* (i.e., consistency). In other words, does it measure what it is supposed to measure and would the instrument give the same results if replicated (Guba & Lincoln, 1981) (P.340).

Finally, Brown (2004) offers five principles of language assessment:

Practicality / Reliability / Validity / Authenticity / Washback.

- ✓ **Practicality:** "An effective and practical test is a test that
  A) is not excessively expensive
  B) stays within appropriate time constraints
  C) is relatively easy to administer
  D) has a scoring/evaluation procedure that is specific and time-efficient" (Brown, 2004, p. 19).

- ✓ **Reliability:** Reliability is related to consistency and dependability. "Whenever a test is administered, the test users would like some assurance that the results could be replicated if the same individuals were tested again under similar circumstances. This desired consistency (or reproducibility) of test scores is called reliability." (Crocker and Algina,1986, p. 105 as cited in Fluncher& Davidson, 2007). However, a test's unreliability can be attributed to a variety of circumstances:

  -Student-Related Reliability: **"**the students' personal state and circumstances such as illness, motivation, and tiredness" (Mousavi, 2002, p. 804).
  -Rater Reliability: when human error, subjectivity, and bias enter into the scoring process because of having two or more scorers or unclear scoring criteria, fatigue, and bias.
  -Test Administration Reliability: **"**the conditions in which the test is administered affect the results. This includes photocopying variations, the amount of light in different parts of the room, variations in temperature, and even the condition of desks and chairs" (Brown, 2004, p. 21).
  -Test Reliability**:** the test itself may cause unreliability as in the case of long and timed tests that might affect the students' performance.

It is worth mentioning a number of methods of computing reliability that try to estimate the extent to which an observed score is near to a true score:

1- "Test–retest: the same test is administered twice and the scores are compared.
2- Parallel forms: two forms of the same test are produced and the correlation between the scores on the two forms is taken as a measure of reliability.
3- Split half: within a single test, half of the items are taken to represent one form of the test and correlated with the items in the other half of the test" (Fluncher& Davidson, 2007, p. 105).

- ✓ **Validity:** "the extent to which inferences made from assessment are appropriate, meaningful, and useful in terms of the purpose of the assessment" (Gronlund, 1998, p. 226). The important question now is: how can we know if a test is valid? Brown (2004:22) mentions several different kinds of evidence that may be invoked to support validity:

- Content-Related Evidence**:** "it is often referred to as content validity." (Mousavi, 2002; Hughes, 2003). We can argue that a test has content validity when test-takers are required to perform the behavior that is being assessed.
- Criterion-Related Evidence: "the extent to which the criterion (implied objectives) of the test has actually been reached."
- Construct-Related Evidence: "A construct is any theory, hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions." (Brown,2004, P.25). In the field of assessment, construct validity asks, "Does this test actually tap into the theoretical construct as it has been defined?"
-Consequential Validity: "Consequential validity encompasses all the consequences of a test, including such considerations as its accuracy in measuring intended criteria, its impact on the preparation of test-takers, its effect on the learner, and the (intended and unintended) social consequences of a test's interpretation and use." (Brown,2004, P. 26).

- Face Validity: It is the extent to which "students view the assessment as fair, relevant, and useful for improving learning" (Gronlund, 1998, P. 210). In addition, "Face validity refers to the degree to which a test looks right, and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgment of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers" (Mousavi, 2002, P. 244). As stated by (Brown, 2004, P. 27) “Face validity will likely be high if learners encounter:

a) a well-constructed, expected format with familiar tasks
b) a test that is clearly doable within the allotted time limit
c) items that are clear and uncomplicated
d) directions that are crystal clear
e) tasks that relate to their course work (content validity)
f) a difficulty level that presents a reasonable challenge.”

- ✓ **Authenticity:** (Bachman and Palmer,1996: 23) define authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a target language task.” Essentially, when you say that a test task is authentic, you are saying that it is very much similar to "real world” tasks.

“In a test, authenticity may be present in the following ways:

a) The language in the test is as natural as possible.
b) Items are contextualized rather than isolated.
c) Topics are meaningful (relevant, interesting) for the learner.
d) Some thematic organization to items is provided, such as through a story line or episode.
e) Tasks represent, or closely approximate, real-world tasks.” (Brown, 2004,p. 28)

- ✓ **Washback**: It is "the effect of testing on teaching and learning" (Hughes, 2003, p. 1). One kind of washback is the information that "washes back" to students in the form of useful diagnoses of strengths and weaknesses. Washback also plays the role of windows of insight for the teacher into further work and progress. “Assessment places the needs of the students at the center of the teacher’s

planning"(Penaflorida, 2002, P.344). As Messick (1996) says, "washback refers to the extent to which the introduction and use of a test influences language teachers and learners to do things that they would not otherwise do that promote or inhibit language learning" (P.241).

5. Evaluation of the English secondary school certificate test

5.1. Test specifications:

"Test specifications are the most detailed level of test architecture." (Fulcher,2010, p.127). Alderson, Clapham and Wall,1995) call test specifications 'blueprints' (P.9).

