People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

Mentouri Brothers University, Constantine

Faculty of Letters and Languages

Department of Letters and the English Language

An Evaluation of English Language Testing in the Baccalaureate Examination: The Case of the Tests Administered to Technology Streams from 2001 to 2006

Thesis Submitted to the Department of Letters and the English Language in Candidacy for the Degree of 'Doctorate Es-Sciences' in Applied Linguistics

Submitted by

Mr. Mohammed NAOUA

Supervisor:

Prof. Ahmed MOUMENE

Board of Examiners

Chairman:	Prof. Hacene SAADI	Mentouri Brothers University, Constantine
Supervisor:	Prof. Ahmed MOUMENE	Mentouri Brothers University, Constantine
Member:	Prof. Hacene HAMADA	ENS, Constantine
Member:	Prof. Salah KAOUACHE	Mentouri Brothers University, Constantine
Member:	Dr. Haoues AHMED SID	Mentouri Brothers University, Constantine
Member:	Dr. Amar BOUKRIKA	Jijel University

2016

Dedication

I dedicate this thesis to the memory of my dearest mother

To the memory of my father and my brother Ali

To my wife for her sustained support and encouragement

To my children Oualid and Nina

Acknowledgements

This work would not have been possible without the encouragement and sustained support of my supervisor Prof. Ahmed Moumene. I really owe an enormous debt of gratitude to him for his ongoing guidance and valuable guidelines throughout the writing of the thesis.

I am also indebted to the board of examiners, namely Prof. Hacene SAADI, Prof. Hacene HAMADA, Prof. Salah KAOUACHE, Dr. Haoues AHMED SID, Dr. Amar BOUKRIKA for accepting to read, review and evaluate this work

Our thanks are also due to Mr. Mime, the Chief examiner of Eloued rating center for responding to the interview questions. In the same way, my gratitude is extended to the raters in this center who provided us with valuable information concerning the scoring process in the BAC examination.

Abstract

The Ministry of Education in Algeria set several objectives for teaching English in secondary education. In technology specialties, for instance, the syllabuses were designed to enable learners to use this language in specific target domains, or to get access to scientific documentation while pursuing their further studies. Measuring the extent to which these objectives have been attained requires for testing and assessment. The scores obtained by these pupils in seven 'Baccalauréeat' sessions rank them bottom of the list lagging far behind all the other specialties in secondary schools. Seeing that these pupils study at the same institutions; use the same manuals and are almost instructed by the same type of teachers, this study attempts to focus on their BAC English tests for which we have formulated four hypotheses investigating the relationship between low achievement in these specialties on the one hand, and the scoring inconsistencies in the BAC English rating centers, the test construct underrepresentation, content irrelevance and the slim scope of sampling from the instructional domain on the other. The hypotheses have been verified by the data that we have collected by means of the descriptive method instruments such the questionnaire, the interview and the documentary sources. The questionnaire was administered to a population of 63 raters gathering for the purpose of scoring the BAC English tests in Eloued Rating Center. The interview was conducted with the chief examiner of the same center. We have also supplemented our data with documentary sources such as the pupils' scores, their BAC English test papers and their instructional syllabus. The findings of the study have come to challenge the validity of technology pupils' test score interpretations and the purposes for which the scores have been used. The main purpose of this study is to identify the factors responsible for technology pupils' underachievement in English and to propose a set of recommendations intended to improve the process of English language testing in the Baccalaureate examination.

Key Words: Constructs – Evaluation – Reliability - Technology -Testing –Validity

Abbreviations

AERA: American Educational Research Association

APA: American Psychological Association

BAC: The Baccalauréeat

CALL: Computer Assisted Language Learning

CALT: Computer Assisted Language Testing

CAT: Computer Adaptive Testing

CBT: Computer Based Testing

CC: Communicative Competence

CCT: Communicative Competence Trend

CLA: Communicative Language Ability

CRT: Criterion-referenced Tests

CSs: Communication Strategies

CTT: Classical Test Theory

EAP: English for Academic Purposes

EE: Electrical Engineering

EOP: English for Occupational Purposes

ESP: English for Specific Purposes

EST: English for Science and Technology

F V: Facility Value

ID: Item Difficulty

LCT: Language Competence Trend

LK: Language Knowledge

LSP: Language for Specific Purposes

ME: Mechanical Engineering

MSs: Metacognitive Strategies

NCME: National Council on Measurement in Education

NRT: Norm-referenced Tests

ONEC: Office National des Examens et Concours

SLA: Specific Language Ability

SPBK: Specific Purpose Background Knowledge

SPECs: Specifications

SPLA: Specific Purpose Language Ability

UCH: Unitary Competence Hypothesis

List of figures

List of Graphs

List of Tables

Contents

Chapter One

Approaches to Language Testing

Chapter Two

Constructional Constituents of Language Tests

Chapter Three

Stages of Test Development

Chapter Four
Investigating Rater Reliability

Chapter Five

Investigating Test Validity

Chapter Six: Field Study

Validating the Score Interpretations, Relevant to Technology Streams' BAC English Tests

Chapter Seven

Findings Implications and Recommendations

General Introduction

General Introduction

Background of the Study

Testing is one of the main characteristics of human social life. Throughout history, people have been put to tests in order to examine their suitability for a given position or to measure their standing on different types of construct (Bachman, 2004b; Spolsky, 1995, 2008). Testing practices date back to the Hun San Dynasty (206 BC - AD 220) in Imperial China where the emperors used these instruments as a means for providing the civil administration with talented officials on the basis of merit and excellence rather than on their social background or patronage (Kunnan, 2008; Spolsky, 2001-2005).

In modern societies, testing has come to play a very powerful and influential role in people’s lives. This is because its results "can create winners and losers, successes and failures, rejections and acceptances” (Shohamy, 2001, p.113). Consequently, if tests are used for the purposes for which they have been designed, they will certainly yield positive consequences for the stakeholders and serve as door-openers or gateways to different opportunities and positions. Conversely, if these instruments are not used for the purposes they have been intended for, they can have detriment consequences on test takers serving as gatekeepers, limiting their chances of success, or of joining academic or occupational positions (Alderson, Clapham & Wall, 1995; Bachman & Purpura, 2008).

In education, testing is used as an instrument for monitoring the learning progress or for evaluating the educational system as a whole (Alderson & Buck, 1993). In the same way, the scores resulting from this process can be used to make important decisions about individuals and institutions (Bachman, 2005, 2007; Messick, 1996). These decisions can involve test takers' selection, placement, promotion, certification, retention at the same educational level, or even exclusion from schooling. Similarly, these results can also be used in the categorization of schools according to the extent of candidates' achievement. The high ranking schools can, for instance, be labelled as 'superior' or successful schools; while the low-scoring ones can be identified as 'inferior' or 'failing' schools (Popham, 2001, 2003).

As far as language is concerned, the process of testing attempts to make inferences about test takers' levels of language ability; and to make predictions about their capacity of using this language in real target domains. This process consists of two main components: the 'what' and the 'how'. The 'what' refers to the construct(s) that we intend to measure; and the 'how' pertains to the methods, techniques, or facets used to assess these construct(s) (Bachman & Purpura, 2008; Kane, 2013; Shohamy, 2008).

Statement of the Problem

The Ministry of Education in Algeria set several aims for the teaching of English in secondary education. These have been adapted to respond to the requirements of learners in each specialty, or stream. In literary streams, for example, the intent was to enable the pupils to use this language for general communicative purposes (Ministry of Education, 2004). In scientific specialties, the focus has been laid on written communication for the pupils will, according to the Ministry, use English for research writing and experimentation reporting. However, the ultimate objective of teaching this language in technology streams has been to enable learners to use it for specific purposes and in constrained target domains relevant to their fields of specialism (Ministry of Education, 1995).

Measuring the extent to which these objectives have been attained requires testing and assessment. The examination of technology pupils' evaluation records manifest significant differences between the scores they obtained in achievement tests and those obtained in the BAC English tests (Orientation Centre of Eloued, 2001-2006). The scores obtained in achievement tests suggest that these pupils' level in English is similar to that of their colleagues in the other streams. Conversely, apart from June 2001 BAC session, in the following sessions (September, 2001-2006) their results in English rank them at the bottom of the list lagging far behind all the other specialties in secondary education (see appendix A). This leads us to raise the following concern: why have technology pupils in Eloued been achieving the worst results in the BAC English tests?

Aim of the Study

The main aim of this study is to conduct an empirical analysis in order to diagnose the factors responsible for technology pupils' low achievement in the BAC English tests from 2001 to 2006. The results of the analysis will then be incorporated in Toulmin's (2003) argumentation framework to examine the extent to which the interpretations provided for the scores obtained by these pupils are real indicators of their level of language ability. The study will conclude with a set of recommendations intended to improve the process of English language testing in the Baccalaureate examination

Research Questions

Research methodologists distinguish between research problems and research questions. The former refer to some type of difficulty that a researcher encounters or experiences during his study of a given topic or phenomenon; and for which he seeks to find a solution (Kothari, 2004). However, the latter refer to some "specific question[s] asked in the course of investigation to which a specific answer or set of answers is sought…before arriving at possible hypotheses" (Tavakoli, 2012, p. 49). So, in order to investigate the source of technology pupils' low achievement in the BAC English tests, this research tends to answer the following questions:

1- Do technology pupils in Eloued really have low levels of language ability; or have not they been provided with the opportunity to display their standing on this competence?

2- Have the BAC English tests in these streams measured the constructs that test developers intended to measure?

3- How difficult were the test tasks designed for these specialties? In other words, have they fallen beyond the pupils' mental capacities?

4- Have these pupils been provided comparable and equitable testing conditions as their colleagues in the other specialties?

5- Have their BAC English tests been pre-evaluated to be certain that they are free from bias?

6- How consistent were the scoring procedures in the BAC English rating centers?

Hypotheses Formulation

In addition to the problem, hypotheses can be considered as the main elements of scientific research because of their role in linking theory to investigation, which results in more discoveries in knowledge (Cohen, Manion & Morrison, 2007; Goode &Hatt, 1952; Kerlinger, 1973). A hypothesis can be defined as "a proposition which can be put to a test to determine its validity…It may prove to be correct or incorrect. In any event, however, it leads to an empirical test" (Goode &Hatt, 1952, pp.56-7). Kerlinger (1973) identifies three main reasons for the indispensability of hypotheses to scientific research. First, they represent the operational devices of 'theory'. Secondly, these devices can be tested to be shown true or false. Third, testing hypotheses can lead to the "advancement of knowledge" (p. 18).

Hypotheses can be classified into two types: alternative and null hypotheses. The former postulate that there is a relationship between dependent and independent variables; while the latter assume that that no relationship exists between the variables being studied. In other words, the null hypothesis 'says', as Kerlinger puts it, “you’re wrong, there is no relation; disprove me if you can” (p. 204).

Due to the fact that technology pupils in the region of Eloued study at the same schools with the other specialties, use the same manuals as all the secondary education pupils; and are almost instructed by the same teachers, this research tends to focus on their BAC English tests for which we have formulated four hypotheses. The latter seek to investigate the relationship between low achievements in these specialties on the one hand; and the scoring procedures in the BAC English rating centers, the test construct and content underrepresentation as well as the narrow scope of sampling from the instructional domain on the other.

1- Hypothesis one postulates that if the scoring processes in the BAC English rating centers are reliable and consistent, technology pupils can obtain higher scores in these tests.
2- Hypothesis two relates low achievement in technology specialties to the deficiency of the tests to measure the defined constructs. In other words, if these tests focus on measuring the constructs that test developers planned to assess, the pupils' achievement in English would improve.
3- Hypothesis three: If the test mirrors the content included in the pupils' instructional syllabus, their background and language knowledge could interact positively with the test input.

4- Hypothesis four: If the test developers implement the process of 'ecological sampling' to ensure that all the important parts of the domain are represented in the test, the pupils' specific language ability would be engaged by one part or the other of the test input.

Research Methodology

The *Sage Dictionary of Social Research Methods* (2006) considers research methodology as the 'philosophy of methods', which covers two main components: epistemology and ontology. The former, which the dictionary labels as the ‘rules of truth’, strives at justifying the soundness and dependability of the research findings and its conclusions. The latter concerns "establishing the ‘objects’ about which questions may validly be asked and conclusions may be drawn" (p. 175). Deducing from this definition, we can say that research methodology refers to the theory that outlines how research is systematically conducted starting from the problem identification and concluding with its findings and conclusions. This involves the conceptualization and statement of the problem, hypotheses formulation, specifying the relevant survey methods, defining the appropriate population and data gathering tools with ethical considerations; and stating the criteria for analyzing data and presenting the research results.

Choice of Method

The method in scientific research refers to the procedures and techniques that we employ in order to gather evidence about a given phenomenon (Cohen, et al., 2007; Goode & Hutt, 1952). Research methodologists classify the methods into three broad types: survey, historical and experimental methods (Goode & Hutt, 1952). Survey methods are then reorganized into four types: descriptive, analytical, school survey, and genetic methods. Survey methods which employ the descriptive and analytical techniques as a means for gathering data through observations, tests, questionnaires, schedules or interviews seek "to describe the distribution of phenomena in a sample and population [and] to explain relationships between variables – to explain why things are as they are" (Jupp, 2006a, 284).

Seeing that this investigation is concerned with describing a current phenomenon, (technology streams' low achievement in the BAC English tests) and attempting to explain the relationship between dependent and independent variables based on the data that we intend to collect by means of the interview, the questionnaire, and documentary resources, we found that the most convenient procedures and techniques for conducting this research, are the ones stipulated by the survey (descriptive and analytic) method.

Population and Sampling

The population

The population subject to this investigation is composed of sixty-three (63) secondary school teachers participating in scoring the BAC English test (session 2013) in Eloued rating center (Guémar technical school). These respondents are affiliated to forty secondary education institutions distributed in the 'wilaya' of Eloued. Due to the fact that these subjects are grouped in one rating center; and for the purpose of gathering more efficient data, we decided to survey the entire population.

Data Gathering Tools

Research methodologists identify several data gathering tools in descriptive research (Kerlinger, 1986). These include tests, observations, questionnaires, interviews, and documentary resources. Three of these instruments will be used in this survey. We will use the group questionnaire for collecting information from a group of respondents gathering for the same purpose (raters). Seeing that the majority of raters do not attend the mediation phase of scoring; or what is known as 'la troisième correction', we thought that it would be more beneficial for us to interview the BAC English test chief examiner who oversees this process from its beginning until the arbitration phase. However, in certain cases, questionnaires and interviews do not always provide us with all the types of information relevant to this research. Consequently, we felt the need to supplement data from documentary sources such as technology pupils' instructional syllabus, copies of their BAC English tests (ONEC, 2001-2006) as well as the scores they have obtained in seven BAC sessions (2001-2006).

Definition of Terms

In the context of language testing, the terms assessment, evaluation, measurement and tests are often used interchangeably (McNamara, 2000). This "tends to obscure the distinctive characteristics of each….[This is why,] an understanding of the distinctions among [these] terms is vital to the proper development and use of language tests" (Bachman, 1990, p. 18).

Assessment

Language assessment refers to the process of collecting information by means of tests in order to make inferences by standard and 'explicit rules' about a given aspect of individuals' language ability for the purpose of making a variety of decisions about participants, programs and institutions (Gage & Berliner, 1991; Richards & Schmidt, 2002; Weigle, 2002).

Measurement

Measurement refers to "the process of quantifying the characteristics of persons according to explicit procedures and rules" (Bachman, 1990, p. 18). In language testing, quantification means the assignment of numbers (scores) to individuals' mental traits. Unlike physical characteristics such as length, height, or color, which can be observed and directly measured, mental traits can be inferred and indirectly observed in the way we behave. Of course, the assignment of numbers in measurement should not be done haphazardly, but according to explicit rules and procedures, such as scoring guides or rating scales.

Tests

A test can be defined as an instrument or a procedure that is "designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual" (Carroll, 1968, p. 46). Carroll maintains that the main concern of this instrument "is always to render information to aid in making intelligent decisions about possible courses of action" (p.314).

Evaluation

According to the *Sage Encyclopaedia of Qualitative Research Methods* (2008), "to evaluate is to determine the value of something, that is, to determine its merit, worth, or significance" (p. 683). Similarly, in the point of view of Bachman (2004a) "evaluation which involves making value judgments and decisions can best be understood as one possible use of assessment (p. 9).

Relationship amongst Assessment Terms

The relationship amongst assessment, measurement, evaluation and tests can be determined as follows. In assessment, we collect data by means of tests in order to assign scores to the responses of test takers (measurement). The resulted scores are then used to make decisions about test takers and their teachers (evaluation). In this context, we can say that tests are instruments of measurement, which in its turn, represents one type of assessment; however, evaluation can be considered as one probable use of assessment (Bachman, 1990, 2004a; Douglas, 2012; McNamara, 1996).

Structure of the Thesis

This thesis is organized into seven chapters. The first three chapters describe the process of language test construction and development. Chapters four and five review the literature relevant to the evaluation of tests such as reliability, validity, and validation. Chapter six focuses on the analysis of the information included in the questionnaire, the interview and the documentary sources. Chapter seven lays out the results of the research and proposes several recommendations for the purpose of improving the process of English language testing in the baccalaureate examination.

Chapter one accounts for the chronological development of language testing approaches. This chapter starts with Spolsky's (1979) division of the history of the field, which identifies three main trends: the pre-scientific, the structural-psychometric, and the sociolinguistic-integrative trends. Then, it sketches out how the last trend had been overshadowed by communicative language testing and ESP testing. In addition, this chapter portrays the incorporation of computers in test administration and scoring, and concludes with a description of the purposes for which tests can be designed.

Chapter two describes the three 'layers' for language test construction: models of language ability, test frameworks, and specifications (Fulcher, 2010). Models of language ability specify the broad lines of what it means to know and to measure language traits or performance. Test frameworks, on the one hand, select the constructs to be measured form the theories of language; and generate the specifications on the other. The specifications tell us how to design items and how to compile them into comprehensive tests (Fulcher, 2010; Fulcher & Davidson, 2007, 2009).

Chapter three describes Bachman and Palmer's (1996) three stages of language test development: design, operationalization, and administration. The design stage delineates the general purpose of the test; analyzes the target language domains; collects data about test takers' characteristics, their language abilities, and test tasks so as to ensure three types of authenticity. Test takers' characteristics need to be similar to those of real language users; test tasks need to resemble to tasks in target language domains; and the abilities to be tested need to bear some resemblance to language users' abilities. The second stage outlines how to design tasks and how to assemble them into a comprehensive test. The third stage describes the two phases of administration: item tryout and live test delivery.

Chapter four focuses on explaining the concept of 'reliability' and how it can be implemented in the evaluation of test scores. It describes the phases of scoring; explains how true scores can be computed; outlines the criteria for training and appointing raters; and proposes the different techniques for establishing inter rater and intra rater reliability.

Chapter five describes the concept of validly and the process of validation. It reviews validity from the point of view of two schools: the traditional and modern schools (McNamara &Roever, 2006; Messick, 1989, 1996). The traditional approach takes validity to be consisting of different types; and tests can be validated according to their content, criterion, or construct (Hughes, 1989; Lado, 1961). Conversely, the modern approach takes validity as an overall unitary concept, and emphasizes that what needs to be validated is not the test itself, nor its scores, but the interpretation, uses and consequences of the obtained scores (Messick, 1989, 1995). In reviewing the literature relevant to test validation, this chapter introduces Toulmin's (1958, 2003) philosophy of argumentation and explains how it can be implemented in validating the score interpretations and the consequences resulting from the score uses.

Chapter six 'field work' focuses on examining the interpretations, uses and consequences of the scores obtained by technology pupils in Eloued in seven BAC English sessions (2001-2006). What is worth mentioning here is that in 2001, we witnessed the organization of two BAC sessions: the first in June, and the second in September. The validation process employs Toulmin's model of argumentation (1958, 2003). It introduces technology streams' BAC English test scores as its data. The interpretation and uses suggested for these scores as its claim. The information gathered by means of the questionnaire and interview as its warrant and backing; and the evidence collected from test copies and the official syllabuses as its rebuttal.

Chapter seven 'Findings, Implications and Recommendations' lays out the main results of the research and suggests some solutions to the problems. The implications delineate the main areas of English language test development and evaluation relevant to the Baccalaureate examination. The chapter concludes with a set of recommendations intended to test designers, test users, and educational assessment organizations on the issue of test construction and validation.

Chapter One

Approaches to Language Testing

Chapter One

Approaches to Language Testing

Introduction

Language testers divide the history of language testing into several stages or approaches (Fulcher, 2010; Spolsky, 1979). The latter have been developing by approximations in which each new stage results from the improvement of the previous one. The pre-scientific approach, for instance, had extended from the start of the Chinese Imperial system of examinations until the late fifties. In the next stage which started from the early sixties and expanded until the early seventies, the psychometric-structuralist approach was the dominant. In the third stage, the triumph was for the integrative-sociolinguistic trend. Since the beginning of the eighties, the testing pendulum has fallen in favour communicative language testing, ESP testing as well as computer assisted language testing (Douglass, 2000; Fulcher & Davidson, 2007; McNamara, 2000).

1.1. Spolsky's Outline of the History of Language Testing

Applied linguists and language testers refer to Spolsky's (1979) division of the history of language testing (Brown, 1996; Davies, 2003; Malone, 2008; McNamara, 2003; Shohamy, 2008; Stansfield, 2008). Bernard Spolsky (1979) divides the history of language testing into three main trends: the pre-scientific, the psychometric-structuralist and the integrative-sociolinguistic trends. According to him, these approaches "follow in order but [sometimes] overlap in time" (p. 6).

1.1.1. The Pre-scientific Stage

The Pre-Scientific or the traditional approach was, as its name implies, characterized by the lack of assessment literacy and testing expertise (Malone 2008). The field was in its exclusivity the business of language teachers, for it was taken for granted that if one knows how to teach; he will be able to test and measure learners' language proficiency. In this period, the testing practices were intuitive and subjective, ignoring the qualities of reliability and validity. Language tests focused on written examinations, such as dictation, translation of texts, free compositions and sentence completion. The main characteristics of this stage are summarized by Spolsky (1979):

> The pre-scientific period (or trend, for it is still holds sway in some parts of the world) may be characterized by the lack of concern for statistical matters or for such notions as objectivity and reliability. In its simplest form, it assumes that one can and must rely completely on the judgment of an experienced teacher, who can tell after a few minutes' conversation, or after reading a student's essay, what mark to give…During this period, and in this approach language tests are clearly the business of language teachers, or, in more formal situations, of language teachers promoted or specially appointed as examiners. No special expertise is required, if a person knows how to teach, it is to be assumed that he can judge the proficiency of his students (pp. 6 & 7).

1.1.2. The Psychometric-structuralist Stage

In this stage, two types of experts joined the field of language testing: psychometricians and structural linguists (Spolsky, 1979). Each category perceived language testing as their private property. Psychometricians introduced notions about "the utilizations of numerical data and related logical operations in the service of developing, using, and interpreting the results of the measurement activities" (Clark, 1979, p. 26). On their part, structural linguists introduced models of language ability describing the constructs to be tested. The alliance of the two trends led to the introduction of new concepts such as reliability, objectivity, and validity. The main achievement of this period was the implementation of discrete-point and standardized testing.

1.1.2.1 Discrete-point Testing

Discrete-point testing lends itself to Robert Lado (1961) who hypothesizes that peoples' language ability is made up of various components such as "sounds, intonation, stress, morphemes, words, and arrangements of words having meanings that are linguistic and cultural" (p. 25). These elements constitute of variables that need to be tested. Sounds, intonation, stress, for example, make up the variables of pronunciation. Grammatical structure is broken down into two sets of variables: morphology and syntax; and words are organized according to their linguistic or cultural meaning. Lado points out that though these elements "never occur separately in language [and] … are integrated in the total skills of speaking, listening, reading and writing [however, they] can be profitably studied and described –and tested- as separate universes" (p.25). Concerning the choice between testing isolated elements of language, or the situations in which they are used. Lado emphasizes that while the number of the former is limited and can easily and effectively be sampled and tested "the situations in which language is the medium of communication are potentially almost infinite" (p, 26) which makes it difficult to test all the situations that occur in language use.

Documenting for the testing practices which occurred during this stage, Spolsky (1979) writes:

> The psychometric structuralist trend ... is marked by the interaction and (conflict) of two sets of experts, agreeing with each other mainly in their belief that testing can be made precise, objective, reliable, and scientific. The first of these groups of experts were the testers, the psychologists responsible for the development of modern theories and techniques of educational measurement. Their key concerns have been to provide "objective" measures using various statistical techniques to assure reliability and certain kinds of validity….The second impetus of the "scientific" period , or approach, then, was when a new set of experts [linguists who] added notions from the science of language to those of the science of measurement… The marriage of the two fields, then, provided the basis for flourishing of the standardized language test with its emphasis on what Carroll (1961) labelled the "discrete structure point" item (pp. 6, 7, & 8).

1.1.3. The Integrative-sociolinguistic Stage

This Integrative-Sociolinguistic stage was marked by the association of two trends of linguistics the 'Language Competence' trend (LCT) and the 'Communicative Competence' trend (CCT). Refusing the hypothesis of structural linguists which implies that language ability is made up of discrete elements, and the constructs that need to be tested are its atomistic variables, the 'LCT' trend assumed that language ability represents an overall inseparable unity 'the unitary language proficiency', and measuring this ability should also be conducted accordingly. According to this school, testing should target the discrete elements of language when they are incorporated in the skills of listening, speaking, reading, writing, or the ability to translate (see Table 1). Reflecting the views of the 'LCT', Carroll (1961) articulates that "the four skills must…be regarded as integrated performances which call upon the candidate's mastery of the language as a whole i.e., its phonology, structure, and lexicon" (pp. 317-8). For this reason, he recommends "tests in which there is less attention paid to specific structure points, or the lexicon then the total communicative effect of an utterance" (p. 319).

Table 1: Carroll's Integrative Testing Grid

Skill	Language Aspect			
	Phonology or Orthography	Morphology	Syntax	Lexicon
Auditory comprehension				
Oral Production				
Reading				
Writing				

Source: Carroll, 1961, p. 316

The second school, the 'CCT' trend was influenced by sociolinguistic theories emerging in the early seventies such as Hymes' theory of communicative competence (1972), Savignon's views on communication strategies in the classroom (1972); as well as Halliday's functional grammar (Halliday 1973; Halliday & Hasen, 1976). This trend which perceived the understanding of language with respect to the social context that it is used in, "accept[ed] the belief in integrative testing, but insist[ed] on the need to add a strong functional dimension to language testing" (Spolsky, 1979, p. 9).

Briefly speaking, in his outline of the history of language testing, Spolsky (1979) identifies three major stages: the pre-scientific, the psychometric-structuralist and the integrative-sociolinguistic stages. The first stage was characterized by the lack of assessment literacy on the part of testers. In the next stage, language testing bore the imprints of psychometricians and structural linguists. In the third stage, the concern shifted from testing isolated elements of language to measuring integrated skills in contextualized situations. Summarizing the state of the art during these periods, Spolsky (1979) comments:

> Originally, testing was simply a teacher's function, although many people believed a teacher's judgment automatically improved when he changed hats and was identified as examiner. Next, experts on testing moved into the field with their principles. It was soon shown that psychologists alone could not develop good language tests: some linguists like Lado showed that the job needed to be shared and to depend on two kinds of expertise. Finally, a group of psycholinguists and sociolinguists, with somewhat imperialistic notions, are starting to claim the field for themselves. Language testing, they seem to be saying, is too important to be left to language testers. (p. 11)

1.2. Communicative Language Testing

The alliance between the 'LCT' and the 'CCT' concerning integrative tests did not last for a long time. The reason for this divergence was related to their conceptualization of the language ability to be tested. The 'LCT' takes language proficiency to be consisting of a 'single unitary ability' (Oller,79); whereas developments in sociolinguistics led the 'CCT' to view language ability as multi-componential, made up of numerous constructs each of which can separately be tested (Alderson, 2000a; Bachman & Palmer, 1996; Canale & Swain, 1980; Hymes, 1972, Savignon, 1972).

At the beginning of the eighties, the language testing pendulum fell completely in the favour of the 'CCT' (Bachman, 1991). Documenting for the rise of a new stage in language testing, Moller (1981) remarks that it is now "perhaps time to identify a fourth phase in language testing, closely linked to the third, the sociolinguistic-communicative phase" (p. 39). Communicative Language testing came as a reaction to the failure of the discrete-point and integrative testing approaches to give "any convincing proof of the candidate's ability to actually use the language, to translate the competence (or lack of it) which is demonstrating into actual performance 'in ordinarily situations'" (Morrow, 1981, pp.15-16). Morrow's views were later supported by Weir (1990) who argues that integrative tests focused on measuring candidates' linguistic competence, but did not inform us about their communicative performance. However, the "serious limitation" of integrative testing "was its failure to recognize the full context of language use - the contexts of discourse" (Bachman, 1990, p.82).

1.2.1. Definition of Communicative Language Testing

According to Morrow (1981), Communicative Language testing can be defined as:

> The assessment of the ability to use one or more of the of the phonological, syntactic and semantic systems of the language (1) so as to communicate ideas and information to another speaker/ reader in such a way that the intended meaning of the message communicated is received and understood, and (2) so as to receive and understand the meaning of a message communicated by another speaker/writer that the speaker/writer intended to convey. This assessment will involve judging the quality of the message and the quality of the expression and of its transmission, and the quality of its reception in its transmission (p. 40).