When evaluating a test, one should ask the question of how well the test specifications reflect the purpose of the test and its objectives. Test specifications means that "a test should have a structure that follows logically from the lesson or unit you are testing" (Brown, 2004, p. 33). For Brown, a good test is one where "the objectives are incorporated into its structure that appropriately weights the various competencies being assessed" (p.42).

In the case of the 12th grade final exam, it is done for program exit with items that test the students with what they have learned during the year, so it serves its purpose as an achievement test. Besides, the items are derived from particular material addressed in a curriculum and are adapted or taken form the students' book and activity book, which means that the structure covers and follows, as much as possible, the lesson or unit being tested (in this case the curriculum). Finally, the objectives to test Reading, Vocabulary, Grammar, and Writing are met since the test design is divided into a number of sections that assess the aforementioned skills giving equal weight to each section.

Moving to the test structure and design, test specifications can offer a simple and practical outline of the test, and it comprises:

1. A broad outline of the test,
2. What skills you will test,
3. What the items will look like." (Brown, 2004, p. 50)

Brown (2004) also mentions that many tests have a design that:

1. divides the test into a number of sections (corresponding, perhaps, to the objectives that are being assessed),
2. offers students a variety of item types, and
3. gives an appropriate relative **weight** to each section (Brown, 2004, p. 33).

The specifications for the 12th grade final exam are as follows:

- The broad outline of the test:

Number of questions	11 questions, 44 question items
Score/mark	400 marks
Time	2 hours and a half
Type	Written, multiple-choice test

- The skills tested:

  - Reading (comprehension)
  - Vocabulary
  - Grammar
  - Writing

- What the items will look like:

Here we can see test applicability and test difficulty (Henning, 1987), since the actual test format and the tasks are familiar and are developed taking into consideration the abilities and other characteristics of the intended examinees. (See Appendix 1 for the format of the test items).

As for the design of the test, it is divided into a number of sections, within each one objective is being assessed. First, there is a section for assessing reading comprehension with two texts and four question items related to the texts. Second, we have a section for testing vocabulary that contains four questions. Third, for assessing grammar, there is a section on its own with six different questions. Finally, there is a section for assessing writing where students are asked to write a paragraph about a certain topic. Moreover, the test offers students a variety of

items types in each task such as (translate, match, short-answers, fill in the gap, write, and choose). Furthermore, the marks allocated to each section are relatively equal.

5.2. Items evaluation:

The process of evaluating assessment instruments is far too complex to be reduced to five principles. However, for the evaluation of the 12th grade final exam only these five principles (practicality, validity, authenticity, reliability, and washback) will be taken into consideration due to their centrality and importance.

5.2.1. ***The first section of the exam is a reading text*** *that aims at assessing reading comprehension:*

-***The first task is a Passage-based task*** that requires examinees to read a text in order to answer a number of questions related to it. (Carr,2011, P. 27). "This is a type of performance assessment in which students read or review textual materials and then respond to a series of open-ended questions" (O'malley & Pierce, 1995, p.13).

It is a short-answer wh-question and the mode of responding is a *short response* (Brown,2004, p. 51) where students need to provide short answers of no more than one sentence. It should be noted that Hingle &Linnington (2002) found that "open-ended questions evoked longer responses from more able students, but seemed to confound less able students." (P. 358).

-**Practicality:** This task is practical because: 1- administrative details are clearly established beforehand with clear instructions. 2- ease of administration with no material or equipment needed other than papers. 3- students have to produce short answers, so they can complete the task reasonably within the allocated period of time. 4- scoring system is feasible in the teacher's time frame since there is a group of teachers and they are provided with an

<u>First section:</u> 1.Read the following text then do the tasks below:

Animals live everywhere on earth, in every Terrain and in all climates. The place where an animal lives is called its habitat and most animals can only survive in one or two different habitats. For example, lizards live in hot climates and would die if they were moved to the Arctic. Some animals migrate between two habitats at different times of the year.

Kangaroos are marsupials, which means that when young are born, they are carried by their mothers in a pouch. There are 45 species of kangaroos and they live in every kind of habitats, from open plains to forests and Rocky deserts. Tree kangaroos live in remote and mountainous forest regions and have adapted to life in trees. Unlike other kangaroos, they cannot move very fast on the ground.

The jerboa lives all over Asia and Northern Africa. There are 25 different species of jerboa and they are specially adapted to live in extremely dry climates. They have very short front legs and long back legs which enable them to hop quickly over the ground to escape predators. **First task**

Answer the following questions:

1- What kind of climate do jerboa live in?
2- What is an "animal habitat"?
3- How are tree kangaroos different from other kangaroos?

answer key sheet related to the construct being assessed.

-**Reliability:** *Inter-rater reliability* is not that high as multiple teachers work on the evaluation of a task with a number of correct possible answers. As a result, inconsistency of what is acceptable for different scorers might be a problem.
Students reliability: The task is passage-based which means that without reading and comprehending the passage, students can't answer the questions. Students have to read and then provide a written response, which takes time. As a result, anxiety, loss of concentration, fear of timing, and fatigue will affect the students' performance and lower the students' reliability. However, *Test reliability is* high since the text is familiar to the students and is not long and the number of questions is limited to three, which is manageable.