1.2.2. Characteristics of Communicative Tests

Bachman (1991) identifies four characteristics that distinguish communicative tests. These include information gap, task dependency, integration of task and content with a given discourse domain, and the wide scope of the abilities to be measured. The first characteristic requires test takers to "process complementary information through the use of multiple sources of input"(p. 678). Bachman explains that a writing task can be "based on input from both a short recorded lecture and a reading passage on the same topic" (p. 678). Task dependency means that doing tasks in one section depends on the content of the previous one. Third, communicative tests are tightly connected with discourse domain. For example, in tests measuring the language ability of electrical engineering pupils, the content is supposed to be linked to the type of language used in their specific academic domain. Fourth, these tests tend to measure a wide spectrum of language ability such as cohesion, coherence, pragmatics, language use, sociocultural knowledge or strategic competence.

Advances in communicative language testing led to the development of two main issues (Bachman, 1991). The first concerns developing theoretical models describing the different components which make up the concept of 'communicative competence' (Bachman, 1990, Bachman & Palmer, 1996; Canale 1983; Canale & Swain, 1980; Hymes, 1972; Savignon; 1972, 2002); and the second concerns developing test methods describing test task and test takers' characteristics (Bachman, 1990; Bachman & Palmer, 1996). Providing models for language ability enables us to precisely determine the constructs that we intend to measure; however, the description of test task characteristics helps us design tasks, which can engage test takers' language knowledge to mutually interact with the test input.

1.3. LSP Testing

Testing language for specific purposes 'LSP' can be considered as a special case of communicative language testing (Widdowson, 1978, 1983). 'LSP' testing refers to the process of making inferences about test takers' specific language ability, and about using this ability in specific target domains (Douglas, 2000, 2001, 2013).

1.3.1. Definition of ESP

Several definitions have been provided to LSP in the literature. These definitions have focused on highlighting the main features of this approach, such as the analysis of learners' needs, the description of target language situations, the content specificity, as well as the homogeneity of participants (learners/test takers). Echoing the views of Strevens, (1972), Mumby (1978), Widdowson (1983), Hutchinson & Waters, (1987), Dudley-Evans and St John (1998) and Douglas (2000, 2001), Basturkmen and Elder (2004) consider LSP as:

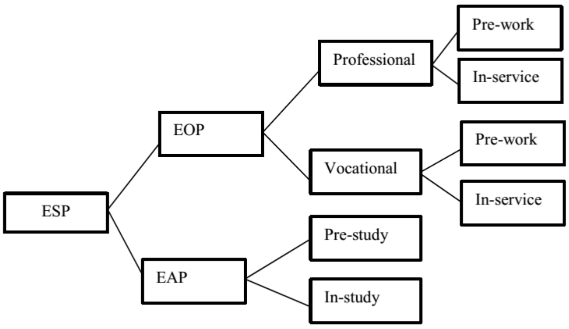
> The teaching and research of language in relation to the communicative needs of speakers of a second language in facing a particular workplace, academic, or professional context. In such contexts, language is used for a limited range of communicative events… Analysis of language in such events generally reveals that language is used in constrained and fairly predictable ways. pp. 672-3.

The authors emphasize that LSP teaching should "focus on the specific language needs of fairly homogeneous groups of learners in regard to one particular context referred to as the target situation" (p.673). Within LSP contexts, Mumby (1978) points out that the terms 'specific' and 'special' should not be used interchangeably. This is because a 'special' course implies that it is not 'ordinary'. However, "the phrase 'specific purpose' [implies] that it is not general"(p.2). Mumby maintains that LSP "should focus on the learner and the purposes for which he requires the target language, and the whole language programme follows from that" (pp. 2-3).

1.3.2. Types of ESP

English for specific purposes can be classified into two broad categories: English for academic purposes (EAP) and English for occupational purposes (EOP) (Douglas, 2010b; Hutchinson & Waters, 1987) (see fig 1). The former concerns the learners who need the language for educational purposes such as pursuing studies in a given academic field of interest. The second category 'EOP' refers to the use of language with the intention of performing part or all of a job (Douglas, 2000). Each type of ESP consists of pre-service and in-service programs. The former "refers to courses designed for learners aspiring to enter particular workplace, academic, or profession situations [while] the latter [is] designed for learners already involved in the target situation" (Basturkmen and Elder, 2004, p. 673). The division of ESP into these types intends to achieve two main purposes. In the first place, it can contribute to determining the specific target language situations so as to provide learners with the appropriate syllabi and teaching material. In the second place, we can devise the specific test tasks to decide "whether applicants have enough control over the target language to succeed in academic studies, [or] to determine whether job applicants or employees can carry out necessary functions in the target language"(Douglas, 2010b, p.3).

Fig 1: Types of ESP



Source: Douglas, 2010b, p. 11.

1.3.3. Definition of LSP Tests

An LSP test can be defined as the:

> one in which content and method are derived from an analysis of specific purpose target language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test takers' language ability and specific purpose content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about test takers' capacity to use language in the specific purpose domain. (Douglas, 2000, p. 19)

It follows from the definition provided above, that an LSP test is a measure that intends to offer information about test takers' capacity of using language for specific purposes in specific constrained target language domains. In order to justify the uses and interpretations of these tests, three types of correspondence need to be implemented. First, the abilities that are intended to be tested should resemble, to some extent, the language abilities of LSP real-life users. Secondly, the characteristics of test takers' should be similar to those of LSP users. In the third place, test tasks need to be similar to the ones that usually occur in specific target situations. Another point which can be related to the third type of correspondence has to do with content coverage. In other words, in addition to content relevance, test developers require to demonstrate that tests ensure a full representation of the constructs intended to be measured. Once these requirements are implemented, test takers' language ability and background knowledge can actively be engaged by the test input (test tasks).

1.3.4. Characteristics of LSP Tests

Douglas (2001) highlights three aspects that distinguish LSP tests from general purpose language tests: authenticity of task, specificity of content and interactiveness between language knowledge and specific content knowledge. The distinctiveness of these aspects in LSP testing will be explained in the following paragraphs.

1.3.4.1. Authenticity of LSP Tasks

Authenticity of task refers to the extent to which LSP test tasks:

> share critical features of tasks in the target specific purpose situation of interest to test takers. The intent…is to engage the test takers' language knowledge in carrying out the test task as far as possible in the same way it would be in responding to target situation (Douglas, 2000, p.46).

Bachman (1991) identifies two types of authenticity: situational and interactional. The former refers to "the perceived relevance of the test method characteristics to the features of a specific target language use situation" (p. 690). This implies that developing situationally authentic test tasks requires us to focus on the characteristics that they share with the target language use situations. This type of authenticity should not be confounded with the one developed by the 'Real-World Approach' which implements authenticity by "sampl[ing] actual tasks from a domain of nontest language use" (Bachman, 1991, p. 691) and incorporating them in test design. Situational authenticity does not sample tasks from real-world situations; suffice for it to design tasks that share significant characteristics with target contexts. Bachman points out that "language testers and teachers alike are concerned with this kind of authenticity, for we all want to do our best to make our teaching and testing relevant to our students' language use needs" (p. 791).

1.3.4.1.1. Interactional Authenticity

In addition to situational authenticity, interactional authenticity or 'authenticity of a task' is one of the main characteristics of LSP tests (Douglas, 2000, 2001). This type refers to the extent to which test takers' language and background knowledge are engaged in performing test tasks (Bachman, 1991;Douglas, 2010b, 1013; Widdowson, 1983, 2003). The incorporation of situational authenticity in LSP tests requires us to demonstrate that the specific test task characteristics correspond to the characteristics of the specific target language situations. However, the gauge of interactional authenticity in test tasks responds to "the extent to which the test taker is engaged in the task, by responding to the features of the target language use situation embodied in the test method characteristics" (Douglas, 2001, p. 47).

1.3.4.2. Specificity of Content

According to Douglas (2001), the specificity of content refers to the features which can:

> affect the level of specificity of a written or spoken text in an LSP test [such as] the amount of field specific vocabulary, … the rhetorical functions of various sections of the text, and the extent to which comprehension or production of the text required knowledge of subject specific concepts" (p. 46).

The implementation of specificity in LSP tests, requires us to conduct a statistical analysis on a range of specific target domains (needs analysis). Then, we decide what degree of specificity is to be included in the test. In explaining this point, Douglas (2001) inquires "is a specific language test for engineers good enough, or must we produce different tests for agricultural, automotive, chemical, civil, electrical, industrial, marine, mechanical, nuclear, and transportation engineers?" (p. 48). Douglas suggests that the notion of specificity of content should also be raised even within the same field of interest. The author stresses that even "within the field of mechanical engineering alone, for example, we might produce separate tests for those in combustion science, dynamics, fluid mechanics, metrology, micro-electromechanical systems, nanostructures, tribology, and thermal engineering" (p. 48). Now, the question that needs to be answered regarding engineering specialties in Algerian secondary education: is the design of one test for civil, electrical or mechanical engineering specialties good enough? Or we are required to design three different tests for each one of these sub-specialties.

1.3.4.3. Interaction between Background Knowledge and Language Knowledge

The interaction between background knowledge and language knowledge is the most important distinctive feature in LSP testing. This is because in general language tests, background knowledge is considered as one of the construct irrelevant variances which lead to measurement errors, and eventually affect the interpretation and uses of test scores (see Chap IV). Conversely, in LSP testing, the interaction between test takers' specific purpose background knowledge and the aspects of their language competence leads them to be engaged by the specific input of test tasks.

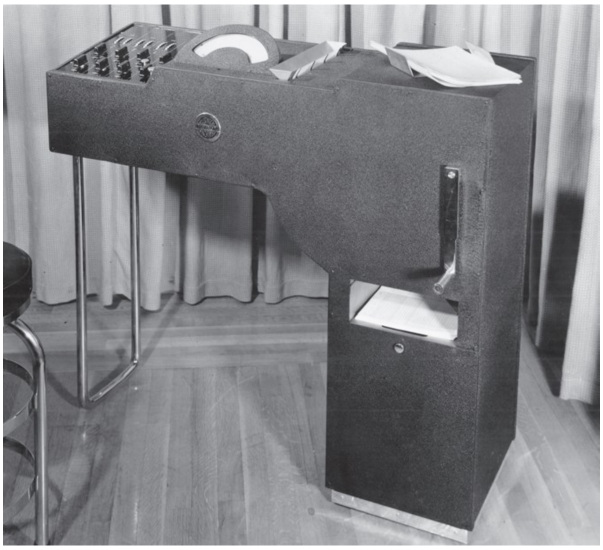
1.4. Computer Based Testing

Developments in information technology have almost affected all the fields of human life; and language testing is no exception. Since their introduction into the field, these devices have come to play a key role in scoring, item banking, test delivery, test construction, administration, and in making inferences about test takers' language abilities (Chappelle & Douglass, 2006; Douglas, 2000, 2010b). The use of machines in language testing dates back to 1935 when IBM 805 scoring machine became commercially available (Williamson, Bejar & Mislevy, 2006). This machine which was manufactured by 'Information Business Machines Company' (IBM) was first incorporated in multiple-choice scoring in the USA. Since then, more sophisticated devices have been introduced into the field to the point that every aspect in language testing is now affected in a way or the other by computer technology (Chappelle, 2003; Chapelle & Douglas, 2006; Douglas, 2000, 2010b; Fulcher,2003; 2010; Williamson, Bejar & Mislevy, 2006)

1.4.1. IBM 805 Scoring Machine

As we have mentioned previously, IBM 805 is a scoring machine which was manufactured by IBM Company in the mid-thirties (see Fig. 2). The machine was devised by a school teacher called Reynolds Johnson (Williamson, Bejar & Mislevy, 2006). When 'IBM' learned of his invention, he was hired by the company to develop the original version of the device. According to Fulcher (2010), this machine "could handle up to 150 multiple-choice items per sheet, and could score between 800 and 1000 test papers per hour, depending upon the skill of the operator"(p. 203). On their part, Chapelle and Douglas (2006) argue that IBM 805 has brought speed accuracy to objective scoring in that it "could score “objective” tests ten times faster than humans, and with greater accuracy. This concept is still in use today, essentially unchanged except with respect to advances in computer technology itself" (p. 34).

Fig 2: IBM 805 Scoring Machine



Source: Chapelle & Douglas, 2006, p. 34; Fulcher, 2010, p.203

1.4.2. Computer Assisted Language Testing

Computer assisted language testing 'CALT' can be defined as "an integrated procedure in which language performance is elicited and assessed with the help of a computer"(Noijion, 1994, p. 38). This procedure offers three main functions: test generation, interaction with test takers and the evaluation of their responses. Chapelle (2010, as cited in Suvorov & Hegelheimer, 2014) provides three reasons for using computers in language testing: efficiency, equivalency, and innovation. Efficiency can be achieved in CALT by reinforcing the criteria of validity, reliability, speed in administration and scoring. Equivalence "refers to research on making computerized tests equivalent to paper and pencil tests that are considered to be “the gold standard” in language testing" (p.

1). Innovation implies that the conceptualization of language ability has to be redefined whenever it is necessary. For example, in this information age, communicative competence "needs to be conceived in view of the joint role that language and technology play in the process of communication" (Chappelle & Douglas, 2006, p. 108). The authors, of course, refer to Rassool (1999) who thinks of CC to be referring to “the interactive process in which meanings are produced dynamically between information technology and the world in which we live” (p. 238).

1.4.3. Computer Adaptive Testing

Computer adaptive testing 'CAT' "requires a digital computer to present each test item, score each response, and then select the next item that will be most appropriate for the candidate" (Green, Bock, Humphreys, Linn & Reckase, 1984, p.347). In CAT, test takers are first presented with a task of medium difficulty. The ones who perform it correctly are presented with another one of greater difficulty; conversely, the examinees who fail to do the first task correctly are presented with one of less difficulty; and the process goes on this way. "Eventually", as Douglas (2000) remarks, "the computer gets a fix on the test taker's ability level and presents only items at that level until predetermined degree of reliability has been achieved, and the test ends" (p.269).

1.4.4. Advantages of the CALT

According to the proponents of the use of technology in language testing, CALT offers more advantages over paper and pencil tests (Chappelle, 2003, 2010). These include measuring the time taken by test takers for doing a given assignment; recording their route through the test; facilitating their easy access to the large amount of stored information; fast psychometric calculations allowing for balanced difficulty indices; providing a variety of multimedia options for test takers; and last but not least, reinforcing the quality of standardization through identical administration procedures (Noijion, 1994). The main advantages of the 'CALT' are listed in Fig.4.

Fig 4.The Main Advantages of the CALT

1. The computer has the ability to measure time. The time which a learner takes to complete a task or even the time taken on different parts of a task, can be measured, controlled and recorded by computer.
2. The computer has the ability to record information about the testee' s routes through the test.
3. The computer can present information in a variety of ways.
4. The computer can provide quick and easy access to a variety of different types of information.
5. The computer can be linked to other equipment. This can allow different types of input and presentation.
6. The computer can encourage the learner's own strategies for evaluation. In particular the information which a computer can collate and present about test performance could help the learner to feel that his own opinions are of importance.
7. The computer can make use of language rules. a. At a relatively simple level the computer can do a spelling check on the learner's text. b. Parsers of varying degrees of sophistication can be used not only to check for syntactic errors in the learner's text, but to provide "communicative" tests as well

Source: Alderson, 1990 as cited in Chapelle and Douglas, 2006, p. 11

In addition to the advantages stated above, language testers raise some concerns on the application of technology in language testing. We can, for example, mention the lack of expertise in computer literacy. This is because 'CALT' requires three types of expertise: the knowledge of the ability to be tested; expertise in testing; and literacy in information technology. Additionally, the multiple-choice item type in CALT is of limited scope; and automatic scoring for constructed answers is not always accurate. Moreover, there are concerns that computers do not focus on measuring the same language ability; instead, they are adapted to measure different ranges of language abilities that individual test takers have.

1.5. Frame of Reference

The question of score interpretation has been a topic of debate amongst measurement specialists and language testers (Popham, 2004). Their divergence is on whether test scores need to be interpreted in relation to the performance of a given group of test takers with respect to the performance of another group taking the same test, or with respect to 'an established standard'. Due to the importance of both views, measurement specialists distinguish two broad frames of reference: criterion-referenced tests (CRTs) and norm-referenced testing (NRTs).

1.5.1. Norm-referenced Tests

'NRTs' refer to the instruments which language testers use "to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device" (Popham & Husk, 1969, p. 2). In these measures, test takers' scores are interpreted with reference to a norm group which refers to a large a sample of individuals who share the same characteristics with the students for whom the test is intended. The norm group is usually given the test and the norms (how students actually do perform) "of this group's performance are used as reference points for interpreting the performance of other students who take the test" (Bachman, 1990, p. 72). Measurement practices in these tests relate test analyses to the mean, the median, the standard deviation, and/or percentile rank (Brown, 1996; Ebel & Frisbie, 1991; Miller, Linn & Gronlund, 2009).

1.5.2. Criterion-referenced Testing

1.5.2.1. Historical Perspectives

The early sixties witnessed new educational practices in the field of teaching and evaluation. At the level of teaching, a new term 'instructional technology' was introduced into the jargon of the American armed forces schools (Cartier, 1968; Glaser, 1963; Glaser & Cox, 1968; Glaser & Klaus, 1962; Lindvall & Nitko, 1969). The use of technology instruction was intended to provide the military with highly skilled recruits in specialties such as "jet engine mechanics, supply clerks, and cryptographic technicians"(Cartier, 1968,

p. 27). Unlike traditional methods of evaluation which used to interpret the meaning of test takers' scores with reference to the results obtained by their colleagues on the same measures, instructional technology focused only on what learners have actually achieved as a result of an instructional syllabus. In an article published in the *'Journal of Educational Measurement'*, Cartier (1968) documents for these changes:

> The term instructional technology was introduced into the professional jargon of the Air Training Command and, within a year or two, could be seen in Army and Navy training publications as well. The term was an outgrowth of programmed instruction, but has grown to have a far greater breadth of application and perhaps represents an even more fundamental change of instructional philosophy than programming. Its most important ramifications, in fact, have little to do with instructional media or methods, but more with determination of course objectives and with evaluation of whether the students have, in fact, achieved those objectives. Instructional technologist is not interested in how well one student compares with the class mean score (the norm) at graduation, but solely in whether each individual student can demonstrate the ability to perform each and every one of the essential job behaviors (the criteria)…Students are differentiated from each other only by the amount of instruction they need in order to pass. (pp. 27 & 28).

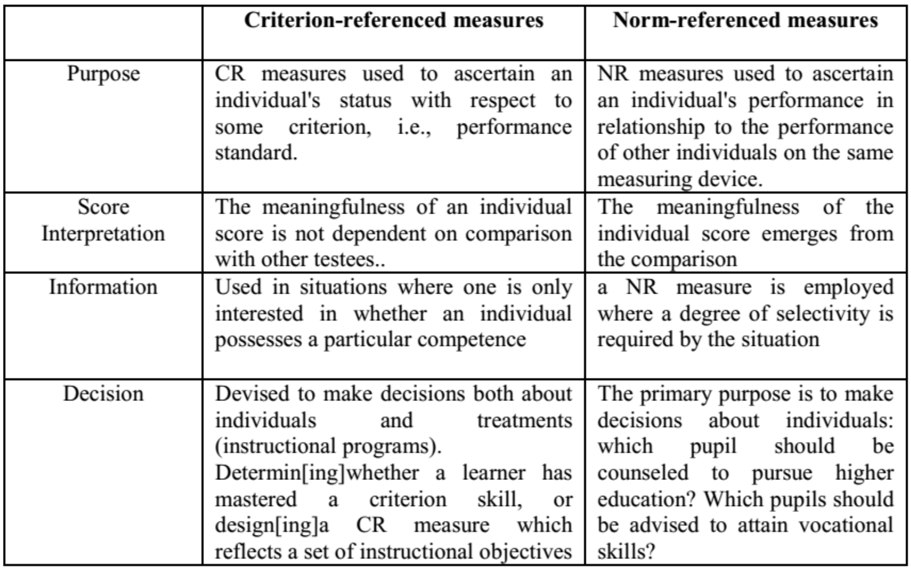
1.5.2.2. Definition of Criterion-referenced Tests

In a seminal article published in the *'American Psychologist journal'* in 1963, Robert Glaser described the new measurement practices "which assess students achievement in terms of a criterion standard providing information as to the degree of competence attained by a particular student which is independent of reference to the performance of others" (p. 519) as criterion-referenced tests. Such a test can be defined as the "one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards. Performance standards are generally specified by defining a class, or domain of tasks that should be performed by the individual" (Glaser & Nitko, 1970, pp. 57-8). The authors emphasize that tasks in criterion-referenced tests need to be sampled in a way to ensure content coverage and representation. On their part, Popham and Husek (1969) consider these measures as the ones "which are used to ascertain an individual's status with respect to some criterion, i.e., performance standard" (p. 2). According to them, the reason for describing these tests as criterion-referenced lies in the fact that test takers are compared to some specified criterion rather than to the scores obtained by their colleagues on the same test. In these measures, Popham and Husek point out that "the meaningfulness of an individual score is not dependent on comparison with other testees. We want to know what the individual can do, not how he stands in comparison to others" (p.2).

1.5.3. Distinction between Norm-referenced and Criterion-referenced Tests

In addition to the features of variability, reliability, validity and item construction and analysis, norm and criterion-referenced tests can be distinguished, as included in Table 2, with reference to the purpose for which the test is developed; the manner in which it is designed; the type of information we are interested in; the meaningfulness of its scores; and the decisions we intend to make (Glaser, 1963, 1969; Glaser & Nitko, 1970; Popham & Husek, 1969).

Table 2. Basic Distinction between Criterion and Norm-referenced Tests



	Criterion-referenced measures	Norm-referenced measures
Purpose	CR measures used to ascertain an individual's status with respect to some criterion, i.e., performance standard.	NR measures used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device.
Score Interpretation	The meaningfulness of an individual score is not dependent on comparison with other testees..	The meaningfulness of the individual score emerges from the comparison
Information	Used in situations where one is only interested in whether an individual possesses a particular competence	a NR measure is employed where a degree of selectivity is required by the situation
Decision	Devised to make decisions both about individuals and treatments (instructional programs). Determin[ing]whether a learner has mastered a criterion skill, or design[ing]a CR measure which reflects a set of instructional objectives	The primary purpose is to make decisions about individuals: which pupil should be counseled to pursue higher education? Which pupils should be advised to attain vocational skills?

Source: Bachman, 1990, p 72: Popham and Husek, 1969, pp. 2&3

1.6. Types of Tests

The classification of language tests fall under five broad types: achievement, prognostic, placement, diagnostic, and proficiency tests. This classification is based on a number of considerations such as the purpose for which tests are intended, the type of information they provide and the decisions to be made (Bachman, 1990, 1991; Brown, 1996; Ebel & Frisbie; 1991; Gronlund,1987).

1.6.1. Achievement Tests

According to Henning (1987), achievement tests are instruments which enable us to "measures the extent of learning in a prescribed content domain, often in accordance with explicitly stated objectives of a learning program" (p. 6). On its part, the '*Dictionary of Language Testing*' edited by Davies, Brown, Elder, Hill, Lumley and McNamara (1999) defines an achievement test as a tool that is "designed to measure what a person has learned within a given time. It is based on a clear and public indication of the instruction that has been given"(p.2). These definitions and others imply that these tests are tightly linked to the formal instructional syllabus for the information they provide on what students have learnt; and on the appropriateness of the test content to the stated objectives.

Language testers distinguish two types of achievement tests: final or standardized achievement tests and progress tests (Gronlund, 1987). Final achievement tests "are those administered at the end of a course of study. They may be written, or administered by the ministries of education, official examining boards, or by members of teaching institutions" (Hughes, 2003, p. 13). This type is also called 'standardized achievement test' which is administered in large scale examinations (Ebel & Frisbie; 1991; Gronlund,1987). The difference between these tests and standardized tests is the frame of reference. In other words, the question is whether to interpret the meaning of students' scores with reference to the scores of norm groups, or to relate them to a given criterion. The second type (progress tests) refers to the measures which inform us of "the progress that students are making" (Hughes, 2003, p. 13) during the course of their school year.

1.6.2. Placement Tests

This type of test is designed for assessing learners' levels of language abilities so that they can be placed in the appropriate course of study. The content of these tests is not necessarily based on the syllabus taught in the host institutions. The main concern of these measures is to elicit information about the extent to which "students possess the skills and abilities that are needed to begin instruction" (Gronlund, 1987, p. 2). The most famous type of these tests in the Algerian universities, is the one administered at the beginning of every academic year to join intensive foreign language courses so as to place applicants in the appropriate instructional level. What is worth mentioning in this context is that placement tests can also be used "to differentiate students who are ready for instruction from those who are not" (Bachman, 1990, p. 60). In this case, we can speak of 'readiness tests'.

1.6.3. Diagnostic Tests

Unlike placement tests which are held at the beginning of instruction, diagnostic tests are given during the instructional program for the purpose of identifying learners' areas of strengths and weaknesses. Information from these tests can be used for finding corrective solutions to the pupils' learning difficulties. For example, if a large number of pupils fail in a given exam, adjustments can be made in the programs of study or the teaching methods. Conversely, if the scores reveal that only a small number of learners have experienced learning failures, these test takers will be invited for additional lessons which concentrate on specific elements of language rather than on integrated skills (Alderson, 2005; Brown, 2006; Gronlund, 1977; Knoch 2009).

1.6.4. Prognostic (Aptitude) Tests

As their name imply, prognostic tests are "designed to measure students' probable performance in a foreign language which he or she has not started to learn: i.e. it assesses aptitude for learning a language" (Heaton, 1988, p. 173). The use of prognostic tests in formal education dates back to the early 1920's in the USA. At that time, Egalitarian principles inspired from the French Revolution were still prevailing in the American society (Spolsky, 1995). These principles required "that everyone should have the right of access to a high-school education, including foreign-language classes that were offered in them. [However] the tiny amount of time allocated in the USA school curriculum to language study led to a distressingly high failure rate" (Spolsky, 1995, p. 324). Faced with waves of unqualified students "who had been admitted to…classes through a policy of mass education" (p. 324), the educational authorities asked psychological and language testers to develop tests which can screen the free access to foreign language classes.

Screening efforts had first been implemented with the use of intelligence tests which intended to predict students' potentials of learning. Then, the responsibility of designing prognostic tests shifted from the psychologists to become the business of linguists who emphasized that the 'facility' of learning a foreign language is "a fairly specialized talent (or group of talents)' independent of the traits included under 'intelligence' " (Carroll, 1960, as cited in Spolsky, 1999, p. 334).

1.6.5. Proficiency Tests

Unlike prognostic tests which attempt to measure learners' specific language items, with reference to a given instructional program, proficiency tests are concerned with measuring individuals' global level of language ability regardless of the program of study or instruction they might have covered during their formal education. The intent here is to decide whether candidates are really proficient in a given language program so that they would be liable for a given occupational or instructional position. Yet, there are proficiency tests, such as the 'Test of English as a Foreign Language' (TOFEL), the 'Cambridge First Certificate' in English (FCE), or the 'Cambridge Certificate of Proficiency' (CPC) which "do not have any occupation or course of study in mind" (Hughes, 2003, p. 12).

Proficiency tests are also used as gate-keeping instruments to control the flow of immigrants and asylum seekers (Shohamy, 2001, 2008). The USA, Britain, the Netherlands, and Australia for instance, require immigrants to pass proficiency tests "as a condition for obtaining entry to the country and/or to residency and ultimately citizenship" (Shohamy & McNamara, 2009, p.1). The implementation of this policy emerges from "the belief that language proficiency, as exemplified through these language tests, is an expression of loyalty and patriotism and should be a requirement for residency, and especially citizenship" (Shohamy, 2007, p. 149).

Conclusion

In its development, the field of language testing attempted to answer three main questions: what to test? How to test it? And who is qualified to test it? In the pre-scientific stage, the testing practices were traditional, intuitive, and unscientific. The psychometric-structuralist stage witnessed the cooperation between two types of experts: structural linguists and psychometricians. The first type considered language ability to be consisting of discrete elements each of which can be measured as a separate universe; and the second type introduced to the field notions about objectivity, dependability, reliability, and validity. In the integrative-sociolinguistic stage, two other categories of linguists dominated the field: the 'LCT' and the 'CCT". Each group viewed language testing as their field of specialty. The former hypothesized that language proficiency is made up of single unitary ability; and its discrete components need to be measured within integrated skills. The latter accepted the fact that the language ability to be tested is undividable; but insisted on testing it within the aspects of language use. Since the eighties, the field of language testing has been dominated by the views of the 'CCT' which conceptualizes language ability to be consisting of various sectors each of which can be measured separately. This era witnessed the flourishing of communicative language testing, task-based, competency-based and ESP testing. However, with the incorporation of computer assisted testing, the field seems, once again, to be shared between applied linguists, and information technologists.

Chapter Two

Constructional Constituents of Language Tests

Chapter Two

Constructional Constituents of Language Tests

Introduction

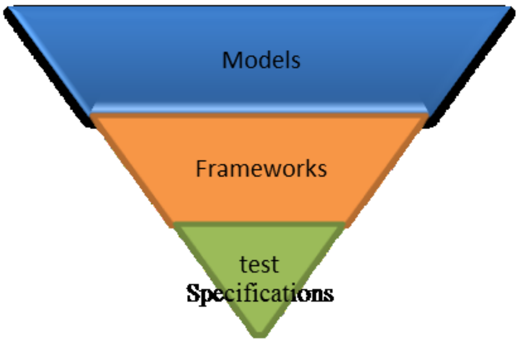
The construction of tests is similar to the construction of buildings. Both tests and buildings undergo similar constructional processes, such as design, structure, development, and use. Buildings are primarily constructed for special types of occupants and users. In the same way, tests are designed for special categories of test takers. However, both tests and buildings are sometimes used for purposes which are not designed for (Davidson, & Lynch, 2002; Fulcher, 2010; Fulcher & Davidson, 2007, 2009, 2012).