-**Validity:** "The major source of validity in a classroom test is content validity: the extent to which the assessment requires students to perform tasks that were included in the previous classroom lessons" (Brown,2004, P.32). In this sense:
Content validity is present in the fact that the task is identical to previous tasks included in the student's book. On the other hand, Brown also stated, "The content validity of an existing classroom test should be apparent in how the objectives of the unit being tested are represented in the form of the content of items, clusters of items, and item types. Do you clearly perceive the performance of test-takers as reflective of the classroom objectives? If so, and you can argue this, content validity has probably been achieved" (2004, p. 33). Based on this quote, the task aims at assessing reading comprehension in the form of a text followed by wh-questions. However, answering the questions correctly does not reflect comprehension as students might copy the sentences from the text without understanding them. Then, answering does not reflect comprehension which is the objective being assessed. In this case, content validity is not evident.
Face validity is high since the task is structured to elicit the optimal performance of the students in relation to the material he has already studied, and students are well prepared for the test. Add to that, facility is reflected in the task clear instructions, logical organization of structure, appropriate timing and moderate level of difficulty.

Authenticity: The task is authentic as its language is as natural as possible. The items are contextualized and the topic is interesting and familiar.

- ***The second task is related to vocabulary:***

- Like the previous task, this task is a *Passage-based task.* However, it is different in the sense that it requires students to understand the text on word level and then choose words from the text that best suit the definitions provided.

- The mode of responding is a *short response* where students need to provide a word for each definition.

Second task: Find words in the text that mean the following:

1- change to be better suited to a situation

2- animal that kill and eat other animals

3-far from civilization or populated areas

-**Practicality:** This task is practical because: 1- administrative details are clearly established beforehand with clear instructions 2- ease of administration with no material or equipment needed other than papers. 3- students need to choose only one word for each sentence, so they can complete the task reasonably within the allocated period of time. 4- scoring system is feasible since the task does not have more than one correct answer.

-**Reliability:** *Inter-rater reliability* is high as multiple teachers work on the evaluation and re-evaluation of the task with reference to a predetermined rubric with one correct answer.

Test reliability is high since the definitions are clear and simple and can be answered once the text is understood.

-Validity: *Content validity* is present in the fact that students are used to this type of tasks. In addition, the task aims at assessing vocabulary knowledge in the form of finding three words for three definitions. Thus, answering correctly by finding the suitable words reflects comprehension and vocabulary knowledge which is the objective being assessed. *Face validity:* the task clear instructions, appropriate timing and moderate level of difficulty increase the face validity of the task.

-Authenticity: The task is authentic as it is passage-based and the items are contextualized.

- ***The third task aims again at assessing reading comprehension:***

Third task: Rewrite these sentences about the text to correct the information:

1- Jerboas carry their young in a pouch.
2- Lizards would stay alive if they were moved to the Arctic.

-The task is also a *Passage-based task* that requires examinees to depend on the information mentioned in the text in order to correct the sentences. To do so, students should fully understand the passage.
- The mode of responding is a *short response* where students need to re-produce the sentences sans the errors.

-**Practicality:** This task is practical because: 1- administrative details are clearly established beforehand with clear instructions 2- ease of administration with no material or equipment needed other than papers. 3- students need to produce short answers, so they can complete the task reasonably within the allocated period of time. 4- scoring system is not that practical as scorers need to make some judgments about the student' answers because there is more than one correct answer.

-**Reliability:** *Inter-rater reliability* is moderate as multiple teachers work on the evaluation and re-evaluation of the task with more than one correct answer.
Test reliability is low since the task does not include the number of incorrect information making it not clear if students figured all the errors or not.

-Validity: 1-*Content validity* is present in the fact that students are used to this type of tasks. Besides, the task measures what it is supposed to measure which is reading comprehension. *Face validity*: the task is face valid as it is structured in a way that the best students will be modestly challenged and the weaker students will not be overwhelmed.

-Authenticity: the task is authentic as it is passage-based and the items are contextualized.

5.2.2. The second section of the exam is also a reading text with Passage-based tasks that aim at assessing reading comprehension. However, it is different in the task formats used.

The first task is multiple-choice task where examinees are presented with item stem posing the question, followed by several answer choices from which the test takers must choose. "The correct option is called the key, and the incorrect answers are the distractors" (Carr,2011,P.29).

Alexander Pope recognized the depth and **originality** of Shakespeare's work. He said that Shakespeare developed characters himself when other playwrights reflected the work of others. In the 19th century, the romantic poets were inspired by Shakespeare's plays and used the same themes in their poems. At this time, Shakespeare was still considered more as a poet than as a playwright. Samuel Taylor Coleridge, one of the most **famous** romantic poets, noticed that some expressions in Shakespeare's work were about philosophy and psychology. Before Coleridge, these expressions were sometimes considered mistakes.
By the 1920s, Shakespeare was thought of as a playwright rather than a poet. Harley Granville- Barker argued that the works of Shakespeare were best when they were performed in a theater rather than read in a book. The globe theater in London was the place where Shakespeare's plays were performed during his lifetime. In 1997 the, theater was rebuilt and many new **critics** were able to watch and enjoy the plays. Today there are many critics who consider Shakespeare as both a playwright and a poet. ***First task***
Choose the correct answer a, b or c:
1-Romantic poets …. Shakespeare's plays.

didn't like affected influenced by
2 According to pop's opinion, Shakespeare…. A)imitated old playwright's characters
b)made his own characters
c)reflected the works of other playwrights

The mode of responding is *receptive or selective* (Brown,2004, p.56).