Test development builds upon two main factors: architectural layers and developmental processes (Alderson, 2000a; Bachman, 2007; Fulcher, & Davidson, 2007, 2009; Fulcher, 2010). When architects decide to design buildings they normally "choose the materials they intend to use in construction" (Fulcher, & Davidson, 2009, p. 123), such as cement, gravel, sand, metal and so on. In the same way, when test developers decide to design a test, they should also take all the material which can help them in building these measures into consideration.

2.1. Constructional Layers of Test Design

Language testers classify the 'material', which they use in test construction into three main architectural layers: models of language ability, test frameworks, and test specifications (see Fig. 3). According to Fulcher and Davidson (2007), the models, which stretch on top of the reversed pyramid, refer to the "abstract theoretical descriptions of what it means to be able to communicate in a second language"(p.36). The layer, which mediates between the models and the specifications, concerns test frameworks. The latter select the constructs from the models and lay them out in the form of target language use tasks for particular testing situations. The third layer refers to test specifications. These "tell us the nuts and bolts of how to phrase the test items, how to structure the test layout, how to locate the passages, and how to make a host of difficult choices as we prepare test materials"(Fulcher & Davidson, 2007, p.3).

Fig 3: Constructional Layers of Test Design



Source: Fulcher & Davidson, 2007, p. 103; Fulcher & Davidson, 2009, p. 127

2.2. Models of Language Ability

Except for the pre-scientific stage where the traditional testing practices had been the dominant, test design has always been built around a full description of the language ability to be tested (Purpura, 2008). According to Alderson et al., (1995) this ability refers to "some abstract belief of what language is, what language proficiency consists of, what language learning involves, and what language users do with language" (p.16). Language testers emphasize that even if tests are intended to measure a very narrow scope of language, they should be informed with a detailed description of the theory of language in question (Alderson, 2000a; Alderson, et al., 1995; Bachman, 1990, 2007; Davidson & Lynch, 2002; Fulcher, 2010). Echoing this point of view, Bachman and Palmer (1996) state that:

> We believe strongly that the consideration of language ability in its totality needs to inform the development and use of any language test....We recognize that many of the language tests we develop will focus on only one or a few of those areas of language knowledge. Nevertheless, we believe that there is a need to be aware of the full range of components of language ability as we design and develop language tests and interpret language test scores. For example, even though we may be only interested in measuring an individual's knowledge of vocabulary, the kind of test items, tasks, or texts used need to be selected with an awareness of what other components of language knowledge may evoke. We believe, therefore, that the design of every language test, no matter how narrow its focus, should be informed by a broad view of language ability (p. 67).

Since the early sixties, a number of models which attempted to describe language ability have been proposed in the literature. We can, for example, state Lado's model which considers language ability to be composed of elements and skills (Lado, 1961). According to Lado, the constructs that need to be tested are: pronunciation, grammar and the lexicon. This model was later overshadowed by the 'Unitary Competence Hypothesis' (UCH) trend, which hypothesized that language ability represents an overall unity in which language forms can be tested within integrated skills as its constructs (Carroll, 1961, 1964, 1968; Oller, 1979). The 'UCH' trend has, in its turn, been challenged by models of communicative competence, hypothesizing that language ability is multi-componential and each of its components can be tested as a separate universe (Bachman, 1990, 1991; Bachman & Palmer, 1996; Canale & Swain, 1980; Hymes; 1972; Mumby, 1978; Savignon, 1972, 2002).

2.3. Models of Communicative Competence

This section accounts for four models of language ability: Hymes' model (1972), Canale and Swain's model (1980), Bachman and Palmer's model (1996), and Douglas's model (2000). We start with the founder of sociolinguistics 'Dell Hymes' as the first linguist who proposed a model for communicative competence (CC) consisting of four interacting sectors. Then, we move to Canale and Swain (1980) as the first linguists who developed Hymes' model and set the procedures for linking the components of 'CC' with language testing. Later on, we move to Bachman and Palmer (1996) to introduce their most influential and comprehensive model communicative of language ability (CLA) accounting for language knowledge and test method facets; finally, we conclude with Douglas (2000) as the first linguist who proposed a theoretical model for testing specific purpose language ability (SPLA).

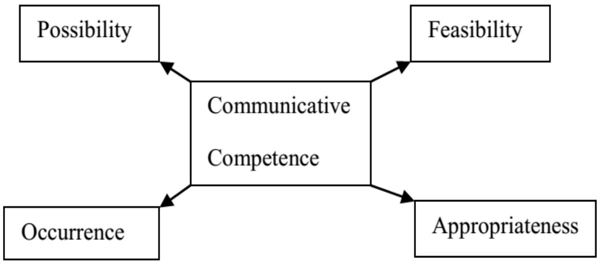
2.3.1. Hymes' Model

According to Dell Hymes (1972), communicative competence (CC) can be defined as the ability of using language not only grammatically but appropriately as well. According to him, this can be manifested during the process of language acquisition when a child starts acquiring:

> competence as to when to speak, when not, and as to what to talk about with whom, when, where, in what manner. In short, a child becomes able to accomplish a repertoire of speech acts, to take part in speech events, and to evaluate their accomplishment by others" (p. 60).

Communicative competence is, as the author emphasizes, "fed by social experience, needs, and motives… [and is] integral with attitudes, values, and motivation concerning language, its features and uses" (p. 20). Hymes' model of 'CC' is made up of four interacting sectors (see Fig 4): possibility, feasibility, appropriateness, and occurrences (Hymes, 1972).

Fig 4: Hymes' Sectors of Language Ability



Adapted from Hymes 1972, p. 63.

The first sector refers to the knowledge of language rules, such as phonology, morphology, syntax and the lexicon. It examines the extent to which communication conforms to the rules of grammar. The second component 'feasibility' refers to the psycholinguistic factors which can affect the human information processing such as "memory limitation, perceptual device, effects of properties such as nesting, embedding, branching, and the like" (p. 67). The third component is appropriateness. The latter examines the extent to which utterances are appropriate to contextual features. The fourth component 'occurrence' refers to the extent to which utterances do really occur in the linguistic repertoire of a given speech community. This means that the probability of occurrences of some utterances can be rare. For example, "saying may 'God be with you' instead of good-bye or bye-bye in ending a routine telephone conversation [may be] rare in a particular community or situation" (Canale & Swain, 1980, p. 16). The components of Hymes' model of 'CC' are summarized in Fig .5:

Fig 5: The Components of Hymes' Model of communicative Competence

1- Whether (and to what degree) something is formally possible;

2- Whether (and to what degree) something is feasible in virtue of the means of implementation available;

3- Whether (and to what degree) something is appropriate (adequate, happy, successful) in relation to a context in which it is used and evaluated;

4- Whether (and to what degree) something is in fact done, actually performed, and what its doing entails.

Source: Hymes, 1972, p. 63

2.3.2. Canale and Swain's Model

Building upon Hymes, 1972; Savignon, 1972; 1976; Widdowson, 1978, 1979; Wilkins, 1976; Morrow, 1977; Candlin; 1978; Mumby, 1978, Canale and Swain (1980) define communicative competence as the

> one in which there is a synthesis of knowledge of basic grammatical principles, knowledge of how language is used in social contexts to perform communicative functions, and knowledge of how utterances and communicative functions can be combined according to the principles of discourse" (p. 20).

Following this definition and as shown in Fig 7, the authors divide their model of 'CC' into three interrelated competencies**:** grammatical competence, sociolinguistic competence and strategic competence (Fulcher, 2010; Fulcher & Davidson, 2007, 2009; McNamara, 1996; Purpura, 2008). Grammatical competence refers to the "knowledge of lexical items and of rules of morphology, syntax, sentence- grammar semantics, and phonology" (Canale and Swain, 1980, p. 29). This competence provides "learners with the knowledge of how to determine and express accurately the literal meaning of utterances "(p. 30). Sociolinguistic competence, according the authors is "crucial in interpreting utterances for social meaning, particularly when there is a low level of transparency between the literal meaning of an utterance and the speaker's intention" (p. 30). This competency is, in its turn, divided into two sets of rules: sociocultural rules of use and rules of discourse. The first set of rules specifies "the ways in which utterances are produced and understood appropriately with respect to the components of communicative events outlined by Hymes (1967, 1968)" (p.30). Canale and Swain specify two roles for sociocultural rules. The chief role focuses "on the extent to which certain propositions and communicative functions are appropriate within a given sociocultural context depending on contextual factors such as topic, role of participants, setting, and norms of interaction" (p. 30). The other role concerns "the extent to which appropriate attitude and register or style are conveyed by a particular grammatical form within a given sociocultural context"(p. 30). Concerning the second constituent 'the rules of discourse', the authors define it with reference to cohesion (grammatical links) and coherence (appropriate combination of communicative functions of groups of utterances). Because of the lack of extensive literature concerning discourse at that time, the authors argue that "it is not altogether clear to us that rules of discourse will differ substantively from grammatical rules (with respect to cohesion) and sociocultural rules (with respect to coherence)" (p.30 [parentheses in original]). This is why they limit the scope of discourse rules in their model to the "combination of utterances and communicative functions and not the grammatical well-formedness of a single utterance nor the sociocultural appropriateness of a set of propositions and communicative functions in a given context" (p.30).

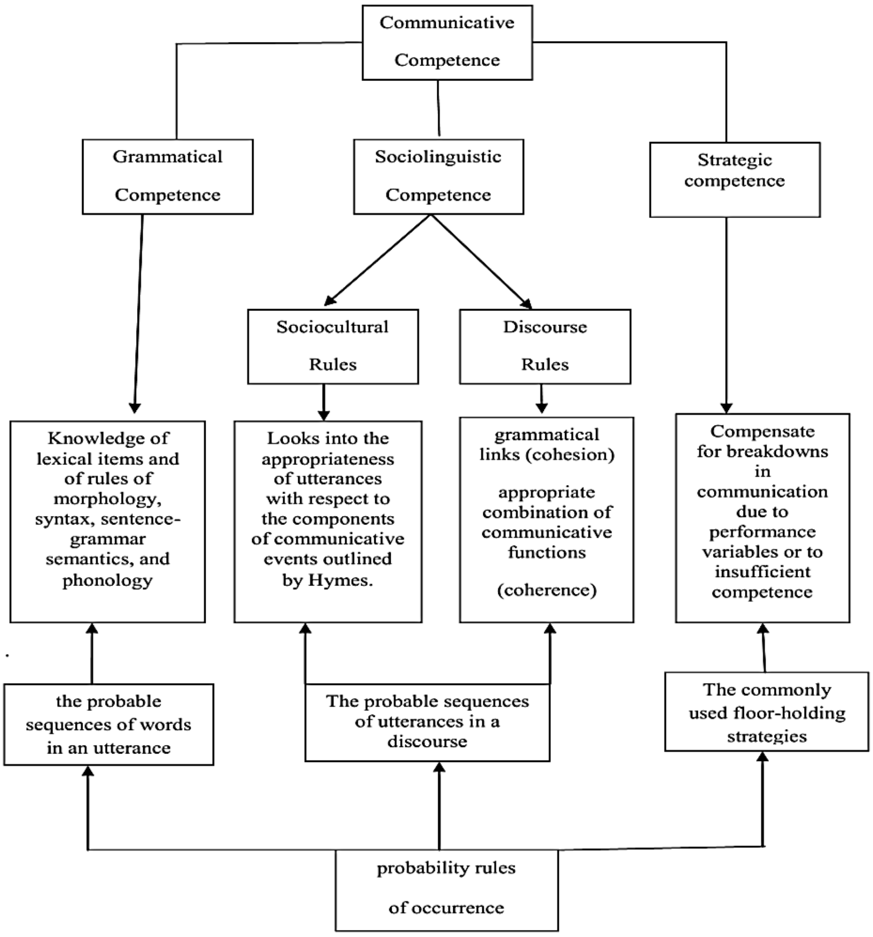
2.3.2.1. Strategic Competence

According to Canale and Swain (1980), strategic competence refers to the verbal and nonverbal communication strategies that language users employ in order to compensate for breakdowns in communication "due to performance variables or to insufficient competence" (p. 30). The authors distinguish two types of strategies: one type is related to grammatical competence; and the other is related to sociolinguistic competence. The first type is called up when language users perceive that they do not master a given language form. So, in order to "keep the communicative channel open" (p. 30), they resort to paraphrasing. The second type is employed in the case of addressing to strangers that one is not sure of their social status.

2.3.2.2 The Probability of Occurrences

Unlike Dell Hymes (1972) who considers the probability of occurrences as an independent sector of his 'CC', Canale and Swain (1980) take it as a subcomponent that exists within each component of their model. According to them, the probability of rules focuses on the extent to which we are aware of the "relative frequencies of occurrence that a native speaker has with respect to grammatical competence (the probable sequences of words in an utterance) sociolinguistic competence (the probable sequences of utterances in a discourse), and strategic competence" (p. 31[parentheses in original]).

Fig 6: Canale and Swain's Framework of Communicative Competence

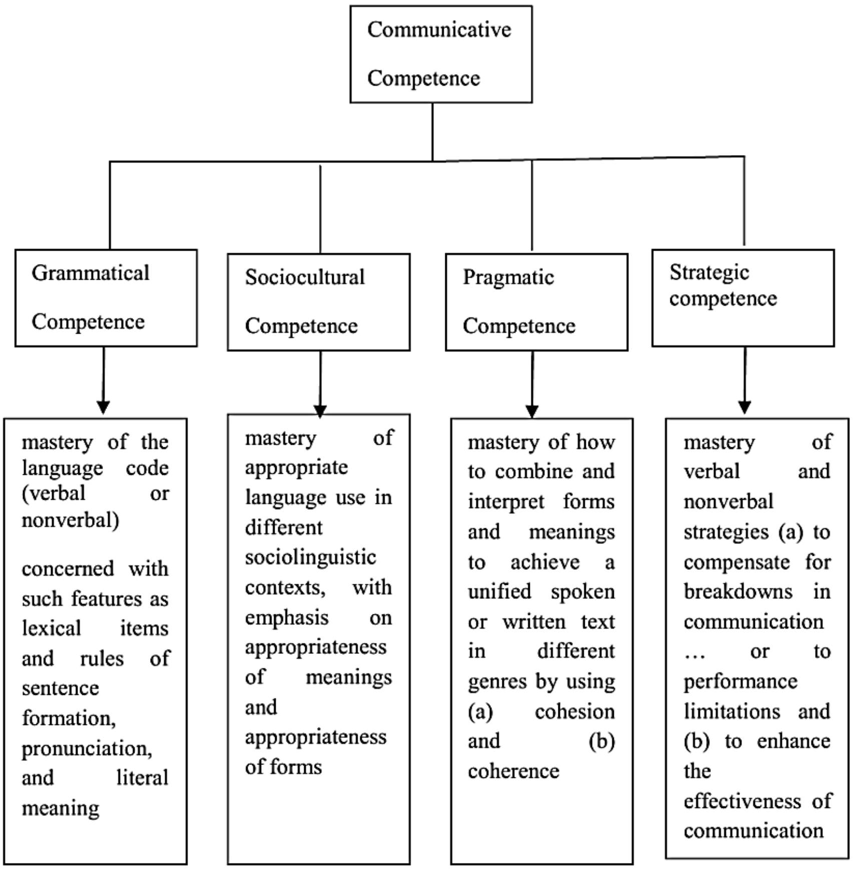


Communicative Competence
Grammatical Competence
Sociolinguistic Competence
Strategic competence
Sociocultural Rules
Discourse Rules
Knowledge of lexical items and of rules of morphology, syntax, sentence-grammar semantics, and phonology
Looks into the appropriateness of utterances with respect to the components of communicative events outlined by Hymes.
grammatical links (cohesion)
appropriate combination of communicative functions
(coherence)
Compensate for breakdowns in communication due to performance variables or to insufficient competence
the probable sequences of words in an utterance
The probable sequences of utterances in a discourse
The commonly used floor-holding strategies
probability rules of occurrence

Organized from Canale and Swain, 1980, pp. 29&30

In a 1983, M. Canale introduced some modifications on Canale and Swain's model, and reorganized it into four constituents (see Fig.7): grammatical competence, sociocultural competence, pragmatic competence, and strategic competence. In this modification, Canale promoted sociocultural and discourse rules to independent competencies. At the same time, he extended the scope of strategic competence to include the purpose of enhancing 'the effectiveness of communication'

Fig 7: Canale's (1983) Modifications on Canale and Swain's Model

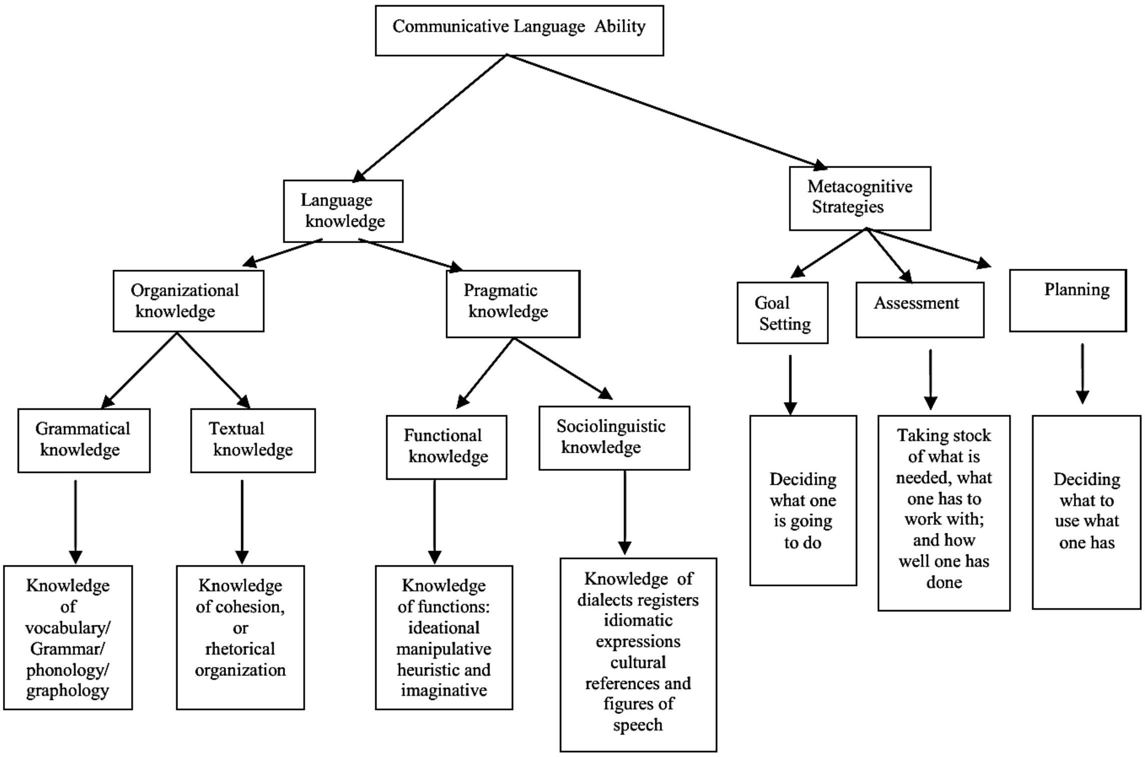

Communicative Competence
Grammatical Competence
Sociocultural Competence
Pragmatic Competence
Strategic competence
mastery of the language code (verbal or nonverbal)
concerned with such features as lexical items and rules of sentence formation, pronunciation, and literal meaning
mastery of appropriate language use in different sociolinguistic contexts, with emphasis on appropriateness of meanings and appropriateness of forms
mastery of how to combine and interpret forms and meanings to achieve a unified spoken or written text in different genres by using (a) cohesion and (b) coherence
mastery of verbal and nonverbal strategies (a) to compensate for breakdowns in communication … or to performance limitations and (b) to enhance the effectiveness of communication

Organized from Canale, 1984, p. 112

2.3.3. Bachman and Palmer's Model

Bachman and Palmer's model (1996) of communicative language ability (CLA) consists of two broad components (see Fig. 8): language knowledge (language competence) and metacognitive strategies (strategic competence). The combination of these constituents, according to the authors "provide[s] language users with the ability, or capacity, to create and interpret discourse, either in responding to tasks on language tests or in non-test language situations" (p. 67).

Fig 8: Components of Bachman and Palmer' s Communicative Language Ability



Communicative Language Ability
Language knowledge
Metacognitive Strategies
Organizational knowledge
Pragmatic knowledge
Goal Setting
Assessment
Planning
Grammatical knowledge
Textual knowledge
Functional knowledge
Sociolinguistic knowledge
Deciding what one is going to do
Taking stock of what is needed, what one has to work with; and how well one has done
Deciding what to use what one has
Knowledge of vocabulary/ Grammar/ phonology/ graphology
Knowledge of cohesion, or rhetorical organization
Knowledge of functions: ideational manipulative heuristic and imaginative
Knowledge of dialects registers idiomatic expressions cultural references and figures of speech

Organized from Bachman and Palmer, 1996, pp.66-8 , 71

2.3.3.1 Language Knowledge

Language knowledge can be defined "as a domain of information in memory that is available for use by the metacognitive strategies in creating and interpreting discourse in language use" (Bachman & Palmer, 1966, p. 67). Language competence is organized into two broad sections: organizational knowledge and pragmatic knowledge. The former "comprises those abilities involved in controlling the formal structure of language for producing or recognizing grammatically correct sentences, comprehending their propositional content, and ordering them to form texts" (Bachman, 1990, p 78). Pragmatic knowledge concerns the ability "to create or interpret discourse by relating utterances or sentences and texts to their meanings, to the intentions of language users, and to relevant characteristics of the language use setting" (Bachman & palmer, 1996, p. 69).

2.3.3.2. Organizational Knowledge

Bachman and Palmer (1996) rearrange organizational knowledge into two competencies: grammatical knowledge and textual knowledge. The former, which includes the knowledge of vocabulary, syntax, phonology, and graphology "is involved in producing or comprehending formally accurate utterances or sentences" (Bachman & palmer, 1996, p. 68). Textual knowledge (discourse competence) concerns what Savignon (2002) labels as "the interconnectedness of a series of utterances or written words or phrases to form a text, a meaningful whole"(p. 9). In textual knowledge, there exist two areas of competence: cohesion and knowledge of rhetorical or conversational organization. The first type is related to the extent to which the 'explicitly marked relationships' among sentences or utterances can affect the production or comprehension of texts or conversations, such as "reference, sub-situation, ellipsis, conjunction, and lexical cohesion as well as conventions such as those governing the ordering of old and new information in discourse" (Bachman, 1990, p.88). However, rhetorical organization includes methods of text development such as narrating, describing, comparing and/or contrasting and classifying; or organizational development of paragraphs, texts, or essays such as topic sentences, supporting sentences or transitional shift to new paragraphs (Bachman, 1990; Bachman & Palmer, 1996).

2.3.3.3. Pragmatic Knowledge

Pragmatic knowledge can be defined as "the acceptability of utterances within specific contexts of language use, and rules determining the successful use of language within specified contexts" (Fulcher and Davidson, 2007, p. 44). This competence examines the extent to which utterances are accepted by other language users in relation to specific settings "these act in a situation, and formulate the conditions stipulating which utterances are successful in which situations" (Van Dick, 1977, p. 190). According to Bachman (1990), pragmatic knowledge focuses on the relationship between utterances and the functions that language users intend to achieve by means of these utterances. Bachman and Palmer (1996) organize pragmatic knowledge into two categories: functional and sociolinguistic knowledge.

2.3.3.3.1. Functional Knowledge

Functional knowledge which "enables us to interpret the relationship between utterances or sentences and texts and the intentions of language users" (Bachman & Palmer, 1996, p. 69) requires the knowledge of four categories of functions: ideational, manipulative, instrumental, and imaginative functions (Bachman, 1990, 1991; Bachman and Palmer; 1996).

2.3.3.3.1.1. Ideational Function

This function enables language users “to express or interpret meaning in terms of [their] experience of the real world” (Bachman & Palmer, 1996, p. 69). Performing such functions can take place when people express or exchange their ideas or feelings about a given topic of interest. The utterances that generally express these functions include expressing ones' feelings, descriptions, explanations, classifications, and so on.

2.3.3.3.1.2. Manipulative Functions

This type of functions which enables language users to affect the world around them falls into three categories: instrumental, regulatory and interpersonal (Bachman, 1990; Bachman and Palmer, 1996).

2.3.3.3.1.2.1 Instrumental Functions

Instrumental functions can be classified into two types. One type is used to get other people do things for us. These functions include, for example, requests, suggests, commands and warnings. The other type is used when people express their intention of doing something such as offers, promises, or threats.

2.3.3.3.1.2.2. Regulatory Functions

These functions are used 'to control the behavior of others ' (Halliday, 1973, p. 18**)** with reference to the force of law, the regulations or the social norms. Examples of such functions include prohibitions and obligations.

2.3.3.3.1.2.3. Interpersonal (interactional) Functions

This type of illocutionary knowledge enables us to establish, maintain, change, or break interpersonal relationships when we meet other people. This function includes salutations, giving permission to engage in doing something, leave taking, compliments, or apologies**.**

2.3.3.3.1.3. Heuristic Functions.

This type enables language users to extend their knowledge of the world around them. It includes the use of language for teaching and learning, for memorizing and retaining information such as rules, words, formulae, or for problem solving.

2.3.3.3.1.4. Imaginative Functions

This type "enables us to use language to create an imaginary world or extend the world around us for humorous or esthetic purposes" (Bachman & palmer, 1996, p. 69). The use of figurative or literary language such as in novels, plays, poetry enables us to create or extend our own environment for humorous or esthetic purposes, where the value derives from the way in which the language itself is used.

2.3.3.3.2. Sociolinguistic Knowledge

The second component of pragmatic competence concerns sociolinguistic knowledge. According to Bachman and Palmer (1996) this competency:

> enables us to create or interpret language that is appropriate to a particular language setting. This includes knowledge of the conventions that determine the appropriate use of dialects or varieties, registers, natural or idiomatic, expression, cultural references, and figures of speech (p.70).

Sociolinguistic knowledge is concerned with the extent to which utterances, language functions or the intention of language users are appropriate to a given social context. The features which enable us to use functional knowledge in appropriateness with the social context include "sensitivity to differences in dialect or variety, to differences in register and to naturalness, and the ability to interpret cultural references and figures of speech" (Bachman, 1990, p. 95).

2.3.3.3.2.1. Sensitivity to Differences in Dialect or Variety

One of the abilities, which enable us to express and interpret utterances appropriately, is the awareness of the conventions that govern social or regional differences in language use contexts. For, example, in a context when one is required to use standard language, the use of geographical or social dialects will be inappropriate.

Conversely, the use of standard language sounds pretentious in contexts requiring the use of dialects (Bachman, 1990).

2.3.3.3.2.2. Sensitivity to Differences in Register

According to Halliday, McIntosh, and Strevens (1964 as cited in Bachman, 1990) "the term 'register'…refer[s] to variation in language use within a single dialect or variety" (p.95). These linguists distinguish three aspects of register ''field of discourse', 'mode of discourse', and 'style of discourse'. The field of discourse or discourse domain as Swales (1987) calls it, refers to the subject matter of language use, or to the terms we use in a given domain, such as the register used in medical fields, in sports or at schools. Variation can also take place as a result of the 'mode of discourse' such as in lectures, interviews, sermons, or electoral campaigns (Bachman, 1990; Bachman & Palmer, 1996). Concerning the 'style of discourse', it is organized into five levels: frozen, formal, consultative, casual, and intimate (Halliday et al., 1964). It is the type of relationship between the conversation participants that determines the register which is appropriate for a given language use context.

2.3.3.3.2.3. Sensitivity to Naturalness

Sensitivity to naturalness requires participant not only to express and interpret utterances as linguistically accurate, but in a native-like way as well.

2.3.3.3.2.4. The Ability to Interpret Cultural References and Figures of Speech

According to Lado (1961) culture is a part of any language. This can, for example, be manifested when conversation participants of the same speech community exchange utterances:

> the cultural meanings that each will encode in language for communication will usually be those which are common to other members of the cultural community. The listener thus grasps from the linguistic utterance the cultural meanings encoded in it by the speaker (p.5).

In the same way, the figures of speech such as metaphors, proverbs, or sayings require more than what is referred to as linguistic bound meaning. The ability to interpret cultural references and figures of speech enables us to express and interpret language use beyond the linguistic bound constraints.

2.3.3.2. Metacognitive Strategies

The second component of Bachman and Palmer's communicative language ability (CLA) refers to metacognitive strategies, also known as strategic competence. This constituent is perceived to include "a set of metacognitive components, or strategies, which can be thought of as higher order executive processes that provide a cognitive management function in language use, as well as in other cognitive activities" (Bachman and Palmer, 1996, p. 70). According to the authors, it is these strategies that enable language users' language knowledge to interact internally with their affective schemata; and allow these internal traits to interact with the external context for the purpose of creating and interpreting discourse. Bachman and Palmer conceptualize strategic competence from two standpoints: the interactional and the psycholinguistic approaches.

2.3..3.2.1. The Interactional Approach

Proponents of the interactional approach define communication strategies (CSs) as "a mutual attempt of two interlocutors to agree on a meaning in situations where requisite meaning structures do not seem to be shared" (Tarone, 1981, p.294). According to this approach, 'meaning structure' encompasses linguistic and sociolinguistic structures (see Fig. 9). As Tarone's definition implies, two main features should be taken into consideration. Firstly and most importantly, communication strategies can be engaged only when two or more people are mutually involved in creating and interpreting discourse. Equally important, the role of these strategies is limited to reinforcing the interactional process (meaning negotiation) and to compensating for deficiencies in linguistic or in sociolinguistic competence (Canale, 1983, 1984; Canale & Swain, 1980; Tarone, 1981).