Hughes (2003) cautions against a number of weaknesses of multiple-choice items:

1-The technique tests only recognition knowledge.

2- Guessing may have a considerable effect on test scores.

3-the technique severely restricts what can be tested.

4. Cheating may be facilitated (2003, 76-78)

"The two principles that stand out in support of multiple-choice formats are, practicality and reliability" (Brown,2004, p.55). They are easy to write and score and they come with their predetermined correct answers. In addition, students do not need a lot of time to choose the correct answer, so they can finish the task in less time. This can be seen in the way the stem is restricted to only the information needed to pose the question.

The problem with MCQ is that even if the task is passage-based with only one correct answer depending on the meaning of the text, test-takers are offered a good chance to guess the correct answer without understanding the meaning or the text. Thus, students have the chance to pass regardless of their actual language ability. For guidelines how to write good items see (Haladyna, Downing and Rodriguez,2002).

Validity and authenticity: The task lacks content validity and authenticity because if guessing is used by the students, the answer will not reflect comprehension.

The second task is about vocabulary:

The task requires the students to match the definitions provided to words written in bold from the text.

Matching seems particularly a good task to be used. However, not to depend on logic or guessing more than meaning, the task includes more alternatives than there are prompts, with three bolded words and only two definitions.

Second task: **Match two of the underlined words from the text to the definition/meaning below:**

1- Well known

2- People who judge the merits of literally, artistic or musical works

The task is practical and reliable as it is clear, simple and easy to score with only one predetermined answer. It can also be solved very fast since the words are highlighted, and there is no need to re-read the whole passage. Unless the students employ guessing, it is considered reliable as it is expected to yield similar results whenever it was administered.

-**Validity and authenticity:** the task is content- valid and face - valid. It is also authentic.

The third task is related to assessing reading (sentence completion)

-the task is limited production (short answers) and passage-based task where students are asked to complete the sentences based on the information from the text.

- Carr (2011:33) mentioned two drawbacks of this kind of task:

1- "They take longer to grade or score than selected response items

2-Most questions will have more than one possible correct answer making it impossible not to include alternative answers that are considered acceptable." This makes it **less practical and reliable.**

***Third task*: Complete the following sentences with information from the text:**

1-In Shakespeare's lifetime his plays were acted in…….

2- Coleridge noticed that Shakespeare's work had some expressions about …….

As for **validity**, content validity is high because the task is a good way of assessing reading comprehension without being affected by other factors. On the contrary, face- validity is moderate since the instructions are not clear and do not specify with what the students are supposed to complete the sentences (are they expected to produce words or sentences?) This affects the difficulty of the task and, ultimately, the face- validity.

The following task is a gap filling one in which words of a specific type are deleted and students need to provide these words from memory.

Gap- filling Task: **Complete the following paragraph/sentences by filling in the gaps:**

1- Tariq's instruments have become famous across Syria…..the Arab world and there is now 2-a great demand…….. these instruments. 3- One of Tariq's sons decided to follow…….father into the business 4-so Tareq …….teaching him how to make the Oud.

"Rational-deletion cloze can be used to assess a number of different constructs" (Carr,2011,P.36) as it is used here.

As for **practicality and reliability,** it is impractical in the sense that no clear instructions are given and students will need a lot of time as they do not know what they are supposed to fill the blank with. However, the task is **content and face- valid** because the student are familiar with this kind of tasks, and the task itself is structured in a way that provides sufficient context for test takers to be able to infer what belongs in each blank. In addition, it is **authentic** as it is contextualized.

The following task is a gap filling one in which words of a specific type are deleted and students need to provide these words from a list in front of them.

Fill in the spaces with words from the list. Use each word once only:

Old art still creativity telling shows

1-People have been singing songs and…… each other stories for many thousands of years.

2-Forms of…..such as sculpture are at least 32,000 years old

3-this……..that even back then 4-people had the……. and ability to invent stories

5-spoken literature is therefor very…….indeed.

-The task is **practical and reliable** with clear instructions to choose each word one time and easy scoring system with predefined set of answers. In addition, the results are reliable as the guessing in such tasks is minimized since the question provides six words for five gaps.

Validity is evident as the task has face and content validity. It is also **authentic** because a context was provided for the gaps.

Question/answers formation:

This task is a short-answer task that asks students to produce answers, not merely to recognize them (Carr,2011, p. 96).

Students are assessed on their ability to form questions for certain answers and provide answers for certain questions.

This task is low in **practicality** because other than clear administrative details and instructions are provided beforehand, the task asks student to produce language within a limited time, which takes time and efforts. Besides, the scoring system is not that feasible and it needs time, as there are many alternative answers that can be used and are logically correct. Furthermore, the task is low in **reliability** because of the use of more than one scorer with a wide range of applicable answers. However, **face- validity** is clearly maintained as the apparent frame of the exercise predicts its purpose. Besides, the task is highly **authentic** due to its frequent use in real life.

Question/answers: **Complete the following dialogue by writing suitable questions or answers, Write at least three words for question:**

1-Rama:……..

Lama: Our last holiday was very exciting.

2-Rama:…….

Lama: we went to Cairo.

3-Rama:……

Lama: we arrived very late last night. 4- Rama: why did you arrive late?

Lama:……..

5.2.3. The following section is related to grammar:

Carr (2011, p.101) stated that in a grammar section, different item types should be placed together and this is seen clearly in the test being evaluated.

> **The first task *related to grammar* is a transformation task** where students have to change the sentence according to the instructions given in the question.

Transformation task: **Rewrite the following sentence as required in brackets:**
1-my room is too small
(use I wish……)
2-Fares didn't take his tooth out himself.(use the causative have)
3-can I go out with my friends
(report using Hanni asked his mother…….)
4-thousands of tourist visit historical monuments in Damascus.
(make passive voice)

The task is **practical** because: 1- administrative details are clearly established beforehand with clear instructions as the writers provide the students with what is required of them 2- ease of administration with no material or equipment needed other than papers. 3- Students need to produce short answers and the information is provided to them, so they can complete the task reasonably within the allocated period of time. On the other hand, 4- scoring creates a disadvantage, as scorers need to take into consideration different correct answers.

Assigning more than one scorer to the task might affect the results **reliability** as there is no one defined correct answer and teachers are required to make judgments.

The task is completely **valid** to measure learners' linguistic competence. However, it is low in **authenticity**.

The second task related to grammar is sentence completion:

As with the previously discussed short-answer questions, this task is moderate in practicality since it needs time to be done and the scoring is not that cut through. Scorers have to evaluate the answers on the bases of being logical and grammatically correct. Besides, students are asked to complete the sentences without any additional information except for 'to use a clause'.

Sentence completion: **Complete the following sentences using clauses:**
1-she went to school although,…….
2-if you broke the law,……

The not- so- clear instructions thus, affect **practicality, reliability, and face validity** of the task. The task is also not **authentic** since the sentences are not put within a context.

The third task related to grammar is a multiple-choice task where examinees are presented with two choices and they need to choose the suitable one:

Multiple-choice task: C**hoose the correct words in brackets:**
1-i am good (at, with) math, but I can't do calculations very quickly.
2-too much salt is bad for me, but I couldn't (do up, do without) it
3-the reported asked the old man (what, who) the secret of his healthy life was
4-Nadia's letter was so difficult to read (so that, because)she had written it quickly

-As mentioned above, this type of tasks is **practical and reliable**. However, usually with multiple-choice items more options are provided in order to minimize the factor of guessing. This is not the case in this task as students are only given two words to choose from, which is relatively easy and facilitates guessing and cheating. However, it is worth to mention practicality in the easy administration and scoring processes since there is only one correct answer for each sentence. The task is also deemed **valid** but not **authentic**.

The fourth task of assessing grammar is error correction:

Students have to correct the verbs between brackets by re-writing them in the correct tense.

Error correction: **Correct the verbs in brackets:**
1-she (feel) tired because she has been travelling for 2 days.
2-we (not spend) much time together last year.
3-he (drive) nearly 1000 km by the time he stopped for a break.
4-i (know) Ahmad since I was a child.

-The task is **practical and reliable** as it is clear, simple and easy to score on the assumption that there is only one predetermined answer. It can also be solved very fast since students need to produce short-answers of a couple of words and the incorrect words are highlighted so that the students' attention is drawn to them. The task also has **validity** as it is content and face- valid. However, the items are not contextualized, so it is not **authentic.**

5.2.4. Translation task:

"Translation may fall under limited or extended production، depending on the length of the material to be translated. (Carr,2011, P. 43). In the case of the 12th grade exam, it is used as a limited production task as students have to translate only two sentences from and to English.

Translation Task: **Translate the following sentence into Arabic:**
1-Gibran was deeply affected by Blake's works, which helped to shape his writings and paintings.
Translate the following sentence into English:
2- عاصمة البلد هي غالباً المدينة الأكبر والأكثر سكاناً

Translation is a good way to assess the ability to translate. However, when used to evaluate other language abilities, it is probably rather debatable because it may encompass any or all language areas, making it difficult to draw firm conclusions about any one of them in particular (Carr,2011, P.43). This lowers the **reliability** of the task.

In this task, scoring is the most difficult because there are many ways with which students can deal with the sentences. Any answer that is logical, grammatically correct and gives the right meaning is deemed correct. It should also be taken into consideration the de-contextualization of the sentences and the difficulty posed on students because of this fact. The previous points lower the task **practicality, face validity, and authenticity.** As for **content validity**, we can say that it is evident.