Fig 9: The Interactional Approach to Strategic Competence

> (1) A speaker desires to communicate meaning x to a listener;
>
> (2) the speaker believes the linguistic or sociolinguistic structure desired to communicate meaning x is unavailable, or is not shared with the listener; thus
>
> (3) the speaker chooses to (a) avoid-not attempt to communicate meaning x-or (b) attempt alternate means to communicate meaning x. The speaker stops trying alternatives when it seems clear to the speaker that there is shared meaning.

Source: Tarone, 1981, p. 298.

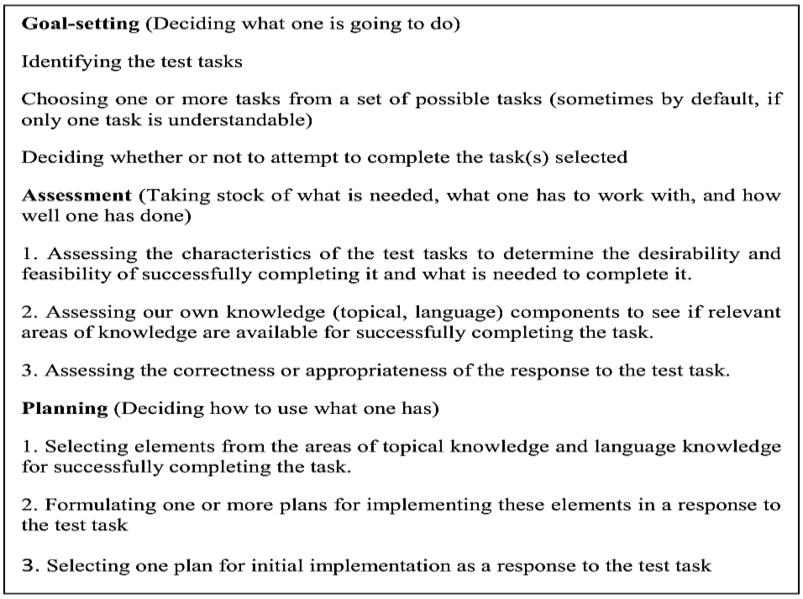
2.3..3.2.2. The Psychological Approach

The psychological approach, on its part, argues that constraining communication strategies (CSs) to compensatory or interactional roles seems to narrow their broad perspective **(** Dörnyei, 1995; Dörnyei & Lee Scott, 1997**).** Proponents of this approach claim that 'CSs' are first and foremost mental processes which enable language users to accomplish communicative goals through an action plan. Building upon this view, Bachman (1990) distinguishes between strategic competence and the psychophysiological mechanisms. He defines the former as "the mental capacity for implementing the components of language competence in contextualized communicative language use" (p. 107); and describes the latter as "the neurological and physiological processes [responsible for] the execution phase of language use" (p. 107). In Bachman and Palmer's model, the neurological and physiological processes have been incorporated to form an integral component of the metacognitive strategies (strategic competence).

2.3.3.2.3. Phases of Metacognitive Strategies

Bachman and Palmer (1996) identify three phases of metacognitive strategies: goal setting, assessment and planning (see Fig. 10.). In the goal-setting phase, the test taker decides what he is going to do. This involves identifying and selecting the test task that he intends to perform; and ultimately deciding whether or not to complete that task. In the assessment phase, the test taker identifies the characteristics of the test tasks. Then, he assesses the "desirability and feasibility of successfully completing [them]" (Bachman and Palmer, 1996, p.73) along with the topical or language knowledge required for doing the items. Moreover, this phase enables us to evaluate the extent to which communicative goals (test tasks) have been achieved. In the planning stage, test takers decide how to make use of their knowledge in order to accomplish their communicative goals (test tasks) in the most successful way. This involves selecting relevant areas of knowledge which are supposed to be included in the plan, "formulating one or more plans whose realization will be a response to the task, and selecting one plan for implementation as a response to the task" (p. 73).

Fig 10: Areas of Metacognitive Strategies



Goal-setting (Deciding what one is going to do)

Identifying the test tasks

Choosing one or more tasks from a set of possible tasks (sometimes by default, if only one task is understandable)

Deciding whether or not to attempt to complete the task(s) selected

Assessment (Taking stock of what is needed, what one has to work with, and how well one has done)

1. Assessing the characteristics of the test tasks to determine the desirability and feasibility of successfully completing it and what is needed to complete it.

2. Assessing our own knowledge (topical, language) components to see if relevant areas of knowledge are available for successfully completing the task.

3. Assessing the correctness or appropriateness of the response to the test task.

Planning (Deciding how to use what one has)

1. Selecting elements from the areas of topical knowledge and language knowledge for successfully completing the task.

2. Formulating one or more plans for implementing these elements in a response to the test task

3. Selecting one plan for initial implementation as a response to the test task

Source: Bachman and Palmer, 1996, p. 71

2.3.4. Specific Language Ability

In 'LSP' testing, we are concerned with making inferences about test takers' specific language ability and of measuring their capacity of using language in specific target domains (Douglas, 2000). There is an agreement amongst applied linguists that 'LSP' is not a type of language, but it refers to an approach of teaching/learning. More specifically, they consider LSP as a special case of communicative language teaching whose syllabus and teaching objectives are built around specific target situations (Basturkmen, 2006, 2010; Basturkmen & Elder, 2001; Widdowson, 1979, 1983, 2003). However their divergence is on whether 'LSP' testing bases its concept on a specialized language content or on a theoretical description of' 'specific' language ability. The first trend, represented by Widdowson (1978, 2003), Hutchinson and Waters (1987), Basturkmen, (2006, 2010) and Davies and Elder, (2004), maintain that "all uses of English, as any other language, are specific [and] all uses of language serve a particular purpose" (Widdowson 2003p. 61). Consequently, the main distinction between LSP and general communicative language teaching, according to this trend, lies in the constrained scope of the purpose that we intend to achieve by means of language learning or testing. The other trend, represented by Bachman (1990, 1991); Bachman and Palmer, (1996); Alderson and Bachman, (2000-2006) and Douglas (2000, 2010a, 2010b, 2013), perceive that the process of LSP testing stands on theoretical grounds describing the components of 'specific purpose language ability' and their interaction with the external context.

2.3.4.1. Douglas' Model of Specific Language Ability

As we have mentioned above, applied linguists do not have the same point of view regarding LSP testing. Some of them think that this field remains atheoretical and its implementation is based only on practical grounds. Others, such as, Douglas (2000, 2001, 2005, 2010b, 2013) strongly argue that:

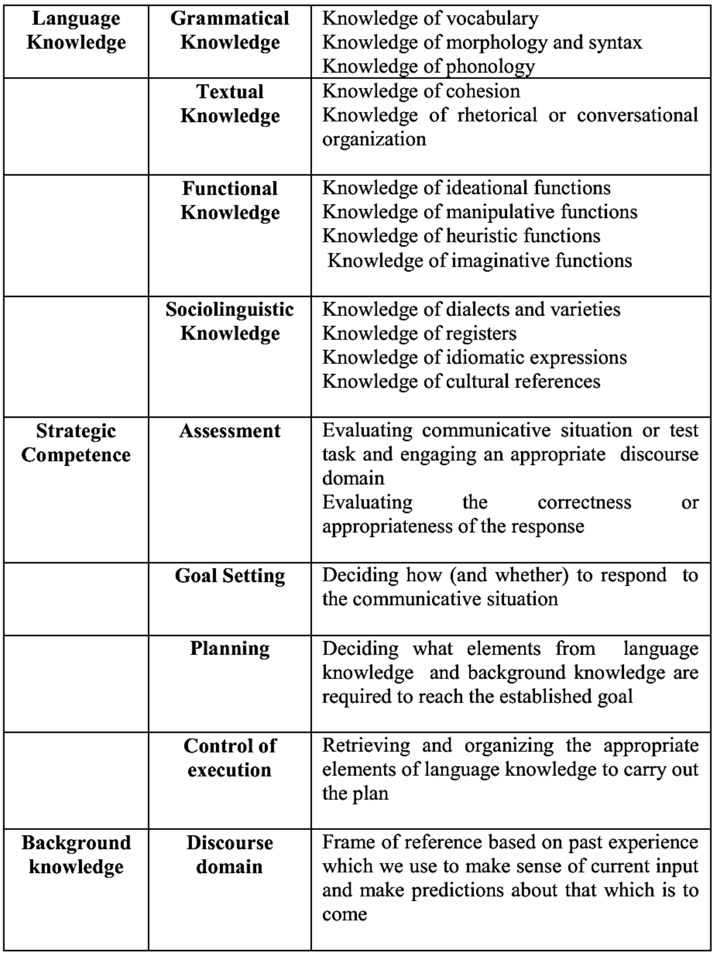
> these assertions are not true, that there is a theoretical justification for ESP, that ESP is different from general purpose language, that language knowledge and specific purpose background knowledge are both part of the ESP construct, and that specific purpose language testing is not only possible but necessary (Douglas, 2010b, p. 3)..

This view is supported by Alderson and Bachman (2000) who point out that Douglas (2000) has "formulated a theoretical framework that provides a basis for developing and using assessments of language for specific purposes" (p. ix).

2.3.4.2.Components of Specific Language Ability

Built upon Bachman (1990) and Bachman and Palmer (1996), Douglas (2000, 2001, 2010a, 2010b, 2013) thinks of specific language ability to be consisting of three main constituents: language knowledge, specific purpose background knowledge, and strategic competence (see Table 3). Language knowledge derives largely from Bachman's language competence (1990) and Bachman and Palmer's language knowledge (1996). Specific purpose background knowledge refers to the knowledge that test takers or language users have acquired as a result of their academic study or of their work at given field of interest. Strategic competence refers to the metacognitive and communicative processes that enable test takers' language knowledge to interact with their background knowledge on the one hand; and it also makes it possible for their internal abilities to interact with the external context on the other.

Table 3: Components of Specific Language Ability

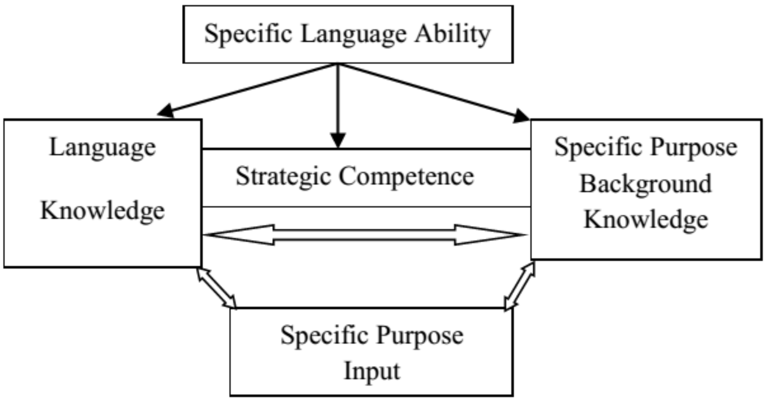


Language Knowledge	**Grammatical Knowledge**	Knowledge of vocabulary Knowledge of morphology and syntax Knowledge of phonology
	Textual Knowledge	Knowledge of cohesion Knowledge of rhetorical or conversational organization
	Functional Knowledge	Knowledge of ideational functions Knowledge of manipulative functions Knowledge of heuristic functions Knowledge of imaginative functions
	Sociolinguistic Knowledge	Knowledge of dialects and varieties Knowledge of registers Knowledge of idiomatic expressions Knowledge of cultural references
Strategic Competence	**Assessment**	Evaluating communicative situation or test task and engaging an appropriate discourse domain Evaluating the correctness or appropriateness of the response
	Goal Setting	Deciding how (and whether) to respond to the communicative situation
	Planning	Deciding what elements from language knowledge and background knowledge are required to reach the established goal
	Control of execution	Retrieving and organizing the appropriate elements of language knowledge to carry out the plan
Background knowledge	**Discourse domain**	Frame of reference based on past experience which we use to make sense of current input and make predictions about that which is to come

Source: Douglas, 2000, p. 35

Specific language ability, as it is illustrated in fig 11, "results from the interaction between specific purpose background knowledge and language ability, by means of strategic competence engaged by specific purpose input in the form of test method characteristics" (Douglas, 2000, p.40).

Fig 11: Specific Language Ability


Specific Language Ability
Language Knowledge
Strategic Competence
Specific Purpose Background Knowledge
Specific Purpose Input

Organized from Douglas, 2000, p. 40

2.3.4.3.1 Language Knowledge

Language knowledge is made up of four components: grammatical, textual, functional, and sociolinguistic knowledge (for more details about this competency, see Bachman and Palmer's Model, pp. 50-52).

2.3.4.2.2. Specific Background Knowledge

Using English for academic or occupational purposes "requires not only linguistic proficiency and [language] knowledge but also knowledge and understanding of work-related and disciplinary concepts" **(**Basturkmen, 2006, p. 137) known as background knowledge (Douglas, 2000, 2013). This type of knowledge can be defined as the "disciplinary concepts from the students' field of study" (Hutchinson & Waters, 1987, as cited in Basturkmen, 2006, p. 137). Applied linguists stress that the more a test is field-specific, the more it will engage test takers' background knowledge to interact with the test input "as field specificity increases, background knowledge will have a proportionately stronger effect on test scores" (Douglas, 2000, p. 34). Conversely, if an LSP test is not field specific, language knowledge alone may not help.

2.3.4.2.3. Strategic Competence

In LSP testing, strategic competence refers to the processes that serve as a mediator "between the learner's internal traits of background knowledge and language knowledge and the external context, controlling the interaction between them" (Douglas, 2000, p. 76). Douglas informs us that this process takes place as soon as test takers or real-life language users start to assess the characteristics of the language use or testing situation by:

> engaging an appropriate discourse domain, or creating a temporary one; they establish goals for responding to the situation: they make a plan for meeting the goals, deciding what elements of knowledge will be required, and they control the execution of the plan by retrieving the required knowledge and organizing it into a coherent response" (p. 76).

Extending Bachman and Palmer's (1996) concept of strategic competence, Douglas (2000, 2001, and 2010b) organizes these processes into two major classes: metacognitive strategies (MSs) and communicative strategies (CSs). The two types are hierarchically engaged in that it the 'MSs' which "direct the language user's interaction with the context, while communicative strategies are called on by the metacognitive strategies to take over direction when the features of the context are specifically identified communicative" (Douglas, 2000, 76-77).

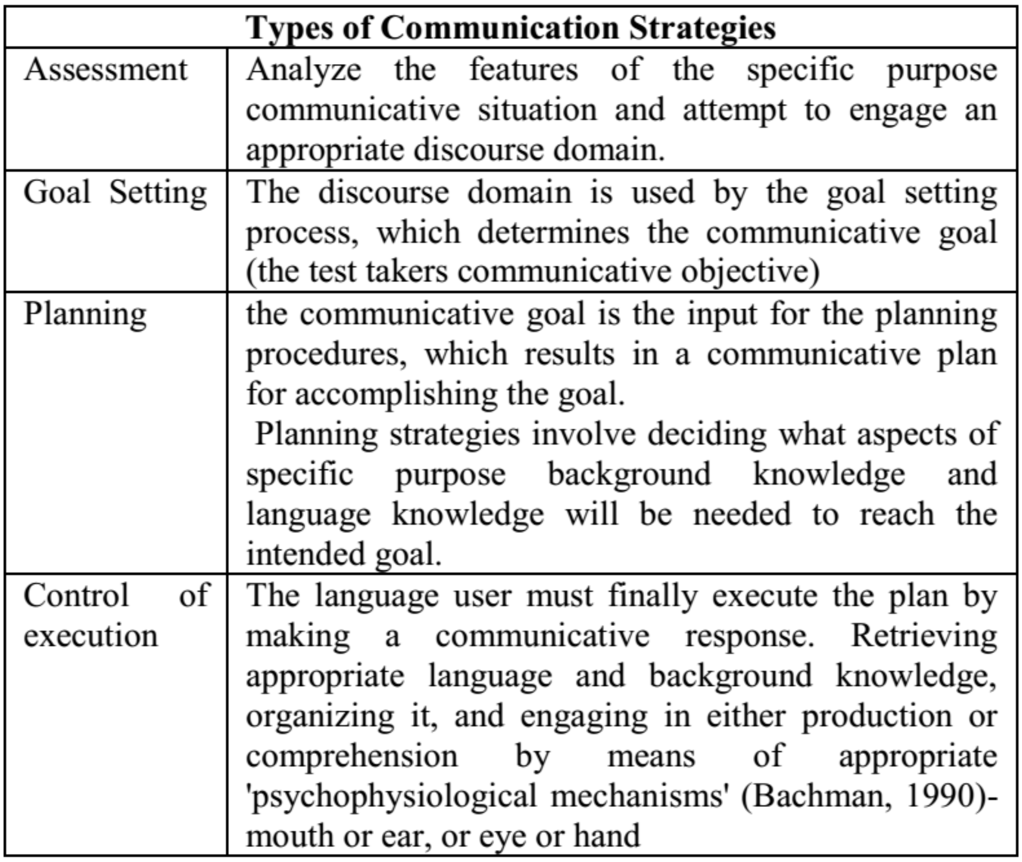
2.3.4.2.3.1. Metacognitive Strategies

The term 'metacognitive strategies' in Douglas' model is not used in the same way as the one used in Bachman and Palmer's model. The latter use the term to refer to all the components of strategic competence "we conceive strategic competence as a set of metacognitive components, or strategies, which can be thought of as higher order of executive processes that provide a cognitive management function in language use, as well in all cognitive activities" (Bachman & Palmer, 1996, p. 70). Conversely, in Douglas' model metacognitive (MSs) strategies are engaged only when the test taker perceives the language situations or test tasks are non-communicative. Douglas (2000) points out that the distinction between metacognitive and communicative strategies is of great importance, especially in the design of test tasks that do not require language responses on the part of test takers. For this reason, he restrains the scope of metacognitive strategies to be "directly responsible for performance in situations not requiring language, such as carrying out a laboratory procedure, or operating an overhead projector" (p. 77).

2.3.4.2.3.2. Communicative Strategies (CSs)

As we have mentioned above, the relationship between the 'two-tiered' types of strategies is hierarchical, in that when the higher level (MSs) perceives that test tasks require meaning negotiation, or discourse creation, they engage the lower level 'CSs' to take action. According to Douglas (2000), communicative strategies "work specifically with language by bringing relevant knowledge into use at the right time, and in the right relationship to the resources demanded by the task" (p. 79). In LSP testing, communicative strategies are made up of four components: assessment, goal setting, planning, and control of execution (see Table .3). The incorporation of Douglas' communicative strategies starts with the assessment of the specific target situation/ test tasks in order to engage the 'appropriate discourse domain'. Then, the language user/ test taker determines the objectives set to be achieved. In the next step 'planning', the test taker decides what constituents of specific background and language knowledge are to be called for achieving his communicative goal. The last step involves the execution of the previous plan which ends in written or oral responses to test tasks.

Table: Areas of Communication Strategies



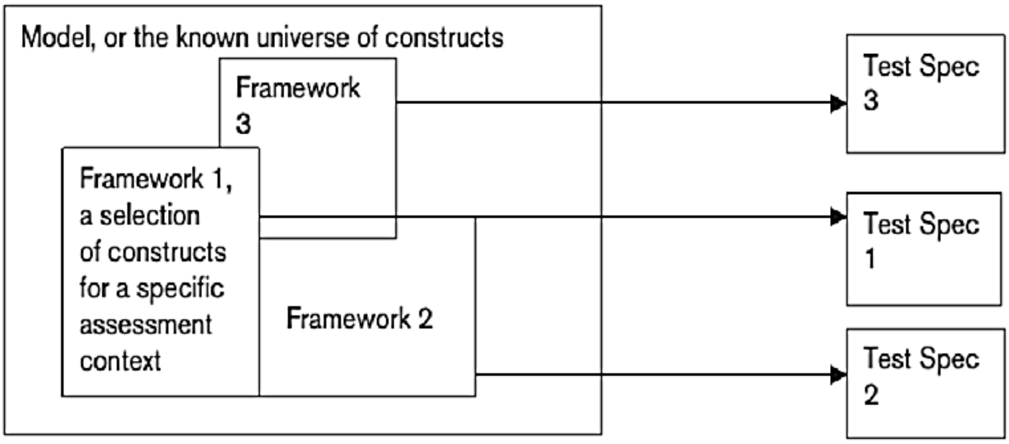
Types of Communication Strategies	
Assessment	Analyze the features of the specific purpose communicative situation and attempt to engage an appropriate discourse domain.
Goal Setting	The discourse domain is used by the goal setting process, which determines the communicative goal (the test takers communicative objective)
Planning	the communicative goal is the input for the planning procedures, which results in a communicative plan for accomplishing the goal. Planning strategies involve deciding what aspects of specific purpose background knowledge and language knowledge will be needed to reach the intended goal.
Control of execution	The language user must finally execute the plan by making a communicative response. Retrieving appropriate language and background knowledge, organizing it, and engaging in either production or comprehension by means of appropriate 'psychophysiological mechanisms' (Bachman, 1990)-mouth or ear, or eye or hand

Adapted from Douglas, 2000, pp. 80, 81 &82

2.4.Test Framework

The second layer in the constructional design concerns test frameworks. The latter refer to the "selection of skills and abilities from a model that are relevant to a specific assessment context" (Fulcher and Davidson, 2007, p.36). This document, as it is illustrated in Fig 12, mediates between the theories of language ability and test specifications**.** It selects the constructs from the models and lays them out for particular testing situations. In the same way, it generates the writing of test and item specifications (Alderson, 2000, 2005; Fulcher, 2010; Fulcher & Davidson, 2007, 2009; Purpura, 2008). The framework states the purpose of the testing event; describes test takers' characteristics; and defines the scope of target language situations on which scores are intended to be generalized. It also "delineate[s] the aspects (e.g., content, skills, processes, and diagnostic features) of the construct or domain to be measured… [and] provides a description of how the construct or domain will be represented" (AERA, APA & NCME, 1999, p. 37[explanation in original]).

Fig 12: Relationship between Models, Frameworks, and Specifications


Model, or the known universe of constructs
Framework 3
Framework 1, a selection of constructs for a specific assessment context
Framework 2
Test Spec 3
Test Spec 1
Test Spec 2

Source: Fulcher and Davidson, 2007, p. 37.

2.4.1. Test Constructs

The field of language testing is made up of two major constituents the 'what' and the 'how' (Alderson & Bachman, 2000-2006; Purpura, 2004, 2008; Shohamy, 2008).The 'what' refers to the constructs or the traits that test designers intend to measure. The other constituent refers to the ‘how' or the test method which describes the characteristics of the test and the participants. This field is concerned with making inferences about test takers' language ability and about their capacity of using language in situations beyond the test itself. Similarly, we need to ensure that the scores obtained from these tests do really reflect the constructs being tested. Consequently, “in order to justify a particular score interpretation”, as Bachman and Palmer (1996) underline, “we need to provide evidence that the test scores reflect the area (s) of language abilities we want to measure, and very little else [and] in order to provide such evidence, we must define the construct we want to measure" (p. 21) in a way that is appropriate for a particular testing situation.

2.4.1.1. Definition of Constructs

Every test has a model of language ability behind it; and every test is designed to measure one or more components of this model, also known as constructs (Alderson, 2000a, 2007; Alderson, et al., 1995; Bachman, 1990, 2007; Bachman & Palmer, 1996). A construct can be defined as "a psychological concept, which derives from a theory of the ability to be tested…Constructs are not psychologically real entities that exist in our heads. Rather, they are abstractions that we define for specific assessment purpose" (Alderson, 2000a, p. 118). In the same way, Fulcher (2010) perceives constructs as the "abilities of the learner that we believe underlie their test performance, but which we cannot directly observe"(p. 96). Unlike physical characteristics which can directly be observed, constructs "are inferred from interrelated observations that a test is designed to measure" (AERA, APA & NCME, 1999, p.5).

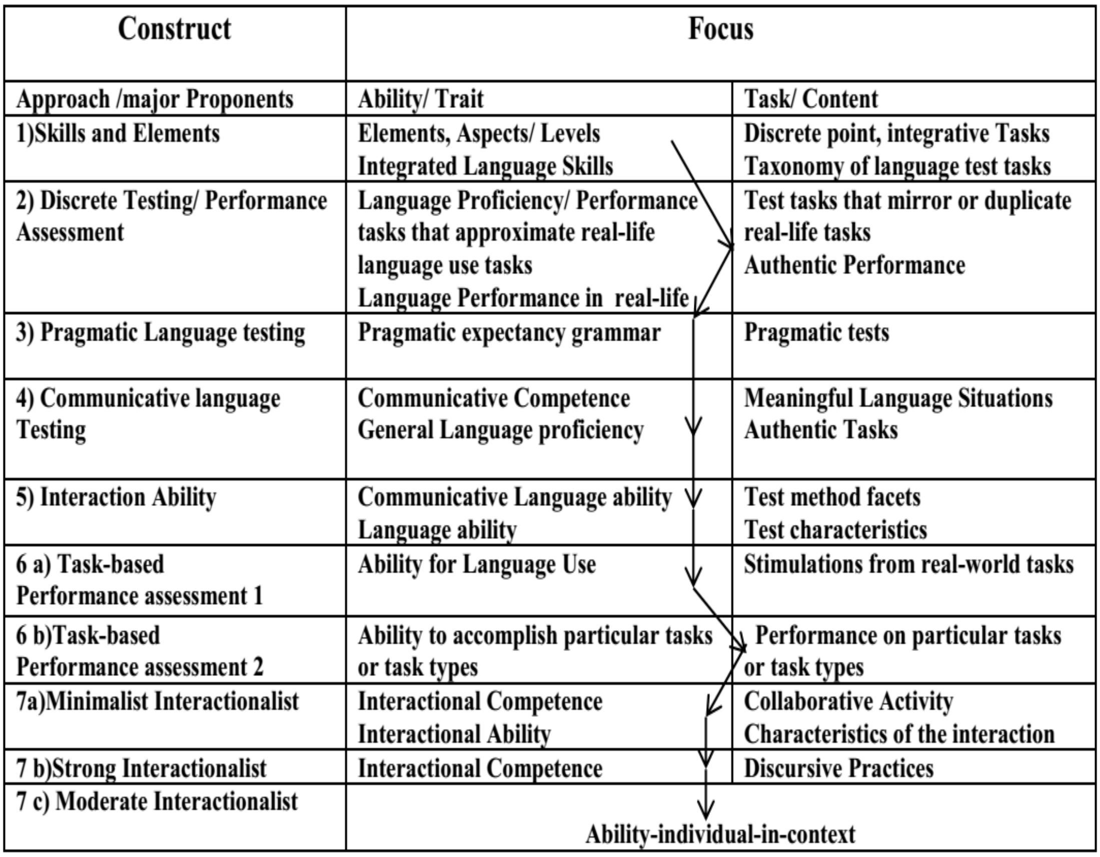
For a clear understanding of constructs, let us first distinguish between physical (observable) and mental traits. Physical characteristics such as length, color, or height "can be experienced directly through the senses, and can therefore be defined by direct comparison with a directly observable standard" (Bachman, 1990, p. 41). Conversely, in the theory of language ability, grammatical, sociolinguistic, functional knowledge or sensitivity to naturalness, for instance, constitute mental traits that one cannot directly observe (Bachman and Palmer, 1996). In the same way, we can speak of different interrelated mental traits such as "skimming, scanning, getting the gist, distinguishing the main ideas from supporting detail, distinguishing statement from example, etc." (Alderson, et al., 1995, p. 17) that underlie the theory of reading. Seeing that we cannot measure these traits directly, we resort to inferring them "through observing [the] behavior that we presume to be influenced" (Bachman, 1990, p. 41) by these constructs.

To illustrate this point, suppose that we want to measure test takers' functional competence. Due to the fact that it is not possible for us to observe how this competence functions, we design test tasks that enable us to see how test takers perform in such situations. This means that mental traits or theoretical constructs need to be operationalized before being measured. Alderson (2000a) recommends that "in designing a test, we do not so much pick the 'psychological entity' we want to measure, as attempt to define that entity in such a way that it can eventually be operationalised in a test" (pp. 118-19). In brief, for general concepts to become measureable, constructs need to be "so defined that they can become ‘operational’" (Fulcher and Davidson, 2007, p. 7) for a particular testing contexts.

2.4.1.2. Approaches to Construct Definition

As shown in Table 5, three approaches to construct definition have been identified in the literature of language testing. These include ability-based, performance-based, and interaction-based approaches (Bachman, 2007; Chappelle, 2008, 2010, 2012; Chapelle, Enright, & Jamieson, 2008, 2010). Ability-based approaches define the construct to be tested "in terms of areas of language ability that test takers have" (Bachman, 2007, p. 57). Performance-based approaches perceive constructs "in terms of what test takers can do in contexts beyond the test itself" (p.57). In other words, ability-based approaches look into the aspects of knowledge that learners have; however, performance-based approaches specify what test takers can do with this knowledge in target language domains. Interaction-based approaches, on their part, relate aspects from test takers' language ability (what they have) with aspects of task performance (what they can do) by means of strategic competence "responsible for putting person characteristics to use in contexts" (Chapelle, 1998, p. 44).