5.2.5. The final question is a writing task:

"Written responses to extended production tasks can range in length from several sentences or a single paragraph up to an entire essay" (Carr,2011, p. 39). based on the number of words provided in the prompt, the composition required can be a paragraph or an essay. Carr (2011) mentioned certain specifications that should be taken into consideration when writing prompts for writing tasks:

Writing task: **Write a composition of no less than 80 words on the following topic:** Recommendation to solve the following problem "very few tourists come to your town because they know nothing about it"

1-give the students clear directions and description of the task format: students should know what they are asked to do. This can be seen in the question provided there.

2-the purpose of the communication should be stated: the genre and rhetorical mode of the writing should be specified. This is not specified in the prompt, as students do not know the genre required.

3-mention the desired length of the expected responses and the topics of the prompts. The topic and an estimated length are given to the test-takers.

4-the criteria for assessment should be known for the students. As stated by (Carr, 2011, p.98) if an essay will be scored on the basis of grammatical accuracy, vocabulary use, content, and organization, those categories should be disclosed to the test takers in the directions. This is not mentioned in this task, as the marking criteria is not included.

The scoring process of this task is the most difficult even with clear and detailed rubric because the scorers need to take into consideration a lot of factors and points.

6. Main points of evaluation of the 12th grade English exam:

1-Practicality: As we have seen, the test is practical in some aspects while not so in others. However, overall, we can say that the test is practical due to many reasons: The first reason is that the test is an achievement test done for exiting the secondary school. This kind of tests in Syria is of great importance and is given special care. The test is administered in a specific time within specific educational departments where a calm examination atmosphere is usually created for the test-takers. The second reason is that the ministry of education makes sure that the test can be administrated smoothly and within a reasonable time frame. In addition, all the materials needed are usually prepared awhile before the students take the exam. The third reason is that the test itself pays attention to individual differences between test-takers as it delivers various types of items ranging from easy to difficult. Finally, the scoring process is done by a group of teachers who are given a reasonable period of time to score in order to minimize the pressure.

2-Reliability: Students reliability is affected in some tasks by many factors such as anxiety, fear of timing, and nervousness, which in general might negatively affect the students' performance. ***Rater reliability*** is evident as human errors are kept to minimum by providing clear and detailed scoring rubric for each item and assigning more than one scorer for each exam paper. ***Test Administration Reliability*** is different according to the conditions in which the test is administered. Nevertheless, taking into consideration the general state of most of the school in Syria (the desks, lights, space, and cleanliness) test Administration Reliability is not that high. Finally, ***Test Reliability*** is high as the items are clear, unambiguous, simple and not too long or require a lot of writing.

3- Validity: Content-related Validity is evident because the test is a curriculum-related with items drawn from the content of instruction covered during the year directly. As for ***Construct-related evidence***, the test is supposed to measure reading, writing, vocabulary, and grammar as stated in the test specifications, and the test items are designed to assess those four areas; so the test is construct valid. ***Face validity*** is high due to the students' familiarity with the task items that are replica of the book activities, the test durability within the allotted time limit, the clarity of directions and instructions and the reasonable degree of difficulty.

4-Authenticity: The degree of authenticity varies between different test items. Some items (III, IV) are contextualized, some (V, X) reflect real life tasks, and some (XI) provide interesting topics. On the other hand, some are decontextualized and do not simulate real world language use.

5-Washback: In the grading system used, test-takers get only a numerical score at the end of the scoring process without any further information about their performance. The test does not tell the students where they have problems and where they need to improve as they are only given a grade.

7. Summary:

A test is an assessment procedure by which teachers can assess student's work and progress. A good test is a test where you can find validity, reliability, practicality, Authenticity, and washback. As we have seen, the 12^{th} grade English final exam is an achievement test for all Syrian 12th graders. It is free, available to all students who are applying, and reasonably simple to administrate and score; all of these factors combine to enhance the test practicality. As for reliability, the importance of the test and its results and the pressure of scoring high marks and meeting the social conditions might get students stressed and anxious, thus lowering student-related-reliability. On the other hand, the test is scored by multiple teachers, which improve the rater-reliability. Add to that, the test has high test reliability. In some parts of the test, it measures students' actual performance with tasks drawn from the course of instruction, which increase the test content and face validity. In the same way, some parts of the test suffer in the authenticity perspective because of the use of items that do not resemble real-world tasks, while others are contextualized and reflect natural use of language. Finally, students are given only a grade with neither further information about their performance nor their weaknesses or strength which is not favorable for future progress. In other words, the washback effect of the test is limited.

References

Alderson,j.c., Clapham,c. and Wall,D. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University press

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L.F. & Palmer, A..S. 1996. *Language testing in Practice*. New York: Oxford University Press.

Brown, D. H. 2004. *Language Assessment: Principles and Classroom Practices.* White Plains, NY: Pearson Education.

Carr, N. T. 2011. *Designing and Analyzing Language Tests: A Hands-on Introduction to Language Testing Theory and Practice.* Spain: OUP Oxford.

Fulcher, G. 2010. *Practical language testing*. London: Hodder Education.

Fulcher, G., & Davidson, F. 2007. *Language Testing and Assessment*. London & New York: Routledge.