Table 5: Dialectic Definition of Constructs



Construct	Focus	
Approach /major Proponents	Ability/ Trait	Task/ Content
1)Skills and Elements	Elements, Aspects/ Levels Integrated Language Skills	Discrete point, integrative Tasks Taxonomy of language test tasks
2) Discrete Testing/ Performance Assessment	Language Proficiency/ Performance tasks that approximate real-life language use tasks Language Performance in real-life	Test tasks that mirror or duplicate real-life tasks Authentic Performance
3) Pragmatic Language testing	Pragmatic expectancy grammar	Pragmatic tests
4) Communicative language Testing	Communicative Competence General Language proficiency	Meaningful Language Situations Authentic Tasks
5) Interaction Ability	Communicative Language ability Language ability	Test method facets Test characteristics
6 a) Task-based Performance assessment 1	Ability for Language Use	Stimulations from real-world tasks
6 b)Task-based Performance assessment 2	Ability to accomplish particular tasks or task types	Performance on particular tasks or task types
7a)Minimalist Interactionalist	Interactional Competence Interactional Ability	Collaborative Activity Characteristics of the interaction
7 b)Strong Interactionalist	Interactional Competence	Discursive Practices
7 c) Moderate Interactionalist	Ability-individual-in-context	

Source: Bachman, 2007, p. 44-5

In addition to these approaches, Bachman and Palmer (1996, as cited in Purpura, 2004) provide three other options for the delineation of constructs with respect to topical knowledge (TK). In cases when we conceive that 'TK' is not of great interest for the instructional syllabus, the definition of the construct will be limited to the different components of the language ability we intend to measure. More importantly, 'TK' in this case can be considered as a construct irrelevant variance responsible for affecting test scores in the negative side. The second option refers to the case in which TK is considered as an integral part of the construct "and where topics or themes contextualize language, provide a social–cognitive context for the tasks, and serve to raise the students' interest level" (Purpura, 2004, p. 159). In the third option, the components of language ability and TK are identified as separate constructs. This is the case when TK is of equal or of more importance than trait-based constructs. This is appropriate in content-based programs, or ESP-based syllabi (Douglas, 2000, 2001, 2010). In technology streams, for instance, the third year syllabi are based on thematic knowledge, such as automation, computing, or mechanization (Ministry of Education, 1998) which requires the construct to be defined with respect to the components of the language ability to be tested as well as the "theme-based language programs, where topic serves as a context for language learning" (Purpura, 2004, p. 159).

2.5. Test Specifications

The third layer of the test construction concerns test specifications. The latter is a detailed document that lays out the blueprints for writing an entire test (Bachman & Palmer, 1996). This document, which is guided by the purpose of the testing event, describes the specific construct to be measured; determines the type and number of tasks used to collect evidence about this construct; distributes the items according to their level of difficulty; and organizes them in a way to represent the content domain. Similarly, the specifications describe test takers' characteristics such as age, gender, their cultural background or levels of language ability; specify the amount of time to be is allotted to the entire test; and set the rules for administration and scoring procedures (AERA, APA & NCME, 1999; Alderson et al., 1995; Gronlund, 1977; Osterlind, 2002, Popham, 1978). Test specifications tell test writers "what the test tests and how it tests it" (Alderson, et al., 1995, p.9). According to Fulcher and Davidson (2007), 'what the test tests' " concerns the identification of the unobservable traits that are intended to be tested and the type of evidence we need to collect in order to make inferences about the abilities being measured"

(p. 67) and 'how it tests it' refers to "the situations in which test takers respond to items or tasks that generate the evidence we need" (p.67).

Test specifications should not be confounded with syllabus specifications (Alderson et al., 1995). The latter is a "public document…which indicates to test users what the test will contain [and] it is directed more to teachers and students who wish to prepare for the test…and to publishers who wish to produce materials related to the test" (p.10). However, test specification is a more detailed and often confidential document that is intended to test constructors (telling them what to include in the test); to test validators (allowing them to examine whether the test has really measured the defined constructs); and to test users (enabling them to validate their decisions). In short, the use of syllabus specification is limited to the persons who want to know what a given test will contain; and cannot be used as a basis for language test construction or evaluation.

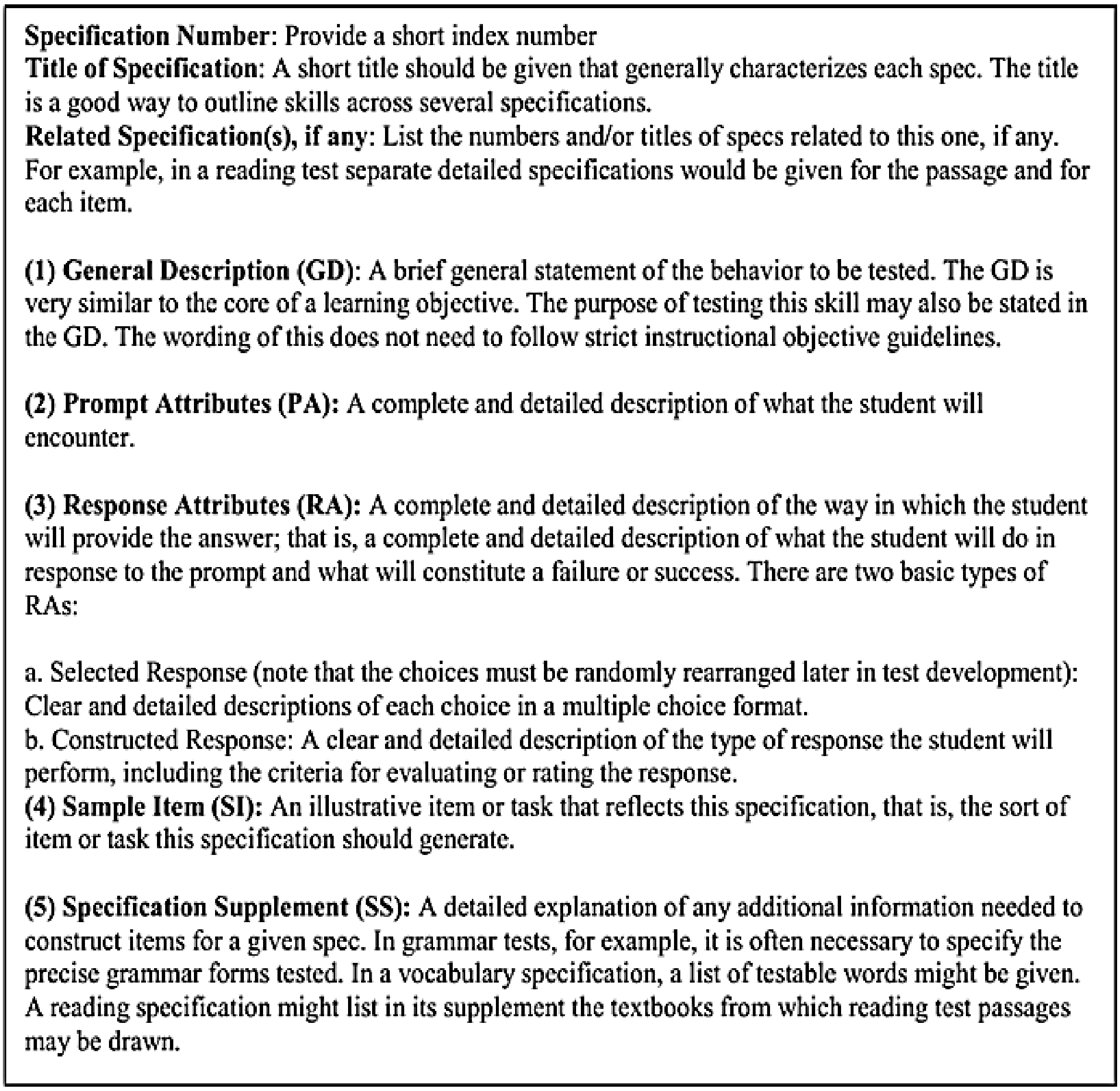
2.5.1. Components of Test Specifications

Several frameworks for test specifications have been proposed in the literature. The most known of these include Popham's specifications (1978) modified by Lynch and Davidson (1994, 1998) and by Davidson and Lynch (2002), Alderson et al's specifications (1995), Bachman and Palmer's blueprints (1996) as well as the specifications provided by the Standards for Educational and Psychological Testing (Chapelle & Douglas, 2006; Fulcher, 2010; Luoma, 2004; Read, 2000).

2.5.1.1. Popham's Specifications

Popham's specifications which have been modified by Lynch and Davidson (1994, 1998) include five components (see Fig. 13): general description, prompt attributes, response attributes, sample item and specification supplement (Davidson & Lynch, 2002). The general description delineates the testing purpose and describes the behavior (construct) to be measured. To distinguish between the framework and the specification constructs, we can say that the scope of former is wider and describes a number of interrelated traits; while the latter is stated in a way to describe a specific behavior. For example, the construct of reading includes a set of specific constructs such as "skimming, scanning, getting the gist, understanding the communicative factions and paragraphs [and so on]" (Alderson, et al., 1995, p. 15). The second component, 'prompt attributes' refers to what test takers are required to do in order to demonstrate their ability in the criterion intended to be tested. The next component 'response attributes' describes the way how test takers respond to test tasks (Davidson & Lynch, 2002). The fourth component, 'sample item' reminds test writers to include sample responses or answers to each test item. The fifth component is a specification supplement which provides a "detailed explanation of any additional information needed to construct items for a given spec" (Lynch & Davidson, 1994, p. 731).

Fig 13: Popham's Test Specifications (1978)



Specification Number: Provide a short index number
Title of Specification: A short title should be given that generally characterizes each spec. The title is a good way to outline skills across several specifications.
Related Specification(s), if any: List the numbers and/or titles of specs related to this one, if any. For example, in a reading test separate detailed specifications would be given for the passage and for each item.

(1) General Description (GD): A brief general statement of the behavior to be tested. The GD is very similar to the core of a learning objective. The purpose of testing this skill may also be stated in the GD. The wording of this does not need to follow strict instructional objective guidelines.

(2) Prompt Attributes (PA): A complete and detailed description of what the student will encounter.

(3) Response Attributes (RA): A complete and detailed description of the way in which the student will provide the answer; that is, a complete and detailed description of what the student will do in response to the prompt and what will constitute a failure or success. There are two basic types of RAs:

a. Selected Response (note that the choices must be randomly rearranged later in test development): Clear and detailed descriptions of each choice in a multiple choice format.
b. Constructed Response: A clear and detailed description of the type of response the student will perform, including the criteria for evaluating or rating the response.
(4) Sample Item (SI): An illustrative item or task that reflects this specification, that is, the sort of item or task this specification should generate.

(5) Specification Supplement (SS): A detailed explanation of any additional information needed to construct items for a given spec. In grammar tests, for example, it is often necessary to specify the precise grammar forms tested. In a vocabulary specification, a list of testable words might be given. A reading specification might list in its supplement the textbooks from which reading test passages may be drawn.

Source: Lynch and Davidson, 1994, p. 731; Davidson and Lynch, 2002, p. 14

2.5.1.2. Alderson et al's Specifications

As it is included in Fig 14 and before the operational writing of the test , Alderson et al's Specifications (1995) require test writers to respond to twelve questions concerning the testing purpose, the characteristics of test takers, the behavior to be measured, the target language use domains, test tasks, type and number of items, time allotment, test administration and the scoring procedures.

Fig 14: Alderson et al's Test Specifications

1. **What is the purpose of the test?**
2. **What sort of learners will be taking the test?**
3. **How many sections/papers should the test have, how long should they be and how will they be differentiated?**
4. **What target language situation is envisaged for the test, and is this to be simulated in some way in the test contest and method?**
5. **What text types should be chosen?.. How difficult or long should they be? What functions should be embodied in the texts? How complex should the language be?**
6. **What language skills should be tested?**
7. **What language elements should be tested?**
8. **What sorts of tasks are required?**
9. **How many items are required for each section? What is the relative weight for each item?**
10. **What methods are to be used?**
11. **What rubrics are to be used as instructions for candidates?**
12. **What criteria will be used for assessment by markers? (pp. 12-13)**

Source: Alderson, et al. 1995, pp. 12-13

2.5.2. Item Specifications

Before talking about item specifications, let us first review some of the definitions provided for the concept 'test item' in the literature of language testing. This is because “by knowing the definition, purpose, and characteristics of test items, one will have at hand a great deal of information about a particular test item, its construction, function, and probable effectiveness” (Osterlind, 2002, p. 18). More importantly, the delineation of these concepts enables us to distinguish them from both instructional items and test tasks. This distinction helps item writers “produce items of quality—that is, test items that meet criteria for good items—than may be yielded with a haphazard [way]” (p.18).

2.5.2.1. Definition of Test Items

Osterlind (1990, as cited in Osterlind, 2002) defines a test item as “a unit of measurement with a stimulus and a prescriptive form for answering; and, it is intended to yield a response from an examinee from which performance in some psychological construct…may be inferred (p. 19). On their part, the Standards of Educational and Psychological Testing consider a test item as a “statement, question, exercise, or task on a test for which the test taker is to select or construct a response, or perform a task" ([AERA], [APA], [NCME], 1999, p. 177). These definitions suggest that test items perform four interrelated functions. The first function is related to measurement: test takers' responses will be interpreted in terms of scores. In this context, Osterlind (2002) explains that “the numerical interpretation for test items is what differentiates them from instructional activities” (p.20). The second aspect has to do with stimuli-responses relationship. In language tests, items perform the function of stimuli that call for prescribed responses. Prescribed responses mean that test takers are guided to respond in a particular format such as multiple-choice format or a constructed-response format. Osterlind warns item writers that they “would violate the definition of a test item if the test taker were not directed to make a particular, predetermined kind of response” (p. 21). The fourth aspect suggests that test items are used as instruments to make inferences about test takers' language ability or performance (Osterlind, 1990, 2002).

2.5.2.2. Definition of Item Specifications

Based on Popham (1978), Lynch and Davidson (1994), Davidson and Lynch, (2002) and Fulcher and Davidson (2007, 2009), Fulcher (2010) defines item specifications as a plan that:

> describe[s] the prompts that are designed to elicit the evidence upon which inferences are made about the targeted abilities of the learners. Minimally, these specifications should state what kind of input material the test takers will encounter, what the instructions look like…[and the] ways in which the task may change, or which alterations are permissible" (127).

On his part, Osterlind (2002) writes that test item specifications refer to "a specialized kind of technical writing used in developing a set of items…[they] are formal, systematized directions from a test developer to the item writer that seek to put the test …specifications into action”(p. 88). It follows from these definitions that item specifications represent a specialized document that gives item writers directions on how to develop test items with respect to format, type, number, degree of congruency with test specifications, characteristics of prompt response, measurement of response attributes as well as assembly directions.

Conclusion

Architectural design of tests delineates three main constituents: theories of language ability, test frameworks, and specifications. The relationship amongst these components is hierarchical. The models which refer to what it means to know and use a language operate at the higher level. Theories or models are made up of interrelated constructs. The second layer refers to test frameworks. These components sample the constructs from the theories of language and operationalize them for particular testing situations. The frameworks mediate between the abstract models and operational specifications. The third component concerns the test blueprints. Generated by the framework, this document tells us how to write items and how to compile them into a comprehensive test.

Chapter Three

Stages of Test Development

Chapter Three

Stages of Test Development

Introduction

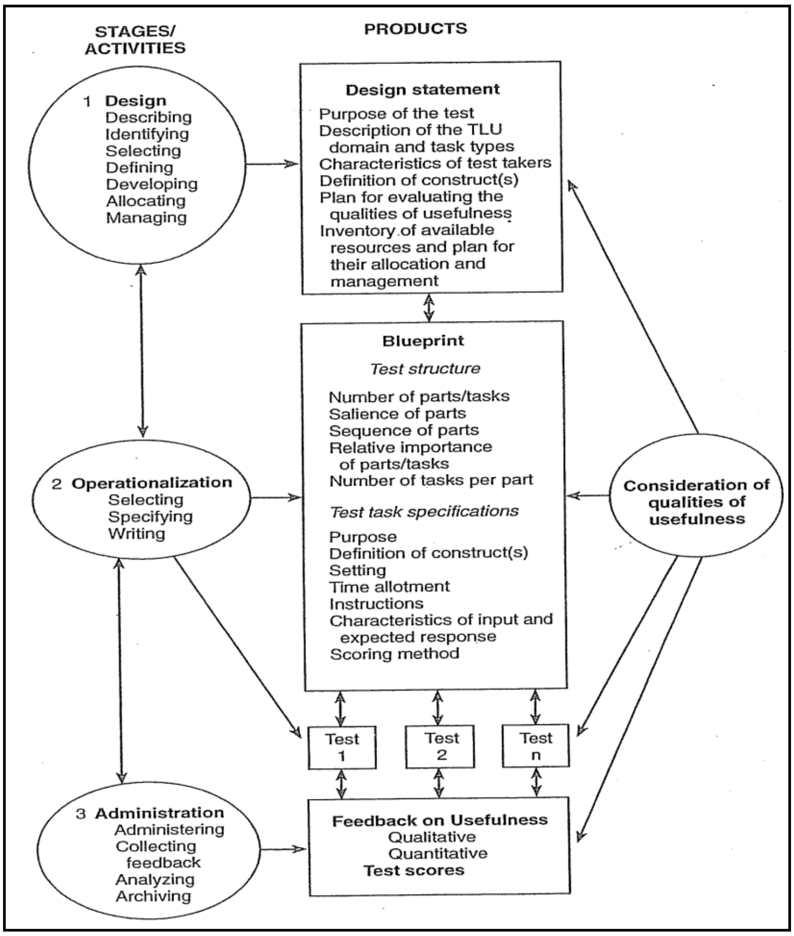
Language test development refers to the process of producing some type of measure in order to assess test takers' levels of language ability; or to examine the extent to which they can use this ability in real communicative contexts. This process starts with initial perception of the need to build a measure and concludes with the design of a concrete test (AERA, APA & NCME, 1999; Alderson, et al., 1995; Bachman, 1990, Bachman & Palmer, 1996; Fulcher, 2003, 2010). Due to that the fact that the scores emerging from tests can be used to make decisions which can affect a large number of people, the developmental "processes might be highly complex, perhaps involving extensive trialing and revision, as well as coordinating the efforts of a large test development team" (Bachman & Palmer, 1996, p.85). In developing such tests, the authors insist that "the qualities of usefulness need to be carefully considered and this consideration should not be scarified in either low-stakes [or] high-stakes situations" (p.85).

Several models for language test development have been proposed in the literature. We can, for example, mention Henning's *'Guide to Language Testing'* (1987), Heaton's *'Writing English Language Tests'* (1988), Bachman's *'Fundamental Considerations in Language Testing*' (1990), Alderson, et al's *'Test Construction and Evaluation'* (1995); Milanovic's *'Language Examining and Test Development'* (2002) or Mislevy et al's *'Evidence Centered Design'* (2003). However, most language testers think of Bachman and Palmer's model (1966) "to be more successful as a powerful intellectual framework… acting as a conceptual mold within which a number of very helpful books have been written…on various aspects of language testing " (McNamara & Roever, 2006, p.34). The authors, of course, refer to the 'Cambridge Language Assessment Series', edited by Alderson and Bachman (2000-2006), which published a number of important books focusing on assessing the four skills of listening, speaking, reading and writing, as well as vocabulary, grammar, language for specific purposes and language through computer technology (Alderson, 2000a; Bachman, 2004; Buck, 2001; Chapelle & Douglas, 2006; Douglas, 2000; Luoma, 2004; McKay, 2006; Purpura, 2004; Read, 2000; Weigle, 2002)

3. Stages of Test Development in Bachman and Palmer's Model

Bachman and Palmer (1996) organize their model of test development into three linear and iterative stages which include design, operationalization, and administration (see Fig 15). The design stage focuses on delineating the guiding purpose; defining the constructs to be tested; collecting information on test takers' characteristics; and examining the extent of authenticity between test tasks and target language use situations. Operationalization which is governed by stage one draws the specifications for writing tasks and test blueprints. The third stage concerns test administration. The latter is conducted at two levels. Phase one specifies the procedures for test tryout and feedback collection; and phase two focuses on live administration.

Fig 15: Stages of Test Development



Source: Bachman and Palmer, 1996, p. 87.

3.1. The Design Stage

The design stage "involves the accumulation of information and making initial decisions about the entire test process"(Purpura, 2004 p.156). According to Bachman and Palmer (1996), this stage is made up of six activities. In the first place, it decides on the purpose of the test and delineates the scope of its construct. Then, it provides a portrayal of test takers' characteristics; and conducts an analysis of the target language use tasks (TLU). It also ensures the correspondence between TLU tasks and tasks in the test. Additionally, it sketches out a plan for evaluating the qualities of usefulness and draws an inventory for the required material and human resources. In summary, the design stage which offers test writers "a principled basis for developing test tasks, a blueprint, and tests enable[s] us to monitor the subsequent stages of development" (Bachman and Palmer 1996, p 88).

3.1.1. Describing the Specific Purpose(s) of the Test

The design stage is guided by the statement of the purpose. The delineation of the purpose defines the scope of the construct(s) and the content to be measured. At the same time, it provides feedback for designing the test blueprints and specifying the characteristics of test takers. In developing such tests, the purpose should be stated in a clear and specific way because no test is valid for all purposes (AERA, APA & NCME, 1999) and "if a test producer wishes to have a test that can fulfill any purpose, we have design chaos" (Chalhoub-Deville & Fulcher, 2003, p. 502).

3.1.2. Target Language Use Domains

Target language use (TLU) domains can be defined as "a set of specific language use tasks that the test taker is likely to encounter outside the test itself, and to which we want our inferences about language ability to generalize" (Bachman & Palmer, 1996, p. 44). The authors organize 'TLU' domains into two categories: real-life domains and instructional domains. The first category refers to the situations where language is used for real communication purposes. The second category contains the situations where language is used for instructional purposes (teaching and learning). The first category is broad and can be specified by outlining the second type. For example, 'English for Business Communication' is a real life domain within which we can draw some instructional domains such as negotiating with clients, bargaining, advertising, and so on (Nunan, 2004).

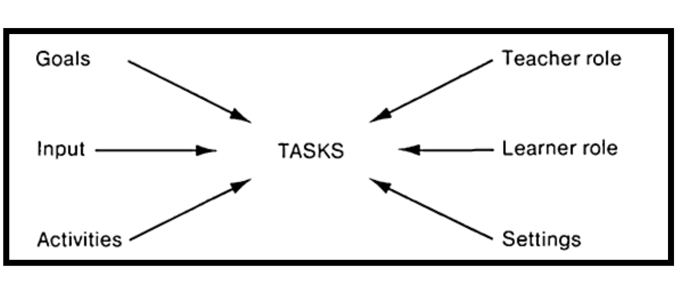
3.1.3. Test Tasks

Before supplying a definition to test tasks, let us first distinguish between 'real world' tasks and instructional tasks. The former refer to "a piece of work undertaken for oneself or for others, freely or for some reward" (Long, 1985, as cited in Nunan, 1989, p. 5). This type includes "the hundred and one things people do in everyday life, at work, at play, or in between" (Nunan, 1989, p. p.5) such as watching TV, doing the shopping, driving one's car, reading a text, or responding to questions. As far as education is concerned, and built upon Richards, Platt and Webber (1986), Nunan, (1989, 1999) and Ellis (2003), Nunan (2004) defines a pedagogical or an instructional task as:

> a piece of classroom work that involves learners in comprehending, manipulating, producing or interacting in the target language while their attention is focused on mobilizing their grammatical knowledge in order to express meaning, and in which the intention is to convey meaning rather than to manipulate form. The task should also have a sense of completeness, being able to stand alone as a communicative act in its own right with a beginning, middle, and an end (p.4)

At the same time, Nunan (1999) distinguishes between instructional tasks and exercises. The former can have a nonlinguistic outcome whereas the outcome of the instructional exercises is always language-based. Instructional tasks can, as illustrated in Fig 16, be identified with reference to six characteristics: the goal, the input, the activity, the setting and teacher's and learner's roles.

Fig 16: Characteristics of Instructional Tasks

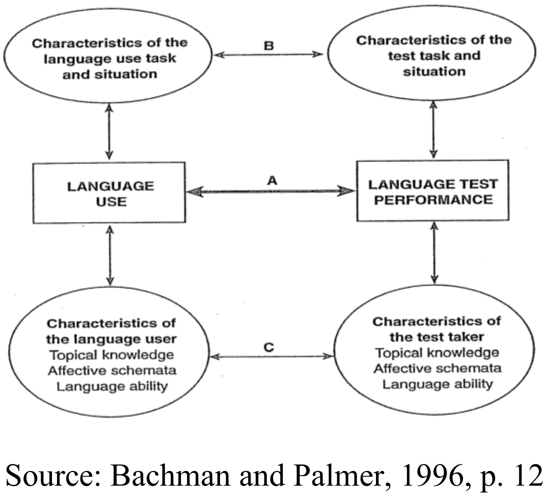


Source: Nunan, 1989, p.11

Returning now to test tasks, these can be defined as the activities that involve test takers "in using language for the purpose of achieving a particular goal or objective in a particular setting closely associated with, or situated in specific situations, goal oriented" (Bachman and Palmer, 1996, p. 44). The use of language can be manifested either in the form of oral or written responses to some stimuli; or in performing some type of instructions. Rests to mention that in test tasks examinees need to understand "what sort of result is to be achieved" (Carroll, 1993, p. 8); and by what criteria their responses are to be evaluated.

Bachman and Palmer (1996) point out that the design of test tasks needs to respond to three types of correspondence (see Fig 17). One, the characteristics of test takers should be determined on the basis of the characteristics of real language users. Two, language test performance should be outlined according to real language use (how people use language) in target language situations. Three, the characteristics of test tasks need, to a great extent, to correspond to the characteristics of target language use tasks (TLU). Summarizing the requirements of this rule, the authors write that we need to "consider that language used on tests as a specific instance of language use, a test taker as a language users in the context of a language test, and a language test as a specific language use situations" (p. 58).

Fig 17: Types of Correspondence

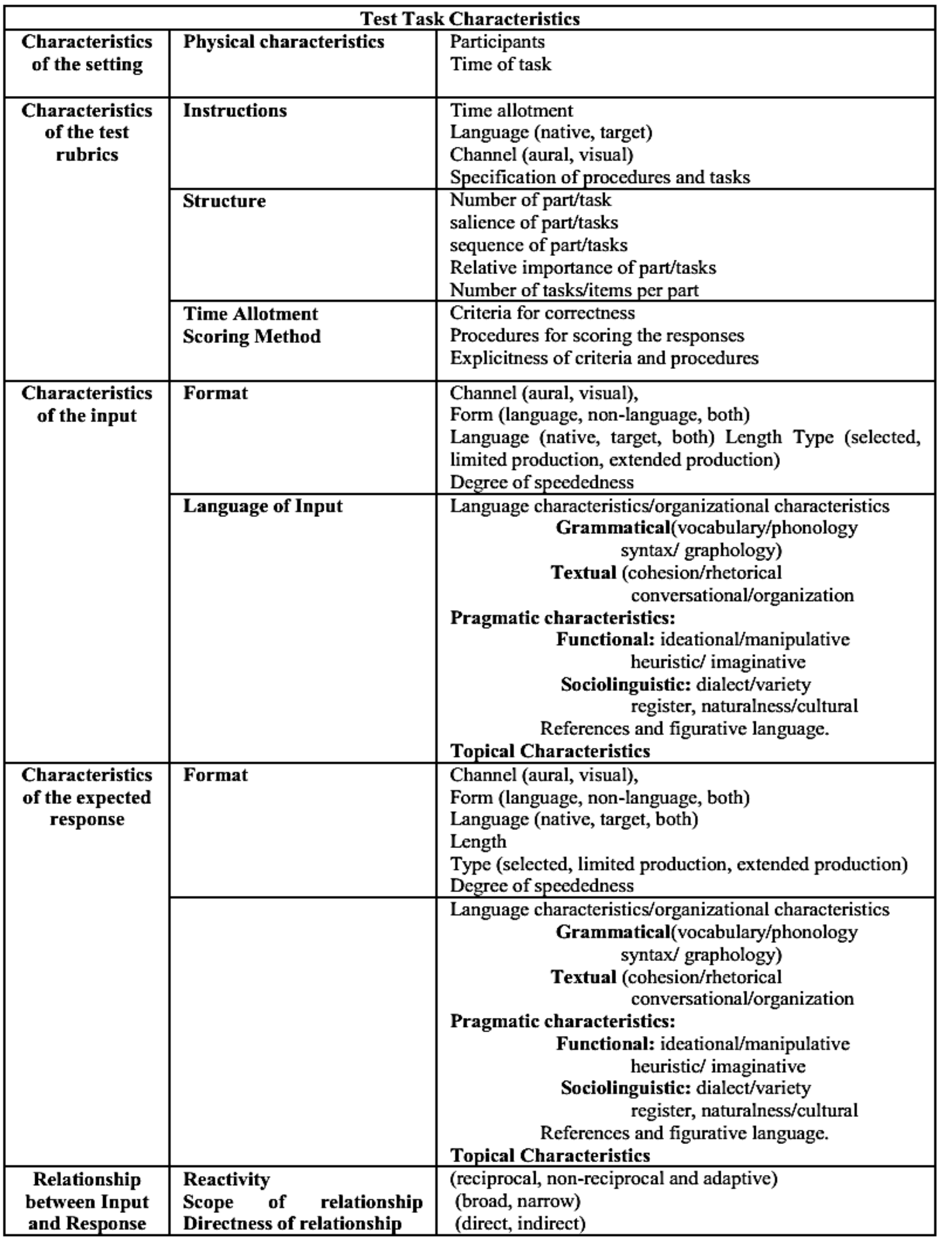

Characteristics of the language use task and situation
B
Characteristics of the test task and situation
LANGUAGE USE
A
LANGUAGE TEST PERFORMANCE
Characteristics of the language user Topical knowledge Affective schemata Language ability
C
Characteristics of the test taker Topical knowledge Affective schemata Language ability

Source: Bachman and Palmer, 1996, p. 12

3.1.4. Test Task Characteristics

Bachman and Palmer (1996) propose a framework for test task characteristics describing five aspects: the setting, the rubric, the input, the expected response, and the relationship between the input and output (see Table 6). The main aim of this framework is to enable test designers to compare "the characteristics of TLU and test tasks to assess their authenticity" (p. 47).

Table 6: Test Task Characteristics



Test Task Characteristics		
Characteristics of the setting	**Physical characteristics**	Participants Time of task
Characteristics of the test rubrics	**Instructions**	Time allotment Language (native, target) Channel (aural, visual) Specification of procedures and tasks
	Structure	Number of part/task salience of part/tasks sequence of part/tasks Relative importance of part/tasks Number of tasks/items per part
	Time Allotment **Scoring Method**	Criteria for correctness Procedures for scoring the responses Explicitness of criteria and procedures
Characteristics of the input	**Format**	Channel (aural, visual), Form (language, non-language, both) Language (native, target, both) Length Type (selected, limited production, extended production) Degree of speededness
	Language of Input	Language characteristics/organizational characteristics **Grammatical**(vocabulary/phonology syntax/ graphology) **Textual** (cohesion/rhetorical conversational/organization **Pragmatic characteristics:** **Functional:** ideational/manipulative heuristic/ imaginative **Sociolinguistic:** dialect/variety register, naturalness/cultural References and figurative language. **Topical Characteristics**
Characteristics of the expected response	**Format**	Channel (aural, visual), Form (language, non-language, both) Language (native, target, both) Length Type (selected, limited production, extended production) Degree of speededness
		Language characteristics/organizational characteristics **Grammatical**(vocabulary/phonology syntax/ graphology) **Textual** (cohesion/rhetorical conversational/organization **Pragmatic characteristics:** **Functional:** ideational/manipulative heuristic/ imaginative **Sociolinguistic:** dialect/variety register, naturalness/cultural References and figurative language. **Topical Characteristics**
Relationship between Input and Response	**Reactivity** **Scope of relationship** **Directness of relationship**	(reciprocal, non-reciprocal and adaptive) (broad, narrow) (direct, indirect)

Source: Bachman and Palmer, 1996, pp. 50-1

3.1.4.1. Characteristics of the Setting

The setting refers to "the physical and temporal test circumstances [which] include the physical characteristics, the participants, and the time of the task" (Chappelle & Douglas, 2006, p. 22). The characteristics of the physical setting include factors such as the location where the test is intended to be held, the noise level, lighting conditions, and degree of comfort (Alderson, 2000a; Bachman & Palmer, 1996). This can also be extended to the delivery of material such as pens, papers, computers, or tapes. The participants which include test takers and administrators highlight the status of each type, and how familiar they are to each other. The third element concerns the timing of the task. This aspect examines the extent to which the standardization of the test administration time is appropriate to the whole number of test takers (Bachman, 1990, 1991).

3.1.4.2. Characteristics of the Test Rubric

The characteristics of the test rubric "consist of the facets that specify how test takers are expected to proceed in taking the test"(Bachman, 1990, p.118). These include four factors: the test structure, the task instructions, the test and task duration, and the scoring procedures. The test structure specifies the number and type of tasks and how they will be combined together to form a test. Concerning the instructions, they represent "the means by which the test takers are informed about the procedures for taking the test, how it will be scored, and how the results will be used" (Bachman & Palmer, 1996, pp. 50-51). For this reason, they need to be clear and explicit. The Instructions comprise three elements: language (native or target), channel (aural or visual), and specification of procedures and tasks. Concerning the third element, it specifies to test takers the way in which they can interact with tasks. For example, should the responses be 'lengthy or brief', with or without illustrations, will be related to the other parts of the test or fully independent. The third component, 'test duration' refers to whether tests are designed in a way that allows all the test takers (whatever their level) to complete the tasks within the allotted period time. As for the scoring method, it specifies how the scores will be assigned to test takers. The scoring method stipulates three features: the criteria for correctness (objective, subjective scoring, or type of rating scales); the procedures for scoring responses (single/ double rating); explicitness of criteria and procedures (the extent to which the two previous factors will be understandable and unambiguous according to test takers). Emphasizing the importance of the 'explicitness of criteria and procedures', Alderson (2000a) argues that if test takers "are to perform to the best of their language ability, [they] need to know how they will be judged" (p. 151)

3.1.4.3. Characteristics of the Input

According to Bachman and Palmer (1996) the input "consist[s] of the material contained in a given test task or TLU task, which test takers or language users are expected to process in some way and to which they are expected to respond" (p. 52). This material is characterized in terms of format and language. The format which refers to the way test tasks are presented to the examinees includes the channel (aural/ visual), the form (language/ non language such as pictures or gestures), language (native, target, or both), length (short, long), type (item, prompt) and degree of speededness which refers to "the rate at which the test taker or language user has to process the information in the input" (p. 53). The last point has to do with the vehicle used to present the material. This can be live such as in lectures designed for note taking or in listening comprehension; and it could also be reproduced if it were intended to be presented by audio or video. Concerning language characteristics, these aspects delineate the components of language competence to be included in the test; for instance, grammatical, textual, pragmatic, or functional knowledge.

3.1.4.4. Characteristics of the Expected Response

The expected response refers to "what the test developers intend that test takers do in response to the [task] they have attempted to set up by means of the rubric and the input" (Douglas, 2000, p. 62.). On their part, Bachman and Palmer (1996) define the characteristics of the expected response as "the physical response we are attempting to elicit by the way the instructions have been written, the task designed and by the kind of input provided" (p. 53). Seeing that some test takers may not understand the instructions, or may respond in a way that is not expected, test developers distinguish two types of responses: expected responses and actual responses. The former refer to what item writers expect of test takers to respond; while the latter may include unexpected information on the part of test takers**.**

3.1.4.5. Relationship between the Input and Response

This feature describes the relationship or the interaction between the input and the expected response with respect to three features: reactivity, scope, and directness. The first characteristic refers to "the degree to which the input can be altered in light of the responses of the language user" (Douglas, 2000, p.63) in terms of reciprocal, non-reciprocal, or adaptive tasks. In reciprocal tasks such as dialogues, interviews, conversations, the test taker receives feedback from the interlocutor on the relevance of the response; and in its turn, the response of the test taker can affect the input provided by the interlocutor. Conversely, when the feedback is not available such as in listening to taped passages, or writing messages, we can speak of non-reciprocal tasks. Concerning the delivery of adaptive tasks, the process starts with the administration of medium difficulty tasks; and it is the test taker's response that determines the extent of difficulty of the subsequent task. If these responses are fairly good, the next task will be of more difficulty; but if test takers fail to do the task, the following input will be a little bit easier.

As far as the scope of the relationship is concerned, Bachman and Palmer (1996) define it as " the amount or range of input that must be processed in order for the test taker or language user to respond as expected" (p. 55). The scope of relationship can be identified as 'broad' or 'narrow'. The former involves a lot of input, such as questions that require examinees to provide a summary to a given text; however, narrow scope relationship, such as matching, or multiple choice prompts, requires test takers to provide only a limited amount of input.

Concerning the directness of the relationship between the input and the expected response, it can be defined as the "degree to which the responses depend on the input as opposed to the language user's own …background knowledge" (Douglas, 2000, p.66). If the tasks include feedback provided in the input, we can talk of direct relationship. This can, for example, occur in listening comprehension tests where the completion of tasks depends fully on the read or taped input. On the contrary, if the responses are not provided in the input and test takers have "to rely on information in the context or in [their] own topical knowledge" (Bachman & Palmer, p.56), we consider this relationship as indirect.

3.1.5. Describing Test Takers' Characteristics

In order to make reliable and valid inferences about the examinees' language abilities, language testers stress that not only should test tasks correspond to real life tasks, but the characteristics of test takers should also correspond, to a great extent, to the characteristics of real-life language users. Test takers' characteristics such as personal attributes, topical knowledge, affective schemata, and levels of language ability refer to the factors that do not form a part of the construct that is intended to be measured, but which do have their impact on the interpretations that we are supposed to provide for students' scores. The first set of correspondence leads to task authenticity; however, the second one reinforces the concept of interactiveness. The failure to consider one of these concepts will question the concept of test usefulness as a whole (Bachman and Palmer, 1996).

3.1.5.1. Personal Characteristics

Personal Characteristics can be defined as the individual's "attributes that are not part of the test takers' language ability but which may still influence their performance on language tests" (Bachman and Palmer, 1996, p. 64). The authors list seven characteristics which include factors such as age, gender, nationality, immigrant status (immigrant or international student), native language, level and type of general education and prior experience with a given test.

3.1.5.2. Test Takers' Topical Knowledge

Topical knowledge refers to the type of knowledge that test takers have previously acquired from their real-life experience and which they bring to a given testing context (Luoma, 2004). Test tasks that are built on the assumption that topical knowledge of test takers is homogeneous tend to fall in the preference of one type of examinees at the expense of the other type. Conversely, when test takers are considered to have diverse topical knowledge, the content of the test should cover different areas of interest. As far as formal education is concerned, topical knowledge can be related to students' fields of specialty. In Algerian secondary education, for instance, tests which include information on mechanics may fall in the preference of students of mechanical engineering. In the same way, topics on business, banking, marketing, or trade may fall in favor of economy and management or accountancy specialties at the expense of the other branches.

3.1.5.3. Predictions about Test Takers' Potential Affective Responses

Affective schemata can be defined as "the emotional correlates of topical knowledge…[which] provide the basis on which language users assess, consciously or unconsciously, the characteristics of the language use task and its setting in terms of past emotional experience in similar contexts" (Bachman & Palmer, 1996, p.65). Affective schemata can determine the way in which test takers will interact with tasks. In other words, these schemata can either facilitate or inhibit the flexibility of students in responding to tasks. To promote the feeling of comfort and security for the purpose of positive interaction between test takers and test tasks, language testers recommend that tests should include or at least start with "tasks at a level of language with which the test taker feels comfortable and at ease" (p. 66).

3.1.5.4. General Level and Profile of Language Ability

This type of characteristics concerns test takers' levels of language ability in performing different tasks and skills. Listing these characteristics enables, on the one hand test developers to design appropriate tests; and to identify the areas of language ability (components of communicative competence) within which students can perform better on the other. Feedback on test takers' levels of language ability can be obtained in the pretesting phase of test administration (see stage three 'Test Administration').

3.1.6. Test Usefulness

Test usefulness can be defined as a function comprising several qualities such as reliability, construct validity, authenticity, interactiveness, impact and practicality "all of which contribute in a unique, but interrelated ways to the overall usefulness of a given test" (Bachman & Palmer, 1996, p. 18). Test usefulness is built upon three principles: (1) what should be reinforced in test development is the overall conception of usefulness rather than its individual components; (2) test qualities should be evaluated in terms of their combined effect on the test; and (3) "the appropriate balance among the different qualities cannot be prescribed in general, but must be determined for each specific testing situation (p. 18).

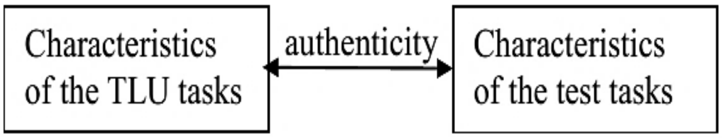
3.1.7. Components of Test Usefulness

As we have mentioned above, test usefulness is a framework which consists of six qualities: authenticity, interactiveness, practicality, impact, reliability, and construct validity. In this section, we will consider the first four qualities; however, because of the importance of reliability and construct validity to this research, we will introduce them in chapters four and five respectively.

3.1.7.1. Authenticity

Authenticity which can be defined as “the degree of correspondence of the characteristics of a given language test task to the characteristics of a TLU task" (Bachman, 1991 p. 111) enables us to establish some type of relationship between test tasks and the domain to which we intend to generalize the interpretations of the scores obtained by test takers (see Fig 18). Additionally, this can help test developers reinforce the concept of interactiveness between test takers and the content of tasks because when examinees feel that test tasks are, to a great extent, similar to TLU tasks, their motivation for working will be maximized.

Fig 18: Authenticity


Characteristics of the TLU tasks
authenticity
Characteristics of the test tasks

Source: Bachman and Palmer, 1996, p. 23

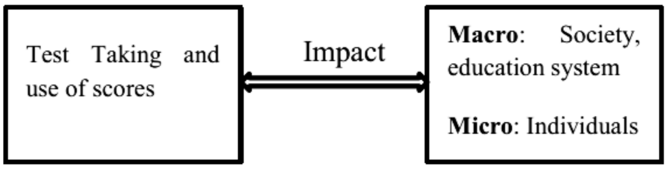
3.1.7.2. Interactiveness

Interactiveness refers to the engagement of test takers' sectors of language knowledge, background or topical knowledge, strategic competence and their affective variables by the test input (Bachman and Palmer, 1996; Chappelle & Douglas, 2006; Douglas, 2000). McNamara and Roever (2006) inform us that test takers usually "feel frustrated by the lack of opportunity" (190) to be engaged by the test tasks which makes "the levels of anxiety experienced depress [their] performance" (p.190).

3.1.7.3. Test Impact

Test impact refers to "the wider effect of tests on the community" (McNamara, 2000, p. 74). In the same way, Bachman and Palmer (1996) consider it as the effect of tests “on society and educational systems and upon the individuals within those systems” (p. 30). This quality operates, as illustrated in Fig 19, at two levels: micro and macro levels. The former refers to the individuals who can be affected by test scores or the purposes for which the scores will be used. The macro level concerns the impact of tests on the educational system and on the society as a whole.

Fig 19: Test Impact


Test Taking and use of scores
Impact
Macro: Society, education system
Micro: Individuals

Source: Bachman and Palmer, 1996, p. 30

Language testers distinguish between test impact and washback (McNamara, 2000; McNamara & Roever, 2006; Messick, 1996; Wall, 1997, 2012). According to them, the scope of the former is wider, in that the latter can be considered as a special instance of the former. Explaining the nature of this relationship, Wall (1997, as cited in Bailey, 2004) thinks of test impact to be “any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole” … whereas washback (or backwash) can be defined as “the effects of tests on teaching and learning” (p. 291).

3.1.7.4. Practicality

Practicality which Bachman and Palmer (1996) define 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). Practicality, as it is illustrated in Fig 20**,** examines whether the existing human and material resources available for a live testing situation can meet the ones prescribed in the blueprints. For example, if the required resources exceed the available ones, we can consider the test as impractical unless more recourses will be allocated. On the contrary, if test developers conclude that the available resources can meet what is specified in the specifications, we can say that the test design and use are practical.

Fig 20: Relationship between Available and Required Resources

Available resources
Practicality=
Required resources
If practicality$\geq$ 1, the test development and use is practical
If practicality$<$ 1, the test development and use is not practical

Bachman and Palmer, 1996, p. 36

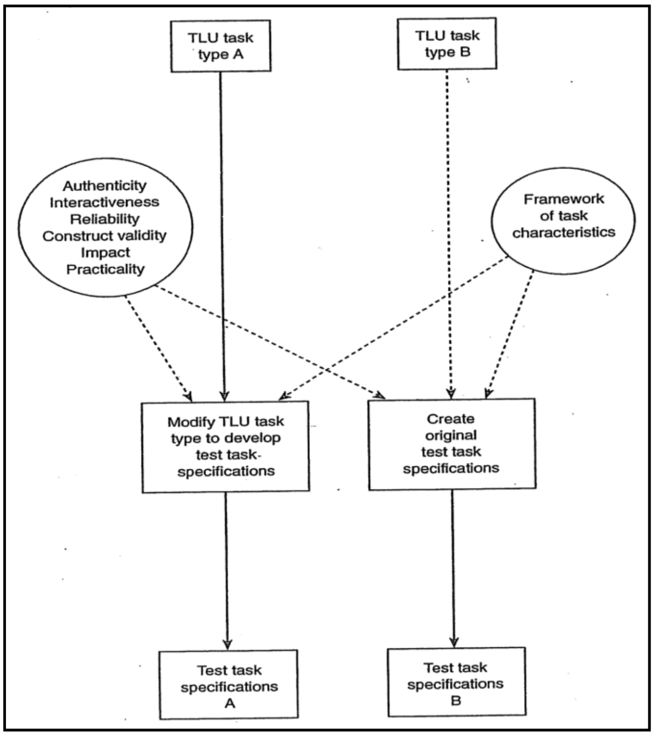
Bachman and Palmer classify these recourses into three broad types: human recourses, material recourses and time. Human recourses include test developers, item writers, clerical support, test administrators, proctors, raters, security forces and so on. Material resources include space such as rooms; equipment such as typewriters, computers or tapes; material such as tables, chairs, papers, pens and the like. The third type of recourses concerns time. Practicality defines time in relation to the development of the test and its tasks. Developmental time starts from the beginning of writing the test and concludes with score reporting; while the 'time for specific tasks' refers to the period of time allocated to test sections, such as the time required for writing the blueprint, the individual tasks, the time devoted to test administration or to the scoring process.

In brief, test usefulness refers to the extent to which the test can be used for what it has been intended (Bachman, 1990, 1991, 2000, 2007; Bachman and Palmer, 1996). The usefulness of a given test is measured in terms of six aspects: authenticity, or correspondence between TLU tasks and tasks; practicality or the extent to which the required resources can meet the ones specified in the blueprints; interactiveness between test tasks and test takers' language knowledge; consistency of scoring; construct validity; and the impact of test scores on participants and institutions. Concerning, the plan for evaluating the availability of required resources, this has been introduced within the concept of practicality of test usefulness.

3.2. Stage Two: Operationalization

Operationalization refers to the process of using the information collected in the design stage for the writing of tasks and compiling them into comprehensive tests. According to Bachman and Palmer (1996), this process takes place during two consecutive phases. The first phase involves writing specifications for individual tasks; and phase two concerns the design of a blueprint for assembling the tasks into a comprehensive test. The operational steps for writing tasks can result from one of two ways (see Fig 21): (1) by modifying and transforming TLU tasks into task specifications, or (2) simply by creating new tasks. This process is finally compared against the checklist of task characteristics and pre-evaluated against the qualities of usefulness. In short, the operationalization stage "describes how an entire test involving several tasks is assembled, and how the individual tasks are specified, written and scored. The outcome of the operationalization phase is both a blueprint for the entire test including scoring materials and a draft version of the actual test" (Purpura, 2004, p. 167).

Fig 21: Developing Task Specifications


TLU task type A
TLU task type B
Authenticity
Interactiveness
Reliability
Construct validity
Impact
Practicality
Framework of task characteristics
Modify TLU task type to develop test task specifications
Create original test task specifications
Test task specifications A
Test task specifications B

Source: Bachman and Palmer, 1996, p. 87.

3.2.1. Test Blueprints

As indicated above, operationalization according to Bachman and Palmer (1996) starts with the design of a test blueprint which is a "detailed plan that provides the basis for developing an entire test" (p.176). Unlike most language testers who see that the design of test specifications precedes the design of item specifications (Alderson, et al., 1995; Gronlund, 1977; Miller, Linn & Gronlund, 2009; Osterlind, 2000), Bachman and Palmer (1996) stress that "in developing a blueprint, we begin with the specifications for the various task types to be included, and determine how best to combine these in a test" (p.

176). In short, according to the authors, test task specifications stipulate the way for writing individual tasks, whereas test blueprints determine how these tasks are compiled into a single test.

3.2.1.2. Components of Task Specification

Task Specifications, as illustrated in Table 7, describe six aspects built around a guiding purpose. The purpose is then extended to delineate the specific construct(s) to be tested; for the broad construct is usually described at the level of the test framework (see the Design Stage). Then, the specifications define the characteristics of task setting, determine the time allotted to do the tasks, specify the way the language of instructions are to be written, describe the relationship between the input and the expected response, and how these responses will be scored.

Table 7: Bachman and Palmer's Test Blueprint

Bachman and Palmer's Test Blueprints : Design Evaluation and use	**Task specifications**	Task Purpose
		Construct description
		Characteristics of the test task setting
		Time allotment
		Instructions for responding to the task
		The characteristics for setting of the task
		Scoring Method
	Characteristics that pertain to the structure of the test	Number of parts/tasks
		salience of parts/tasks
		sequence of parts/tasks
		Relative importance of parts/tasks
		Number of tasks per part
	Qualities of usefulness	Authenticity / interactiveness / reliability / Construct validity/ impact / Practicality
	Use	To permit the development of other tests or parallel forms of the test with the same characteristics.
		To evaluate the intentions of the test developers.
		To evaluate the correspondence between the test as developed and the blueprint from which it was developed.
		To evaluate…the correspondence between characteristics of the TLU tasks and those of the test task

Organized from Bachman and Palmer, 1996, pp. 172-3,6, 7.

On their part, test blueprints delineate the characteristics of task arrangement and how these tasks will be compiled into a single test. The test blueprints determine "the characteristics that pertain to the structure of the test [,and specify] the number of parts/tasks, the salience of parts/tasks, the sequence of parts/tasks, the relative importance of parts/tasks, and the number of tasks per part" (Bachman and Palmer, 1996, p.176).

3.2.1.3. Strategies for Writing Tasks

Bachman and Palmer propose two types of strategies for writing tasks. One type concerns the modification of some TLU tasks and incorporating them in a given test; and the other concerns the creation of original tasks. The choice of one strategy over the other depends on the testing situation and not on the strategy itself. In some situations of ESP, tests such as measuring trainees' ability to communicate with air traffic controllers, we can simply modify the real-life target tasks and transform them into test tasks (Alderson, et al, 1995). In other situations where the specific TLU tasks in real-life may not be appropriate for a given testing situation, we can resort to the creation of new useful tasks. In the same way, the qualities of usefulness can be maximized by the implementation of both types of strategies. In this context, language testers do not "recommend one strategy over the other, since both have the potential for yielding useful tests. Furthermore, whether the test developer decides to use one or the other or both will depend on the situation" (Bachman & Palmer, 1996, p. 174).

2.3. Stage Three: Test Administration

Test administration refers to the process of “giving the test to a group of individuals, collecting information, and analyzing this information, for two main purposes: 1) assessing the usefulness of the test, and 2) making the inferences or decisions for which the test is intended” (Bachman & Palmer 1996, p. 91). This process takes place during two phases: pretesting (try-out) and live (operational) testing. Phase one concerns the collection of feedback for the purpose of item revision and modification and phase two takes "place when the test is used for the purposes for which it was designed"(p. 245). Operational administration allows test developers to make inferences about test takers' language ability, and enables test validators to evaluate the usefulness of the test in order to investigate whether the decisions made by test users are meaningful (valid).

3.3.1. Item Tryout

Item writing derives from a number of considerations such as the delineation of the purpose of the test, the description of the target language use tasks, the design of task specifications and the expertise of item writers. However, no matter how important these features are, "the literature concerning language tests suggests that the examiners' assumptions regarding what they test and their expectations from the respondents often do not match the actual processes which the respondents undergo during testing" (Nevo, 1989, p. 20). For this reason, language testers accentuate that operational assessment needs to be preceded by pretesting and item tryout.

Piloting the test which "refers to all trials of an examination that take place before it is launched, or becomes operational or 'live'" (Alderson, et al., 1995, p. 72) is meant to anticipate the difficulties that may rise during live testing. This is because however "well designed an examination maybe, and however carefully it has been edited, it is not possible to know how it will work until it has been tried out on students" (p. 72). In the same way, Bachman and Palmer (1996) argue that "it is impossible…[to] know how problem free the administrative procedures are without trying them out" (p. 236). For this reason, the Measurement Profession ([AERA], [APA],& [NCME], 1999) strongly recommends in Standard 3.8 that the "test review process should include empirical analyses" (p. 44)

Empirical observation and field testing enable test developers to obtain three types of feedback that expertise in the field cannot afford: feedback about test takers' levels of language ability, feedback about test usefulness and test items; and feedback about test taking strategies and administering procedures ([AERA], [APA],& [NCME], (1999; Alderson et al., 1995). Concerning the first type, test tryout gives us a general overview on examinees' levels of language ability and helps us determine the scope of the constructs that we intend to measure. Additionally, this allows us to discover the components of language ability (organizational/ strategic/ pragmatic/ interactional and so on) that test takers may excel in. Concerning feedback about test usefulness, pretesting enables us to engage in initial evaluation of the test against Bachman and Palmer's six-componential usefulness framework. More importantly, tryout enables us to determine the facility value (F.V.) and discrimination indices (D.I.) of items. Facility value is concerned with the measurement of "the level of difficulty of an item, and the discrimination index measures the extent to which the results of an individual item correlate with results from the whole test" (Alderson et al, 1995, p. 80).

3.3.1.1. Information on Item Facility Value

For a better understanding of the notion of 'item facility value' (F.V.) or item difficulty (ID), let us consider Alderson et al's (1995) explanation of the issue:

> If there are 300 students and 150 of them get the item right, the F.V. of the item is 150/300, which is 50%....This simple measure immediately gives item writers some idea of how easy the item is for the trial sample of students. If 6/300 people get an item right, the F.V. is 2% and it is clear that the item is very difficult indeed. Similarly if the F.V. is 95% (285/300), the item is very easy. Such easy or difficult items are not very informative since they tell us little about the varying levels of ability of the trial group. If examiners…want the students' scores to range from very high to very low, then, they will select items which are as near to an F.V. of 50% as possible because such items provide the widest scope of variation among the individual students (p. 81).

3.3.1.2. Discrimination Indices

Despite the fact that facility value (FV) gives us an overview of item difficulty, this criterion alone does not provide sufficient information which can be used as a basis for the decision to accept or reject an item in a test (Henning, 1998). For this reason, educational measurement specialists, resort to discrimination indices (Gronlund, 1977). Discrimination can be defined as the "tendency of the item to be answered correctly by test takers who are generally strong in the skills or type of knowledge the item is intended to measure and to be answered incorrectly by test takers who are not" (Livingston, 2006, p.422). Consequently, if an item is working well and:

> discriminates between students at different levels of ability….We should expect more of the top-scoring students to know the answer than the low-scoring ones. [But] if the strongest students get an item wrong, while the weakest students get it right, there is clearly a problem, and it needs investigating" (Alderson, et al., 1995, p. 81).

To illustrate this point, suppose that the F.V. of an item is of 50% which means that half of the number of students got it right. However, after we have examined the issue we found that the top scoring students in the test got that item wrong. This means that the item in question has failed to discriminate between the students who ranked at the top of the list and those who ranked at the bottom of the list. For this reason, in addition to the investigation into item difficulty, item developers need to consider the item discrimination indices as well.

3.3.1.3. Feedback about the Administration Procedures

Test tryout enables us to anticipate the problems that may rise during live testing, and to have control over the administering procedures as well. We can, for example, ensure that the testing procedures will be consistent with the ones recommended in the blueprint (Bachman, and Palmer, 1996). Furthermore, We can obtain information on the quality of proctors and the way they communicate test instructions to test takers. Additionally, this process can provide us with feedback about time allocation and test security.

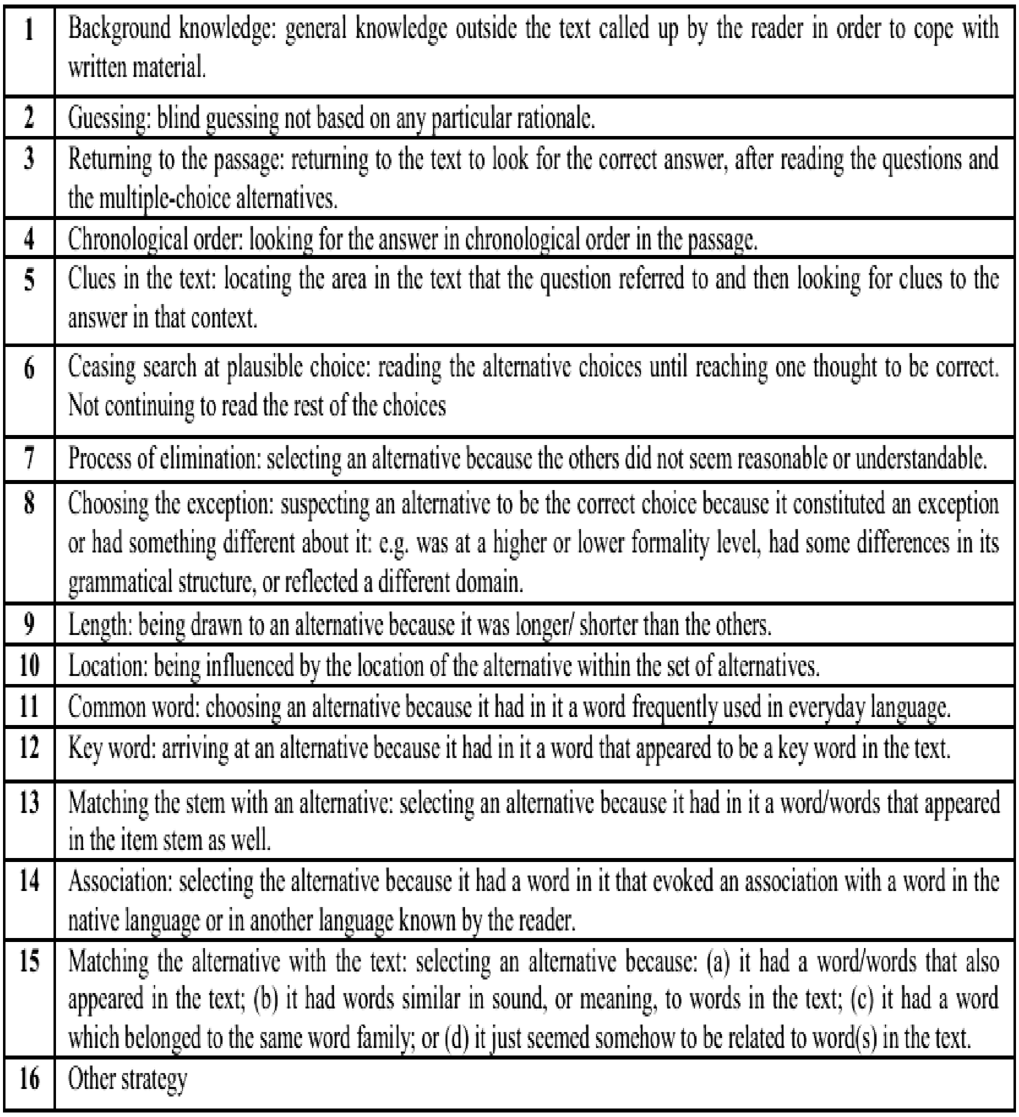
3.3.1.4. Other Types of Feedback

Other sources of feedback such questionnaires, observation or interviews are also used to elicit feedback in the pretesting phase (Nevo, 1989). The main purpose of these data gathering tools is to collect information about test taking strategies. Questionnaires related to this type of feedback fall into four formats: multiple-choice questions, open-ended questions, yes-no questions and rating scales (Alderson, et al., 1995). What is worth mentioning here is that interviews and observations do not differ from the ones administered in empirical research.