Haladyna, T.M., Downing, S. M. and Rodriguez, M.C. 2002**.** A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*,15,209-334.

Henning, G. 1987. *A Guide to Language Testing: Development Evaluation Research.* Cambridge, Mass.: Newbury House.

Hingle, I. & Linington, V. (2002). 'English Proficiency Tests: The Oral Component of a Primary School' in Richards, J.& Renandya, W. *Methodology in Language Teaching: An Anthology of Current Practice*. Cambridge: Cup.

Hughes, Arthur. 2003.*Testing for language teachers.* Second Edition. Cambridge: CUP.

Huerta- Macias, A. 2002. Alternative assessment: Responses to commonly asked questions. in Richards, J.& Renandya, W. *Methodology in Language Teaching: An Anthology of Current Practice*. Cambridge: Cup.

Madsen, Horald S. 1983. *Technique in Testing*. Hongkong: Oxford University.

Messick, S. 1996. Validity and washback in language testing. Language Testing 13, 3, 241-256.

Mousavi, Seyyed Abbas. 2002.*An encyclopedic dictionary of language testing.* Third Edition. Taiwan: Tua Book Company

O'Malley,J. M. & Pierce, L. V. 1995. *Authentic assessment for English Language Learners*. USA: Longman.

Penaflorida, A. 2002. 'Non-traditional forms of Assessment and response to student writing: A step Towards learner autonomy'. In Richards, J.& Renandya, W. *Methodology in Language Teaching: An Anthology of Current Practice*. Cambridge: CUP.