3.3.1.5. Multiple-Choice Questionnaires

The most well-known multiple-choice questionnaire is the one developed by Nevo (1989). This questionnaire which includes sixteen (16) questions requires test takers to tell which strategy they have used in responding to each item. For example, "after you answer each item, check which of the following strategies you used to answer the item" (Bachman and Palmer, 1996, p.241). Students can look at the list of strategies and tick the number which falls in their preferences (see Table 8). The collected information will used in the design of test items.

Table 8: Multiple-Choice Questionnaires



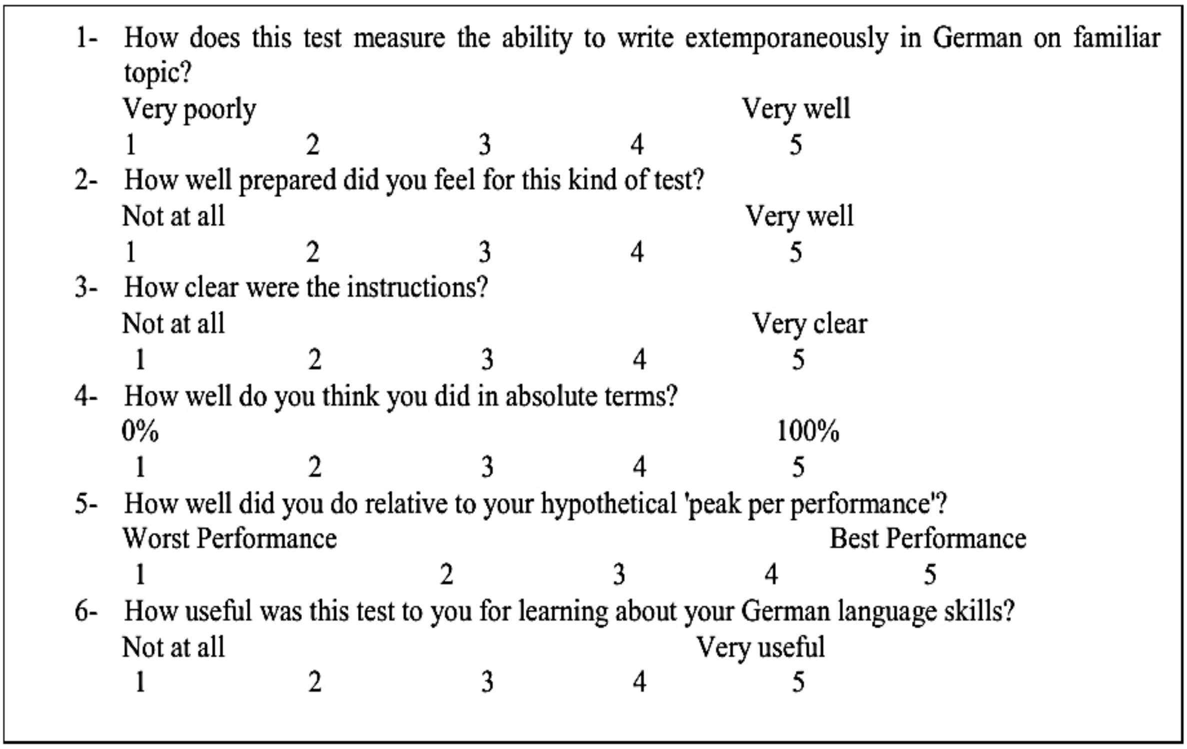
1	Background knowledge: general knowledge outside the text called up by the reader in order to cope with written material.
2	Guessing: blind guessing not based on any particular rationale.
3	Returning to the passage: returning to the text to look for the correct answer, after reading the questions and the multiple-choice alternatives.
4	Chronological order: looking for the answer in chronological order in the passage.
5	Clues in the text: locating the area in the text that the question referred to and then looking for clues to the answer in that context.
6	Ceasing search at plausible choice: reading the alternative choices until reaching one thought to be correct. Not continuing to read the rest of the choices
7	Process of elimination: selecting an alternative because the others did not seem reasonable or understandable.
8	Choosing the exception: suspecting an alternative to be the correct choice because it constituted an exception or had something different about it: e.g. was at a higher or lower formality level, had some differences in its grammatical structure, or reflected a different domain.
9	Length: being drawn to an alternative because it was longer/ shorter than the others.
10	Location: being influenced by the location of the alternative within the set of alternatives.
11	Common word: choosing an alternative because it had in it a word frequently used in everyday language.
12	Key word: arriving at an alternative because it had in it a word that appeared to be a key word in the text.
13	Matching the stem with an alternative: selecting an alternative because it had in it a word/words that appeared in the item stem as well.
14	Association: selecting the alternative because it had a word in it that evoked an association with a word in the native language or in another language known by the reader.
15	Matching the alternative with the text: selecting an alternative because: (a) it had a word/words that also appeared in the text; (b) it had words similar in sound, or meaning, to words in the text; (c) it had a word which belonged to the same word family; or (d) it just seemed somehow to be related to word(s) in the text.
16	Other strategy

Source: Nevo, 1989, pp. 214-215

3.3.1.6. Rating Scales

The other format which is used in obtaining feedback about examinees' test taking strategies refers to rating scales (see Fig 22). These scales differ from the ones designed to score the 'written expression' tasks. These are exclusively used for collecting information about test taking strategies (Bachman & Palmer, 1996).

Fig 22: Rating Scale for Obtaining Feedback about Test taking Strategies



1- How does this test measure the ability to write extemporaneously in German on familiar topic?
Very poorly Very well
1 2 3 4 5

2- How well prepared did you feel for this kind of test?
Not at all Very well
1 2 3 4 5

3- How clear were the instructions?
Not at all Very clear
1 2 3 4 5

4- How well do you think you did in absolute terms?
0% 100%
1 2 3 4 5

5- How well did you do relative to your hypothetical 'peak per performance'?
Worst Performance Best Performance
1 2 3 4 5

6- How useful was this test to you for learning about your German language skills?
Not at all Very useful
1 2 3 4 5

Source: Bachman and Palmer, 1996, p. 242-3

3.3.2. Live Administration

Phase two in test administration concerns operational testing. At this stage, tests are administered for one purpose: it is to make inferences about test takers language ability. At the same time, the resulted scores from this stage enable test validators to examine whether the interpretation of test scores are valid (meaningful); and test users to investigate the appropriateness of the decision they have made on the basis of the obtained scores.

Conclusion of the Chapter

Bachman and Palmer organize the process of developing tests into three linear stages: design, operationalization and administration. The design stage identifies and describes the features that enable us to ensure that the language ability to be measured and the tasks to be designed correspond to a great extent to the abilities of language users in real target language situations. The operational stage describes how to write tasks and how to compile into a comprehensive test. Finally in the administration stage, phase one describes the processes implemented in test tryout; and phase two lays out the procedures for live test delivery.

Chapter Four

Investigating Rater Reliability

Chapter Four

Investigating Rater Reliability

Introduction

In Chapter III, we have introduced Bachman and Palmer's plan of test usefulness which includes six qualities: authenticity, interactiveness, impact, practicality, reliability, and construct validity. We have also pointed out that this plan is used during the construction of language tests for the purpose of anticipating errors of measurement; conducting an initial process of test evaluation; and for ensuring that test scores will be used for the purposes for which they have been intended. In our introduction of this plan, we underlined that the qualities of reliability and validity will be provided in separate chapters. This is because most language testers and educational measurement specialists emphasize that we would better implement these criteria in the post-testing phase as well to scrutinize the extent of the rating consistency; and/or to conduct a validation process to examine the extent to which test score interpretations and uses are real indicators of the language ability being measured (Gronlund, 1977; McNamara, 1996: Weigle, 2002).

4.1 Definition of Reliability

Extensive research has been devoted to the conceptualization of reliability in the literature of measurement (Miller, Linn, & Gronlund, 2009). This conceptualization has associated the concept to the criteria of stability, consistency of scoring and precision of measurement which have "to do with the extent to which any given observation report provides essentially the same information, or generalizes, across different aspects, or facets, of the observation and reporting procedure” (Bachman, 2008, p. 170). Cronbach (1947) defines reliability as “the degree to which the test score indicates unchanging individual differences in any traits" (p.5). In the same way, Guilford (1954) considers it as "the proportion of true variance in obtained test scores." (p.350). The proportion of true variance without which one cannot speak of reliability, is explained by Lado (1961) when he asks, "does the test yield dependable scores in the sense that they will not fluctuate very much so that we may know that the score obtained by a student is pretty close to the score he would obtain if we gave the test again?" (p.33). The previous definitions have been endorsed by the Standards of Educational and Psychological Testing ([AERA], [APA], & [NCME], 1999) emphasizing that reliability "data ultimately bear on the repeatability of the behavior elicited by the test and the consistency of the resultant scores" (pp.25&31). However, when measurement practices of the same traits or constructs yield changing or fluctuating information, we can speak of measurement errors.

4.2 Errors of Measurement

Errors of measurement can be defined as:

> the amount of deviation an examinee's score on a set of test items would exhibit if the test was administered to that examinee an infinite number of times, under identical conditions. The more those scores disperse, the greater the error of measurement (Osterlind, 2002, p. 255).

According to [AERA], [APA],& [NCME], (1999) these errors represent "the hypothetical difference between an examinee's observed score on any particular measurement and the examinee's true or universe score" (p. 25). To illustrate this point, let us consider the following example which is provided in Bachman (2004a). Suppose that we have administered a test a number of times to see whether it can produce consistent measures of the language ability we intend to assess. If the test takers get the same results under the same conditions, we can say that our ratings are reliable. Contrariwise, if the examinees obtain scores that are much lower or higher than the scores they have obtained in the first testing session, we assume that these scores include a component of error.

According to [AERA], [APA], & [NCME], (1999) measurement errors "reduce the usefulness of measures…limit the extent to which test results can be generalized beyond the particulars of a specific application of the measurement process, [and] reduce the confidence that can be placed in any single measurement" (p. 27).

In real life no person is given a test for unlimited number of times; and errors of measurement are generally estimated from a single administration (Osterlind, 2002). For example, in the 'Baccalauréat' examination test takers' input is double-rated; and we expect that the scoring of the two raters will be equal. If it is not equal, we assume, as Kane (2010) puts it, that "our data [scores] are inconsistent" (p.5) and need to be adjusted.

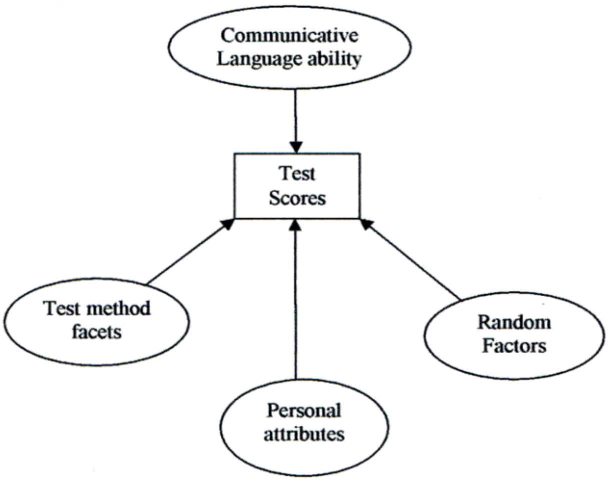
When we administer a test and correct it, we expect the resulted scores to reflect the abilities that we have tested and nothing else. This view is consistent with Messick (1995) who considers the term score to refer to "any coding or summarization of observed consistencies or performance regularities on a test, questionnaire, observation procedure, or other assessment devices such as work samples, portfolios, and realistic problem simulations" (p.741). This definition constrains the term to rating consistencies or performance regularities. However in practice, there are other factors that may influence these scores be it in the positive or in the negative sense. In addition to the constructs being measured, these scores may be affected by the candidates' personal characteristics, their topical knowledge or affective schemata, the construct irrelevant variances, the characteristics of the setting (testing conditions), the scoring criteria and raters' leniency or severity (Brown, 2005, 2012; Fulcher, 2003, 2010; Fulcher & Davidson, 2007, 2012).

Due to the fact that the main concern of reliability is to ensure that test scores are real indicators of the abilities to be measured; and that these scores will be free from measurement errors, this, on the one hand, calls us to survey the source of errors which is supposed to affect these scores; and to investigate how psychometric theories describe and calculate test scores on the other.

4.3 Source of Score Variance

The factors affecting test scores can be classified into four major categories (Bachman, 1990, 2004a; Bachman, & Palmer, 1996). These, as it is illustrated in Fig 23, include the different sectors of the language knowledge we intend to measure; test takers' personal features which do not constitute a part of the construct that we want to assess; criteria relevant to test tasks and rating procedures; and finally, unpredictable random errors (Bachman, 2004a).

Fig 23: Factors Affecting Test Scores



Source: Bachman, 1990, p. 165

Concerning the areas of language ability that we intend to assess, the score variance is related to test takers' different levels of language competence. Differences in scores related to this factor should not be considered as a source of error. On the contrary, this variance is referred to as 'reliable variance'. According to Bachman (2004a) "differences in test takers' performance will be related to differences in test takers' levels of ability

[and] test score variance that is associated with this factor is thus considered to be ‘reliable’ variance” (p.155).

As for the personal characteristics that do not form a part of the ability we want to assess, these include test takers' stable attributes such as differences in age, gender, cognitive abilities, educational, cultural, as well as background knowledge. The type of variance related to these characteristics is systematic (test bias) and not considered as measurement errors since the candidates who differ on these attributes may also perform differently on the test.

The third source of variance is related the test method characteristics and the testing procedures. The impact of these factors on the examinees is not the same. For example, if we consider the bias related to tasks, we can find examinees who prefer multiple-choice tests and there are others who do well on tests that require them to construct their own responses. This means that task design can fall in the advantage of one type of examinees at the expense of the other type. The other factor concerns the testing procedures such as test administration, the time allotted to the test items, as well as the human and material resources. When these elements are not standard, test takers' scores will certainly be affected and bear some source of variance.

The fourth factor is called random errors. Unlike systematic errors which affect only one group of test takers, the impact of random errors is unpredictable. Random errors fall into two main categories. There are errors that are rooted within test takers themselves, and errors that are external to them. The first category includes “fluctuations in the levels of an examinee’s motivation, interest, or attention and the inconsistencies application of skills are clearly internal factors that may lead to score inconsistencies” ([AERA], [APA],

& [NCME], 1999, p. 26). The second category has to do with test administration, scorer subjectivity, scoring procedures, as well as intra rater and inter rater inconsistencies.

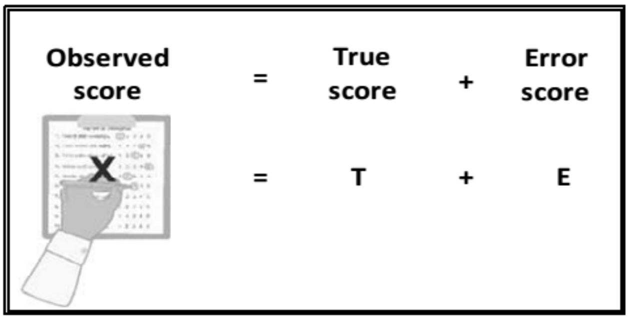
In summary, the factors which can affect test scores consist of reliable variances (differences in levels of language ability), test bias (related to personal attributes), systematic errors (related to test difficulty, test administration and scoring criteria) and unpredictable errors (random, or measurement errors). Bachman (1990) points out that the investigation of reliability responds to two main questions "how much variance in test scores is due to measurement error? and ‘how much variance is due to factors other than measurement error?”(p.238). Bachman’s second question refers to the systematic errors and test bias which can affect only one type of examinees. These can be lifted or at least reduced by minimizing the source of bias. As far as measurement errors are concerned, these can be controlled by the standardization of the testing and scoring procedures, rater training, the increase in the number of observations, as well as the reinforcement of intra rater and inter rater reliabilities (Kane, 2010, 2012a).

4.4. Computation of Test Scores

Classical Test Theory (CTT) theorizes that the scores obtained by test takers reflect the abilities that we want to assess; and other factors that do not form a part of these abilities but which can affect these scores in both senses: positively or negatively (Brennan 1997, 2001, 2010, 2013). For the purpose of measuring the extent to which these scores can be considered as indicators of the traits being assessed, CTT identifies three types of scores: observed, true and error scores. Observed or raw scores refer to the marks that test takers actually obtain as a result of their performance on real-life tests (Osterlind, 2002); or the scores "obtained on a test before any adjustment, transformation, weighting, or scaling is done"(Henning, 1988, p. 196). The true score, according to the author, refers to "the total score minus the cumulative penalties due errors [or] the actual score an examinee would be expected to obtain if no error of measurement were present at the time of testing or scoring" (pp. 197& 198).

In order to compute examinees’ true scores, we need to eliminate the features that influence them. Describing this process, Brennan (2010) assumes that “one can define *T* as the expected value of the observed scores *X*, which leads to the expected value of *E* being zero[or] one can define the expected value of *E* as zero, which leads to *T* being the expected value of *X*” (p.3[Italics in original]). This process can be explained by the following formula where (x) is the observed score, (t) is the true score and (e) is the error score: X= T +E. The illustration of the formula is included in Fig 24.

Fig 24: Computation of Test Takers' True Scores



Source: Tavakoli,2012, p. 62

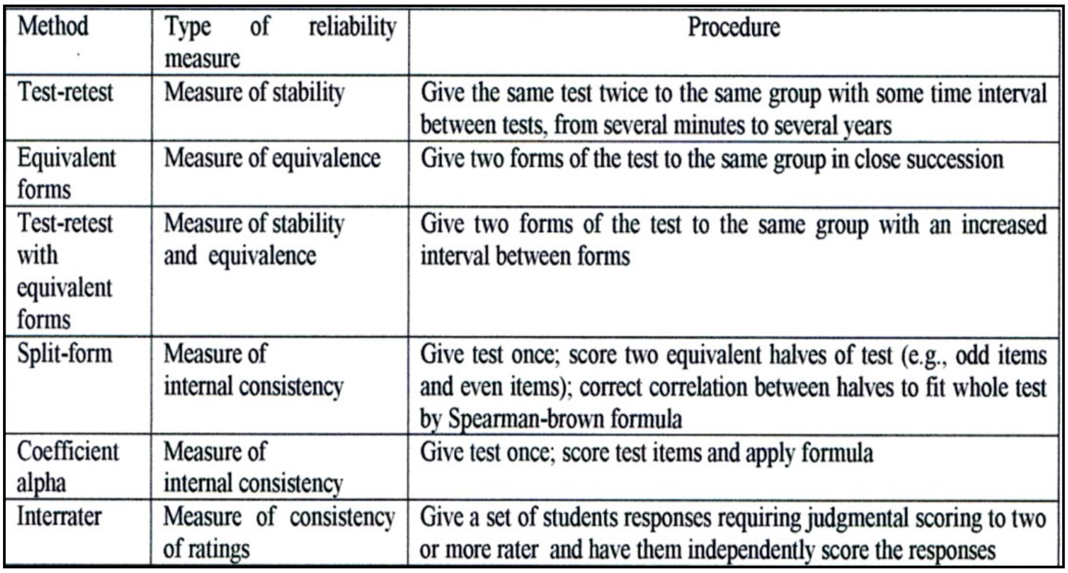
Suppose, for example, that the test takers (A) and (B) obtained the following scores respectively 06/20 and 12/20 on a given test. According to the measurement specialists, these marks can be interpreted as indicators of test takers' levels of language ability, and of other factors such as measurement errors, construct irrelevant variances, construct underrepresentation or deficiency in content relevance and coverage (Messick, 1989b, 1990, 1994). These factors can, of course, influence test takers' marks to become invalidly higher or lower than what they are supposed to be. To calculate test takers' true scores, language testers recommend us to use the following formula:

X (observed score) – E (error score) = T (true score) (Ebel & Frisbie, 1991; Miller, Linn & Gronlund, 2009)

4.5. Methods for Estimating Reliability

The estimation of reliability can be obtained when two observations of the same performance under similar testing conditions yield identical scores. However when these measures bring variable scores, measurement specialists, as indicated in Table 9, suggest substitute procedures for estimating score reliability. The common concept of these methods is that all of them involve the correlation of two sets of scores obtained either from the same assessment procedure or from equivalent forms of the same procedures (Miller, Linn, & Gronlund, 2009).

Table 9: Method for Estimating Reliability



Method	Type of reliability measure	Procedure
Test-retest	Measure of stability	Give the same test twice to the same group with some time interval between tests, from several minutes to several years
Equivalent forms	Measure of equivalence	Give two forms of the test to the same group in close succession
Test-retest with equivalent forms	Measure of stability and equivalence	Give two forms of the test to the same group with an increased interval between forms
Split-form	Measure of internal consistency	Give test once; score two equivalent halves of test (e.g., odd items and even items); correct correlation between halves to fit whole test by Spearman-brown formula
Coefficient alpha	Measure of internal consistency	Give test once; score test items and apply formula
Interrater	Measure of consistency of ratings	Give a set of students responses requiring judgmental scoring to two or more rater and have them independently score the responses

Miller, Linn, & Gronlund, 2009, p. 110

5.5.1. Test-Retest Reliability

In test retest reliability, the test is administered twice to the same examinees within a given period of time. The interval between the first and the second administrations can extend from several minutes to several years (Gronlund, 1977). The examinees' true scores are then computed by correlating the marks of the two administrations. Bachman (1990) identifies two sources of error which can affect the consistency of the test-retest method. These include ‘differential practice effect’ and ‘differential changes in language ability’. 'Differential practice effect' refers to situations when test takers perform better in the second administration because they still remember the test content, due to the brief interval between the two administrations. Concerning ‘differential changes in language ability’, since students learn at different rates, those who can retain what they have learnt for longer periods may do better in the second instance of the exam. For this reason, Gronlund (1977) suggests that the interval between the two administrations should not extend more than two weeks.

4.5.2. Equivalent Parallel Forms Reliability

Unlike the test-retest method which estimates reliability from the administration of the same test on two different occasions, equivalent parallel forms reliability estimates the equivalence of test scores across different forms of the test. These tests which are equivalent in content and construct are administered to the same group in close successions (Bachman, 2004). Once corrected, the results of the two forms will be correlated (see Table 10).

Table 10: Equivalent Parallel Forms Reliability

	1st Administration	2nd Administration
G1	Form A	Form B
G2	Form B	Form A

Source: Bachman, 2004a, p. 168

4.5.3. Split-Half Reliability

In this method, reliability is estimated from a single administration. Before correcting the test, the raters divide it into two halves. In one half, they place the odd-numbered tasks; and in the other half they include the even-numbered tasks. The two halves will be considered as two different tests and scored separately. Finally, the scores of each half will be correlated with the scores of the other half (see Table 11).

Table 11: Split Half Reliability

Sum number of odd items correct	Sum number of even items correct	September 25 Test
Items 1 3 5	Items 2 4 6	Item 1 2 3 4 5 6
Odd score = 40	Even score = 42	Total score = 82

Miller, Linn, & Gronlund, 2003, p. 113

4.6. Instruments for Maintaining Rater Consistency

Estimating reliability through the repeatability of observations is not always functional, especially in large scale assessment such as in the BAC exam. In large scale testing, measuring test takers' language ability is usually implemented by means of one observation for "in real life, no examinee is given a set of test items an infinite number of times, so the measurement error must be estimated from a single administration"(Osterlind, 2002, p. 255).The estimation of reliability in this case is concerned with the "variability that is associated with characteristics of the raters and not with the performance of examinees" (Eckes, 2008, p. 155).

Rater variability which can be defined as "the tendency on the part of raters to…provide ratings that are lower or higher than is warranted by student performances" (Engelhard, 1994, as cited in Schaefer, 2008, p. 465), can manifest itself in different ways (Lumley, 2005; McNamara, 1996; Weigle, 2002;). According to Eckes (2008) raters can differ:

> (a) in the degree to which they comply with the scoring rubric, (b) in the way they interpret criteria employed in operational scoring sessions, (c) in the degree of severity or leniency exhibited when scoring examinee performance, (d) in the understanding and use of rating scale categories, or (e) in the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks. (p. 156)

Due to the fact that much of the variability in scoring originates from “the application of different rating criteria to different samples or the inconsistent application of the rating criteria to different samples” (Bachman, 1990, p.178), language testers suggest six procedures for maintaining high intra rater and inter rater reliability. These include the use of scoring rubrics which explain in detail the criteria to be used in the rating process as well as the use of sample scripts for training raters in the pre-scoring stage. The other four criteria include independent blind double scoring, controlled reading, checks on the rating by room leaders and rater record evaluation.

4.6.1. Rating Scales

In educational measurement, we can speak of two types of scoring: objective and subjective scoring (Alderson, et al., 1995; Bachman, 1990). Objective scoring, as its name implies, requires raters to read the examinees’ scripts quickly and judge them against prearranged criteria. The candidates are “required to produce a response which can be marked either ‘correct’ or ‘incorrect’ (Alderson, et al., 1995, p. 106). Bachman (1990) explains that the correctness of these responses "is determined entirely by predetermined criteria so that no judgment is required on the part of scorers" (p.76). As a result, this type of scoring does not require too much expertise on the part of raters. Objective scoring is used to rate tasks that call for matching, multiple choice, true or false, determining odd words, picking out irregular verbs, classification of verbs according to their final ‘s’ or ‘ed’ and so on. Conversely, the evaluation of speaking or writing skills is much more complicated because it requires a rater to “make a judgment about the correctness of the response based on her (or his) subjective interpretation of the scoring criteria” (p.106). According to Henning (1987), "any rater called upon to make subjective estimates of composition quality or speaking ability in a language is liable to be inconsistent in judgement" (p. 76). Language testers emphasize that in subjective scoring "there is no feasible way to 'objectify' the scoring procedure" (Bachman, 1990, p. 76) unless we use rating scales ([AERA], [APA], & [NCME], 1999; Johnson, Penny & Gordon, 2009; Weigle, 2002).

4.6.1.1. Definition of Rating Scales

This instrument can be defined as a:

> scale for the description of language proficiency consisting of a series of constructed levels against which a language learner’s performance is judged. .…Typically such scales range from zero mastery through to an end-point representing the well-educated native speaker. The levels or bands are commonly characterized in terms of what subjects can do with the language…and their mastery of linguistic features (such as vocabulary, syntax, fluency and cohesion).(Davies et al., 1999, as cited in Fulcher, 2003, pp. 88-9 [parentheses in original]).

4.6.1.2. Types of Rating Scales

In the composition literature, rating scales can be classified into three types: primary trait, holistic and analytic scales (Davies et al., 1999; Luoma, 2004; Weigle, 2002). There are two main characteristics that distinguish these scales (see Table 12). The first is whether to use these instruments to measure a narrow aspect or a large spectrum of language ability. The second concerns whether to assign a single or multiple scores to each script (Davies et al., 1999; Fulcher, 2003; Weigle, 2002).

Table 12: Types of Rating Scales

	Specific to a particular writing task	Generalizable to a class of writing tasks
Single score	Primary trait	Holistic
Multiple score	Multiple trait	Analytic

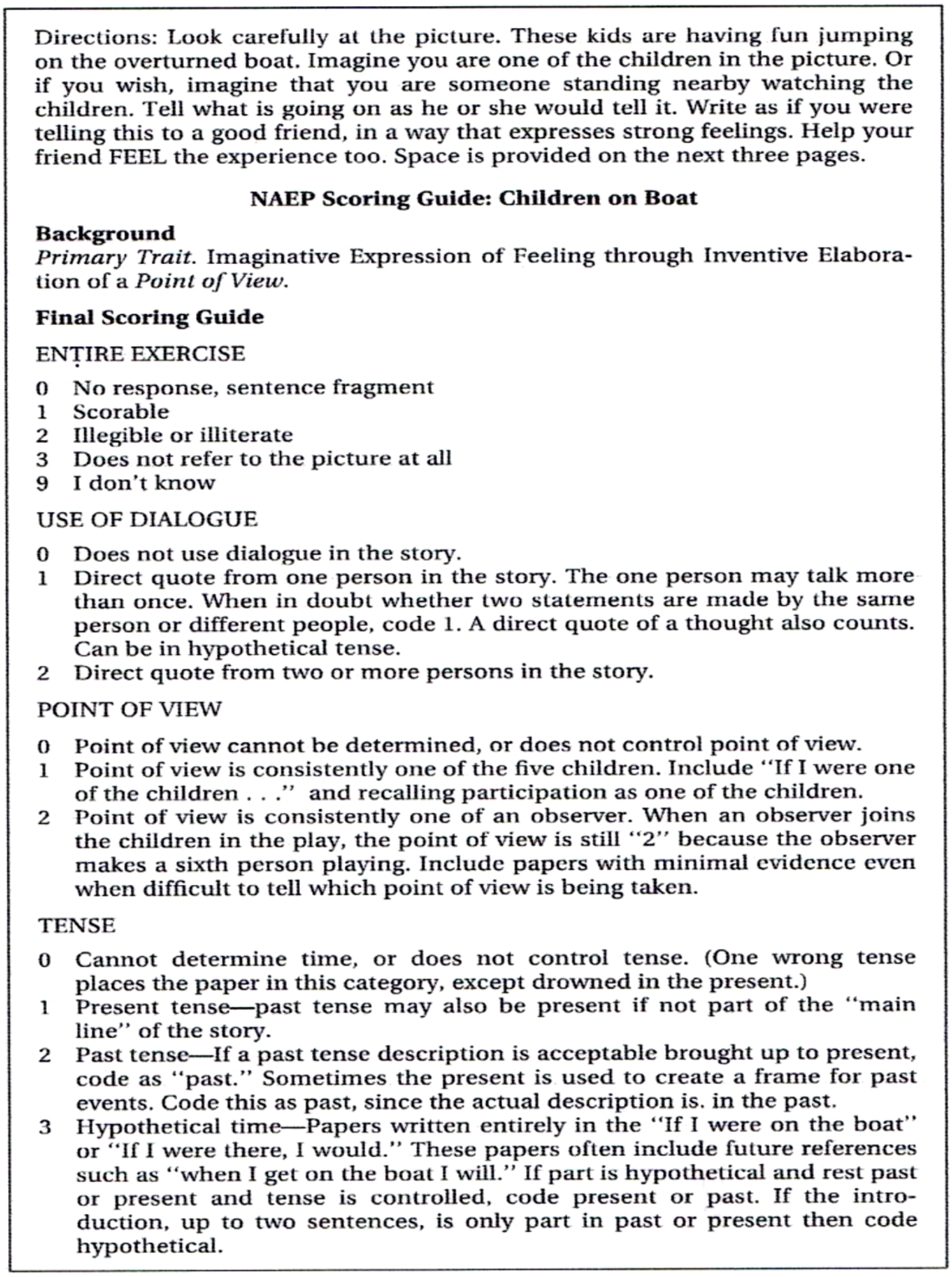
Source: Weigle (2002, p.109).