تقويم اختبار اللغة الإنكليزية لشهادة التعليم الثانوي في سوريا

أ.د علي سعود حسن

كلية التربية

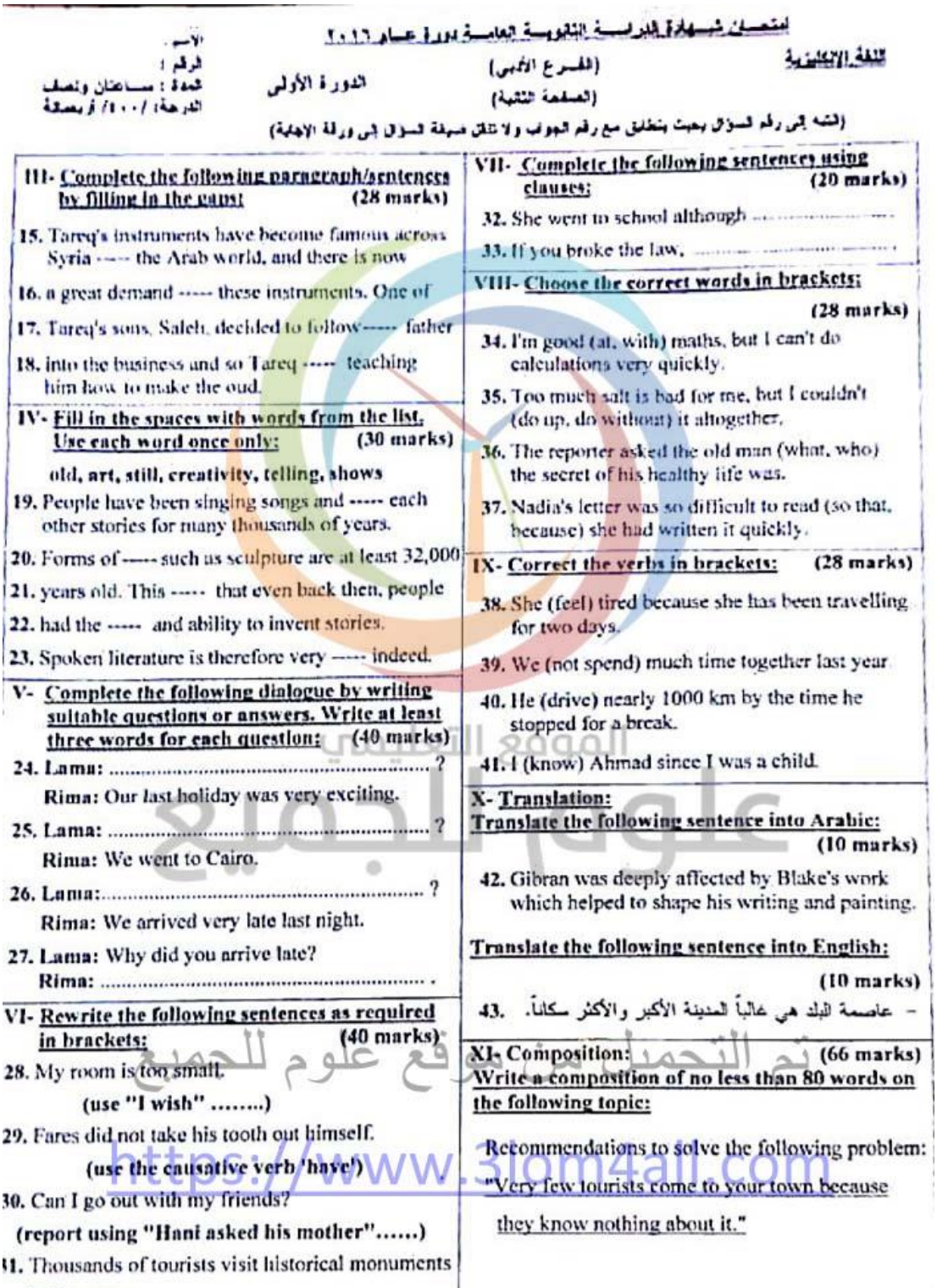
جامعة دمشق

يعد التقييم اللغوي عملية مستمرة تجري في الغالب في كل صف من صفوف تعليم اللغة الإنكليزية بوصفها لغة أجنبية, اذ يقوم المدرسون بتقييم أداء المتعلمين بشكل مستمر عن طريق الإجابة عن سؤال ما, او استعمال كلمة جديدة في تركيب نحوي او تقديم عمل مكتوب. وسواء أكانت هذه الأنواع في عملية التقييم عرضية أم مقصودة فإن لها أهمية كبرى في تقييم أداء المتعلمين.

يقدم هذا البحث شرحا لأنواع مختلفة من الاختبارات اللغوية موضحا الخصائص التي يتحلى بها الاختبار الجيد. كما يبين كيفية اجراء تقييم للاختبارات. وبشكل أدق فإن هذا البحث يقدم تقويما وافيا لأحد اختبارات اللغة الإنكليزية التي تجريها وزارة التربية في الجمهورية العربية السورية ألا وهو الاختبار النهائي لشهادة التعليم الثانوي لعام ٢٠١٦.

تظهر أهم نتائج البحث أن الاختبار قيد الدراسة عملي في بعض جوانبه ولكنه غير عملي في جوانب أخرى. كما أن الاختبار يتحلى بدرجة عالية من الثبات من حيث وضوح بنوده وبعدها عن الغموض وبساطتها وقصر طولها. كما أن الصدق الظاهري للاختبار هو أيضا على درجة عالية من الصدق. غير أن التأثير الرجعي للاختبار كان ضعيفا ذلك أن الطلبة لا يتزودون بالتغذية الراجعة المناسبة.

Appendix 1



لمتحانات شهادة الدراسة الثانوية العامة دورة عام ٢٠١٦

اللغة الإنجليزية (الفرع الأدبي)

الدورة الأولى (الصفحة الثانية)

الاسم :
الرقم :
المدة : ساعتان ونصف
الدرجة : ١٠٠/ مئة

(انتبه إلى رقم السؤال بحيث يتطابق مع رقم الجواب ولا تنقل صيغة السؤال إلى ورقة الاجابة)

III- Complete the following paragraph/sentences by filling in the gaps: (28 marks)

15. Tareq's instruments have become famous across Syria ---- the Arab world, and there is now
16. a great demand ----- these instruments. One of
17. Tareq's sons, Saleh, decided to follow----- father
18. into the business and so Tareq ----- teaching him how to make the oud.

IV- Fill in the spaces with words from the list. Use each word once only: (30 marks)

old, art, still, creativity, telling, shows

19. People have been singing songs and ----- each other stories for many thousands of years.
20. Forms of ----- such as sculpture are at least 32,000
21. years old. This ----- that even back then, people
22. had the ----- and ability to invent stories.
23. Spoken literature is therefore very ----- indeed.

V- Complete the following dialogue by writing suitable questions or answers. Write at least three words for each question: (40 marks)

24. **Lama:** .....?
Rima: Our last holiday was very exciting.
25. **Lama:** .....?
Rima: We went to Cairo.
26. **Lama:**.....?
Rima: We arrived very late last night.
27. **Lama:** Why did you arrive late?
Rima: .....

VI- Rewrite the following sentences as required in brackets: (40 marks)

28. My room is too small.
(use "I wish" .....)
29. Fares did not take his tooth out himself.
(use the causative verb 'have')
30. Can I go out with my friends?
(report using "Hani asked his mother".....)
31. Thousands of tourists visit historical monuments

VII- Complete the following sentences using clauses: (20 marks)

32. She went to school although .....
33. If you broke the law, ...

VIII- Choose the correct words in brackets: (28 marks)

34. I'm good (at, with) maths, but I can't do calculations very quickly.
35. Too much salt is bad for me, but I couldn't (do up, do without) it altogether.
36. The reporter asked the old man (what, who) the secret of his healthy life was.
37. Nadia's letter was so difficult to read (so that, because) she had written it quickly.

IX- Correct the verbs in brackets: (28 marks)

38. She (feel) tired because she has been travelling for two days.
39. We (not spend) much time together last year.
40. He (drive) nearly 1000 km by the time he stopped for a break.
41. I (know) Ahmad since I was a child.

X- Translation:

Translate the following sentence into Arabic: (10 marks)

42. Gibran was deeply affected by Blake's work which helped to shape his writing and painting.

Translate the following sentence into English: (10 marks)

43. عاصمة البلد هي غالباً المدينة الأكبر والأكثر سكاناً.

XI- Composition: (66 marks)

Write a composition of no less than 80 words on the following topic:

Recommendations to solve the following problem:
"Very few tourists come to your town because they know nothing about it."