4.6.1.2.1. Primary Trait Scales

As their name imply, primary-trait scales assume that test takers' performance is made up of multiple constructs which necessitate raters "to make a single judgment about the performance on a single construct, such as 'communicative ability' [and] each descriptor in the rating scale must therefore describe a level within this construct" (Fulcher, 2012, p.378). Primary-trait scales fall into the narrowly defined type. Their main purpose is to see the extent to which learners can write or speak within a specific function of language (e.g. describing a place, salutations and greeting, asking for and granting permission, complaining and so on). As it is illustrated in Fig 25, the design of these scales is tile and labour consuming; in that a scoring rubric should be developed for each individual task. These rubrics consist of several features listed by Weigle (2002):

> (a)The writing task; (b) a statement of the primary theoretical trait (for example, persuasive essay, congratulatory letter) elicited by the task; (c) hypothesis about the expected performance on the task; (d) a statement of the relationship between the task and the primary trait; (e) a rating scale which articulates levels of performance; (f) sample scripts at each level; and (g) explanations of why each script was scored as it was (p.110).

Fig 25: Primary Trait Scoring Scale



Directions: Look carefully at the picture. These kids are having fun jumping on the overturned boat. Imagine you are one of the children in the picture. Or if you wish, imagine that you are someone standing nearby watching the children. Tell what is going on as he or she would tell it. Write as if you were telling this to a good friend, in a way that expresses strong feelings. Help your friend FEEL the experience too. Space is provided on the next three pages.

NAEP Scoring Guide: Children on Boat

Background
Primary Trait. Imaginative Expression of Feeling through Inventive Elaboration of a *Point of View.*

Final Scoring Guide

ENTIRE EXERCISE

0 No response, sentence fragment
1 Scorable
2 Illegible or illiterate
3 Does not refer to the picture at all
9 I don't know

USE OF DIALOGUE

0 Does not use dialogue in the story.
1 Direct quote from one person in the story. The one person may talk more than once. When in doubt whether two statements are made by the same person or different people, code 1. A direct quote of a thought also counts. Can be in hypothetical tense.
2 Direct quote from two or more persons in the story.

POINT OF VIEW

0 Point of view cannot be determined, or does not control point of view.
1 Point of view is consistently one of the five children. Include "If I were one of the children . . ." and recalling participation as one of the children.
2 Point of view is consistently one of an observer. When an observer joins the children in the play, the point of view is still "2" because the observer makes a sixth person playing. Include papers with minimal evidence even when difficult to tell which point of view is being taken.

TENSE

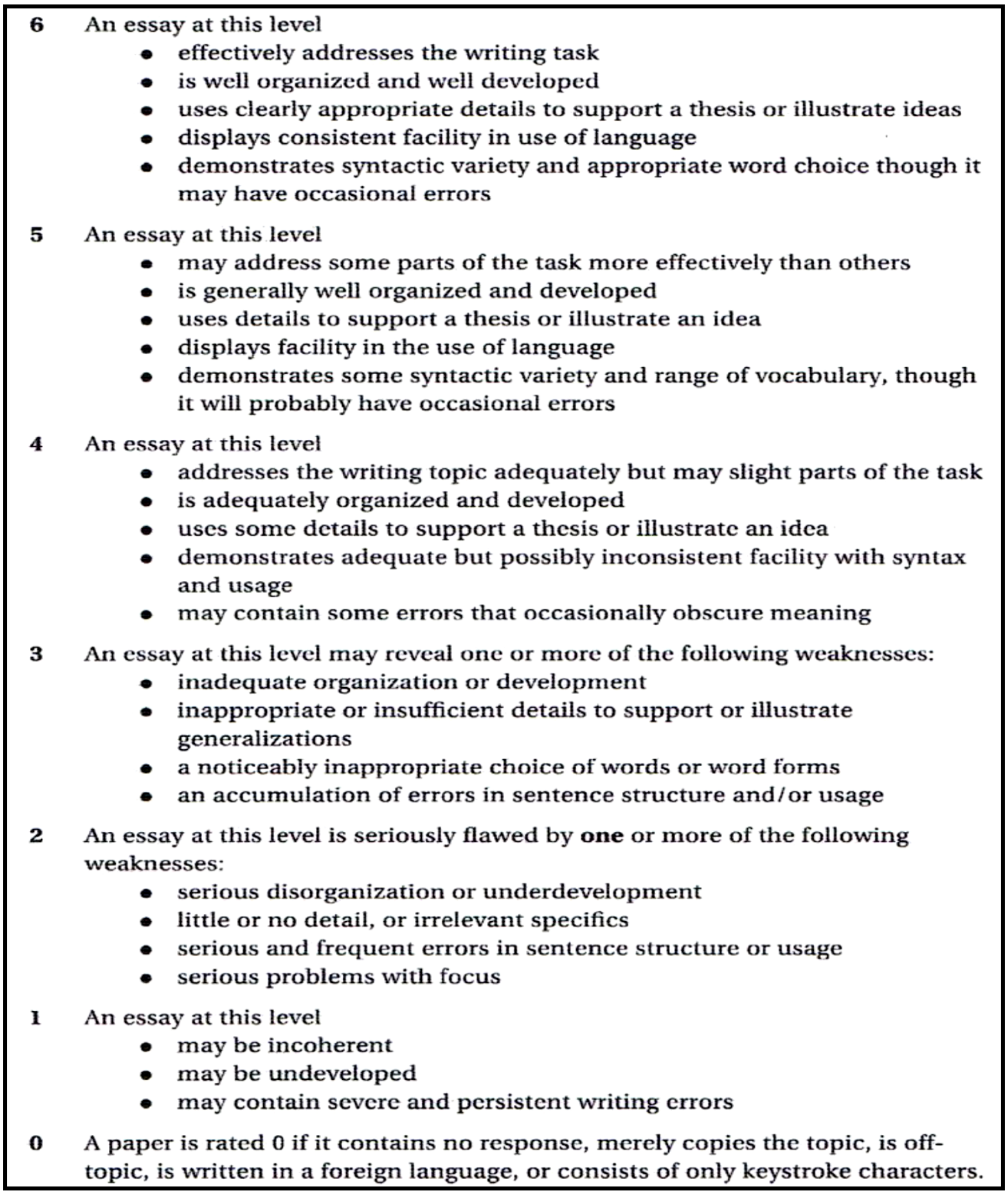
0 Cannot determine time, or does not control tense. (One wrong tense places the paper in this category, except drowned in the present.)
1 Present tense—past tense may also be present if not part of the "main line" of the story.
2 Past tense—If a past tense description is acceptable brought up to present, code as "past." Sometimes the present is used to create a frame for past events. Code this as past, since the actual description is. in the past.
3 Hypothetical time—Papers written entirely in the "If I were on the boat" or "If I were there, I would." These papers often include future references such as "when I get on the boat I will." If part is hypothetical and rest past or present and tense is controlled, code present or past. If the introduction, up to two sentences, is only part in past or present then code hypothetical.

Weigle, 2002, p. 111

4.6.1.2.2. Holistic Scales

Holistic scoring refers to the assignment "of a single score to a script based on the overall impression…each script is read quickly and then judged against a rating scale, or a scoring rubric that outlines the scoring criteria" (Weigle, 2002, p.112). Holistic scoring should not be confounded with ‘general impression scoring’. The main difference between them lies in the availability of a rating scale. In holistic scoring, assessors are required to judge examinees' language performance against a rating scale or scoring rubric; however in ‘general impression scoring’, raters read test takers' responses and assign a single score building their judgement upon their own evaluation which specifies no reliable or explicit criteria (Weigle, 2002). The most famous holistic rating scale is the one developed for the ‘TOEFL’ (see Fig 26).

Fig 26: TOFEL Holistic Rating Scale



6 An essay at this level

- effectively addresses the writing task
- is well organized and well developed
- uses clearly appropriate details to support a thesis or illustrate ideas
- displays consistent facility in use of language
- demonstrates syntactic variety and appropriate word choice though it may have occasional errors

5 An essay at this level

- may address some parts of the task more effectively than others
- is generally well organized and developed
- uses details to support a thesis or illustrate an idea
- displays facility in the use of language
- demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors

4 An essay at this level

- addresses the writing topic adequately but may slight parts of the task
- is adequately organized and developed
- uses some details to support a thesis or illustrate an idea
- demonstrates adequate but possibly inconsistent facility with syntax and usage
- may contain some errors that occasionally obscure meaning

3 An essay at this level may reveal one or more of the following weaknesses:

- inadequate organization or development
- inappropriate or insufficient details to support or illustrate generalizations
- a noticeably inappropriate choice of words or word forms
- an accumulation of errors in sentence structure and/or usage

2 An essay at this level is seriously flawed by **one** or more of the following weaknesses:

- serious disorganization or underdevelopment
- little or no detail, or irrelevant specifics
- serious and frequent errors in sentence structure or usage
- serious problems with focus

1 An essay at this level

- may be incoherent
- may be undeveloped
- may contain severe and persistent writing errors

0 A paper is rated 0 if it contains no response, merely copies the topic, is off-topic, is written in a foreign language, or consists of only keystroke characters.

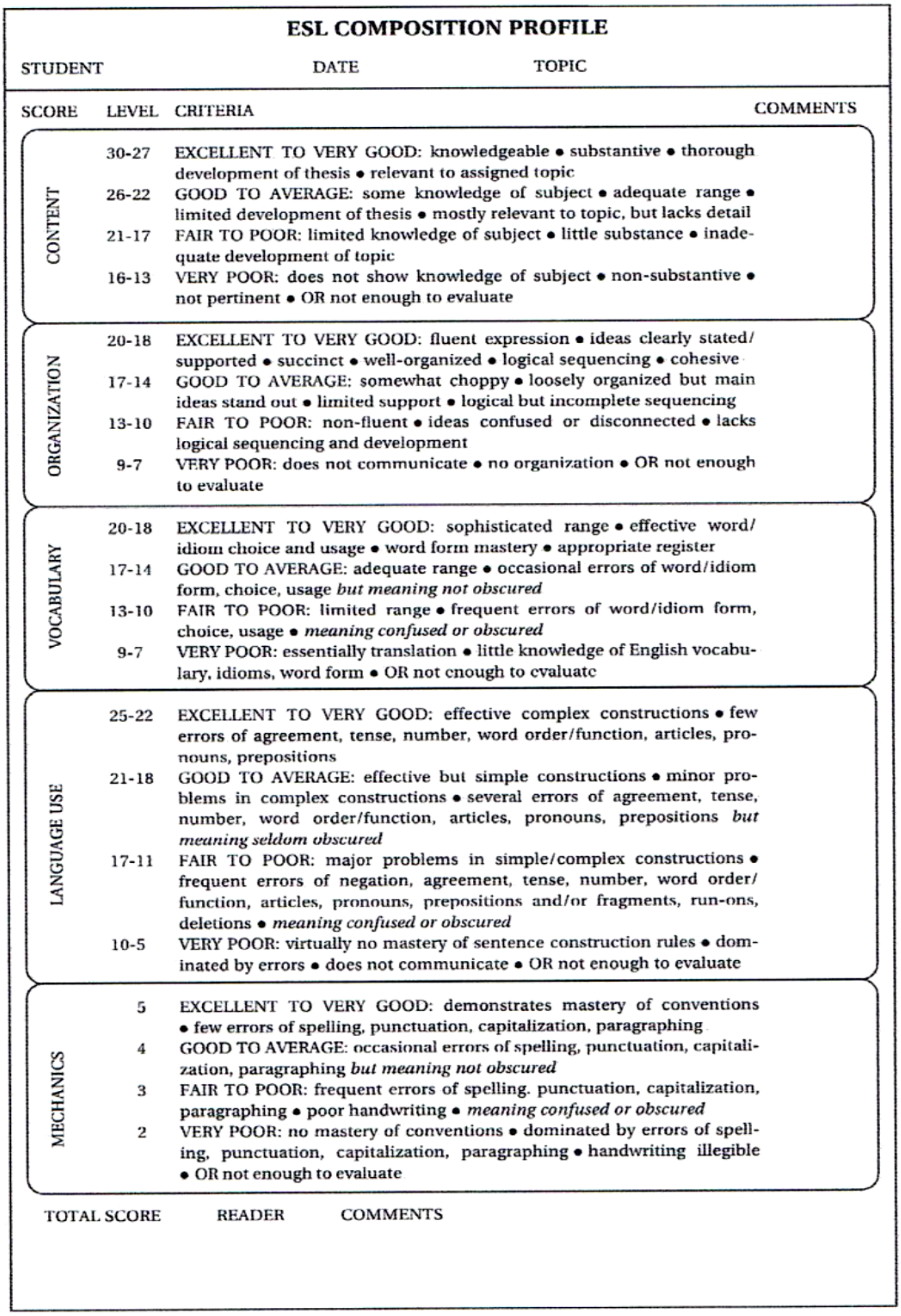
Source: Weigle, 2002, p. 113.

4.6.1.2.3. Analytic Scoring

In analytic scoring, raters read the scripts and assess them on different aspects such as grammar, cohesion, coherence and mechanics and so on. Unlike in holistic scoring when judges assign a single score, analytic scoring requires them to assign various scores according to the examinees' level of success or deficiency in a given language component. According to Weigle (2002), the most famous analytic scale is the one developed by Jacobs et., al (1981) which judges test takers' written performance on five aspects of writing: 'content, organization, vocabulary, language use, and mechanics' (see Fig 27).

In the same way, another analytic scale for measuring examinees' written and oral performance has been developed by Cyril Weir in 1998 (Weigle, 2002; Weir, 1998). This scale , as we see in Fig 28, evaluates test takers' responses on seven aspects: relevance and adequacy of content, compositional organization, cohesion, and accuracy of vocabulary for purpose, grammar, mechanical accuracy I: pronunciation and mechanical accuracy II: spelling (Weigle, 2002; Weir 1998).

Fig 27: Jacobs et al's 1981, Analytic Rating Scale



ESL COMPOSITION PROFILE

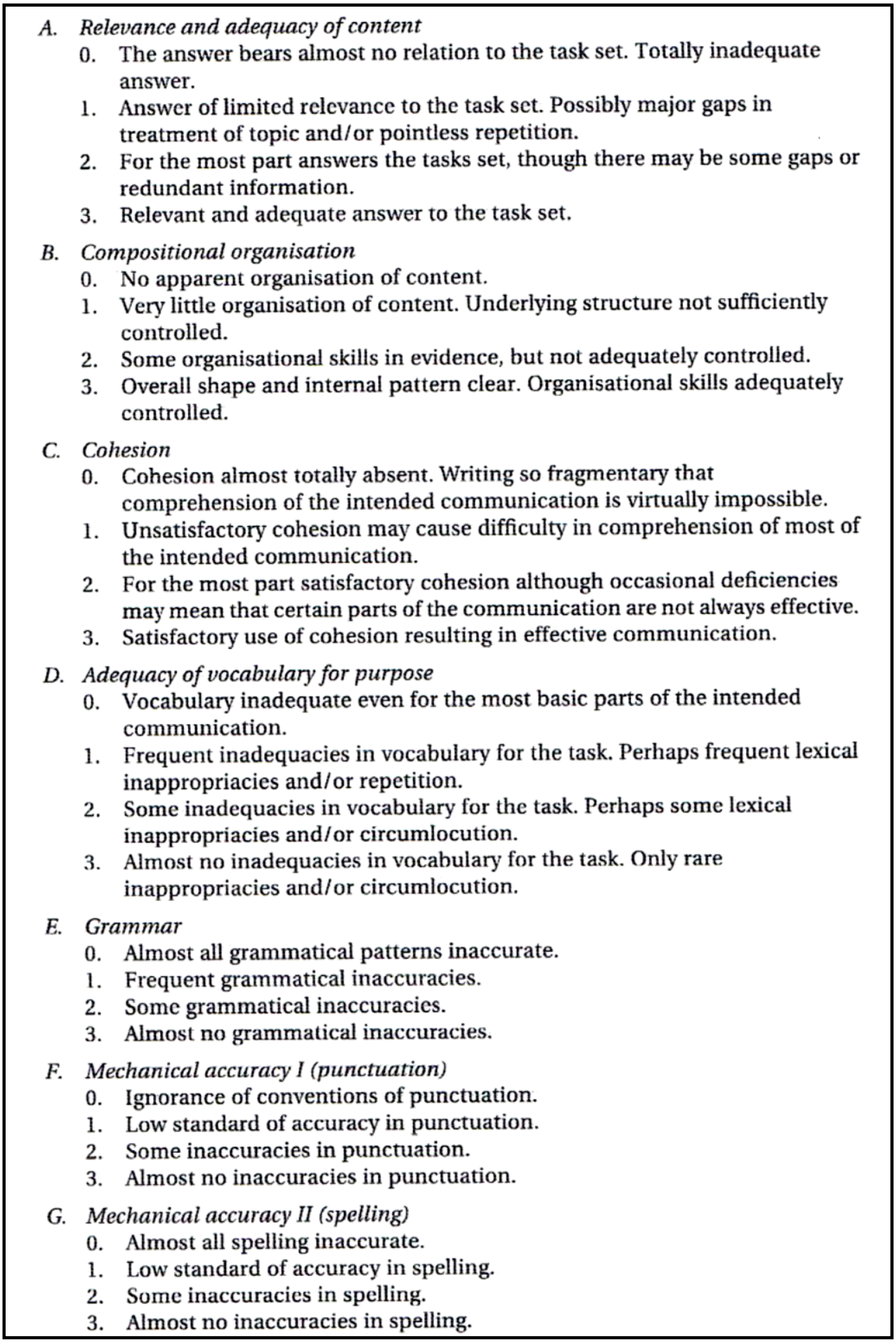
STUDENT DATE TOPIC

SCORE	LEVEL	CRITERIA	COMMENTS
CONTENT	30-27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic	
	26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail	
	21-17	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic	
	16-13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate	
ORGANIZATION	20-18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/ supported • succinct • well-organized • logical sequencing • cohesive	
	17-14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing	
	13-10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development	
	9-7	VERY POOR: does not communicate • no organization • OR not enough to evaluate	
VOCABULARY	20-18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/ idiom choice and usage • word form mastery • appropriate register	
	17-14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured*	
	13-10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • *meaning confused or obscured*	
	9-7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate	
LANGUAGE USE	25-22	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions	
	21-18	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured*	
	17-11	FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/ function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • *meaning confused or obscured*	
	10-5	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate	
MECHANICS	5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing	
	4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing *but meaning not obscured*	
	3	FAIR TO POOR: frequent errors of spelling. punctuation, capitalization, paragraphing • poor handwriting • *meaning confused or obscured*	
	2	VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate	

TOTAL SCORE READER COMMENTS

Source: Weigle, 2002, p.116

Fig 28: Weir's 1998 Analytic Scale



A. *Relevance and adequacy of content*

0. The answer bears almost no relation to the task set. Totally inadequate answer.
1. Answer of limited relevance to the task set. Possibly major gaps in treatment of topic and/or pointless repetition.
2. For the most part answers the tasks set, though there may be some gaps or redundant information.
3. Relevant and adequate answer to the task set.

B. *Compositional organisation*

0. No apparent organisation of content.
1. Very little organisation of content. Underlying structure not sufficiently controlled.
2. Some organisational skills in evidence, but not adequately controlled.
3. Overall shape and internal pattern clear. Organisational skills adequately controlled.

C. *Cohesion*

0. Cohesion almost totally absent. Writing so fragmentary that comprehension of the intended communication is virtually impossible.
1. Unsatisfactory cohesion may cause difficulty in comprehension of most of the intended communication.
2. For the most part satisfactory cohesion although occasional deficiencies may mean that certain parts of the communication are not always effective.
3. Satisfactory use of cohesion resulting in effective communication.

D. *Adequacy of vocabulary for purpose*

0. Vocabulary inadequate even for the most basic parts of the intended communication.
1. Frequent inadequacies in vocabulary for the task. Perhaps frequent lexical inappropriacies and/or repetition.
2. Some inadequacies in vocabulary for the task. Perhaps some lexical inappropriacies and/or circumlocution.
3. Almost no inadequacies in vocabulary for the task. Only rare inappropriacies and/or circumlocution.

E. *Grammar*

0. Almost all grammatical patterns inaccurate.
1. Frequent grammatical inaccuracies.
2. Some grammatical inaccuracies.
3. Almost no grammatical inaccuracies.

F. *Mechanical accuracy I (punctuation)*

0. Ignorance of conventions of punctuation.
1. Low standard of accuracy in punctuation.
2. Some inaccuracies in punctuation.
3. Almost no inaccuracies in punctuation.

G. *Mechanical accuracy II (spelling)*

0. Almost all spelling inaccurate.
1. Low standard of accuracy in spelling.
2. Some inaccuracies in spelling.
3. Almost no inaccuracies in spelling.

Source: Weigle, 2000, p. 117

4.6.2. Rater Training

The process of scoring tests is not limited to expert raters. In the BAC exam, for instance, novice raters are also invited to participate in this process (ONEC, 2012, 2013). In the field of assessment, it is widely recognised that these two types of judges differ in their overall severity and leniency (Bachman, 1990; Bachman & Palmer, 1996; McNamara, 1996). Language testers stress that “reliable ability measures are unlikely to be achieved from untrained raters” (Weigle, 1994, as cited in McNamara & Rover, 1996, p. 124). Introducing raters to the assessment without any type of training is considered problematic in that "if the marking of a test is not...reliable then all of the other work undertaken earlier to construct a ‘quality’ instrument will have been a waste of time” (Alderson, et al., 1995, p. 105). Training tends to achieve two main objectives. One the one hand, it contributes to bringing raters into agreement, or at least into adjacent agreement. On the other hand, it reinforces stability and self-consistency within individual raters (Hamp-Lyons, 2007; Knoch, 1996; Lumley & McNamara, 1995).

4.6.2.1. Standardising Raters' Scoring

Monitoring raters’ judgements can take place at three phases: before, during and after live scoring. The main purpose of the first stage is to ensure a uniform interpretation of the scoring guide. At this stage, the chief examiners introduce the scoring guide, the rating scale and the other marking procedures as training instruments. In order to put the theory into practice, sample scripts will be chosen for the pre-scoring session. The scripts will be divided into two batches: consensus scripts and problematic scripts. Problematic scripts fall, according to Weigle (2002) into three categories: off-task scripts, memorized scripts and incomplete scripts. The first category includes "scripts that are complete but do not address the intended task” (p. 132). The second type refers to the scripts “that have clearly been written from memory rather than in response to the prompts”(p.132). The third category includes "scripts in which the writer has demonstrated an understanding of the important features of the task but was unable to complete the task in the allotted time" (p.132). Each type of the problematic scripts will be blindly double scored; and in case the pre-rating produces adjacent or discrepant scores, a method for adjusting this variability will be implemented.

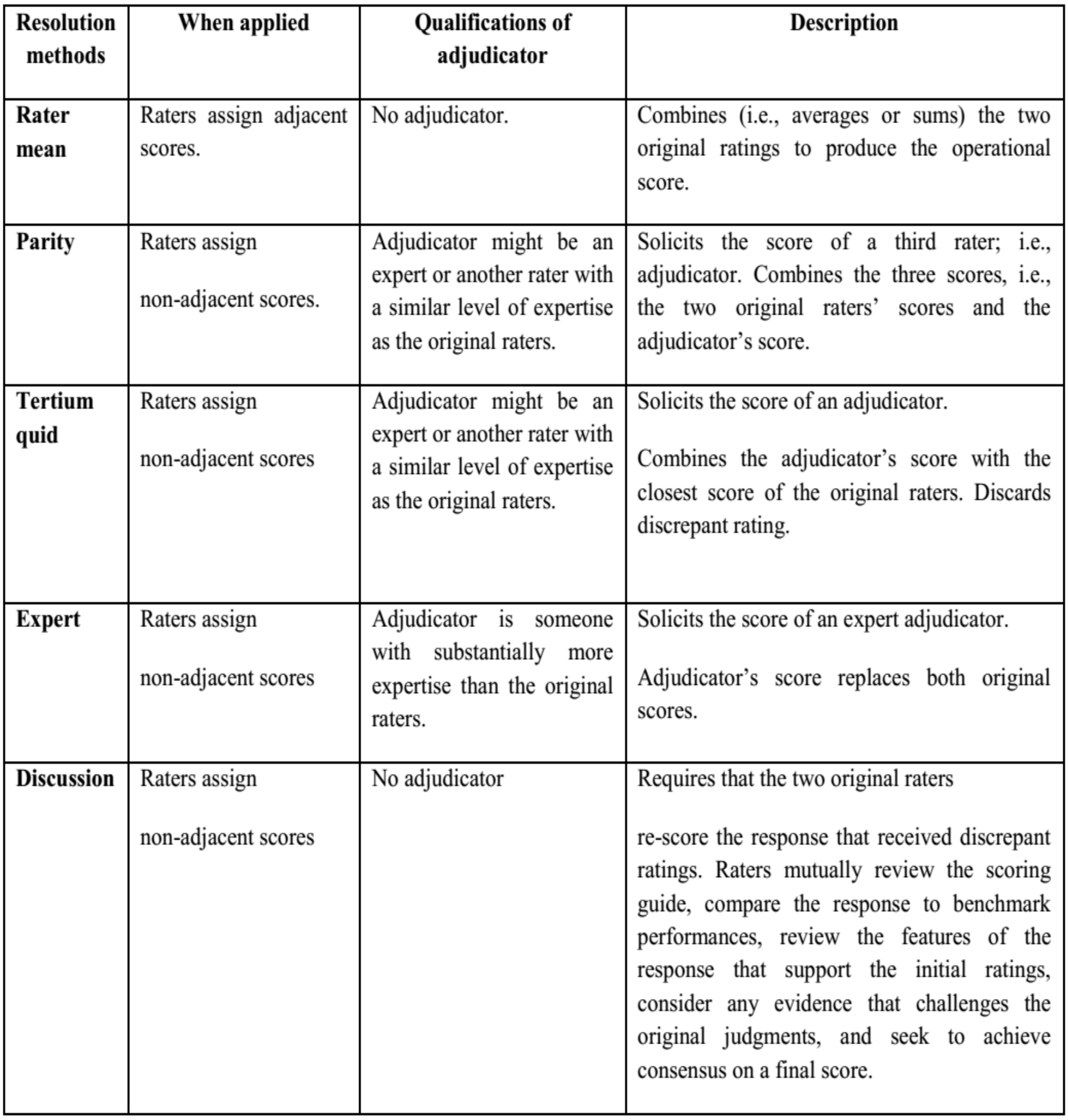
4.6.3. Reinforcing Reliability within Raters (Intra-Rater Reliability)

In any rating, consistency within raters tends to be less accurate for a number of reasons. Some of these factors are internal into the raters themselves, while others are external to them. For example, when the process of scoring extends for long periods of time, the sequencing of corrections or fatigue can affect the precision within these judges. This can also occur in cases where raters are not provided with rating scales or when they find it difficult to interpret the scoring criteria because of the lack of training. Moreover, there are raters who are influenced by superficial features such as handwriting, or the organization of responses. So, in order to ensure the consistency of scoring within a single rater "we need to obtain at least two independent ratings from this rater for each individual language sample" (Bachman, 1990, p. 179).

4.6.4. Reinforcing Reliability between Raters (Inter-Rater Reliability)

Several methods for resolving rater discrepancies have been proposed in the literature of language testing (Johnson, Penny & Gordon, 2009, 2010). These include rater mean, parity, expert, tertium quid and discussion methods (see Table 13). The first method is implemented when two ratings of the same script fall into tolerated variability, or adjacent agreement. In case of discrepant scores, one of the other four methods will be implemented. The extent to which we consider scores adjacent or discrepant depends on the directions of test developers (McNamara, 1996).

Table 13: Major Models of Score Discrepancy Resolution



Resolution methods	When applied	Qualifications of adjudicator	Description
Rater mean	Raters assign adjacent scores.	No adjudicator.	Combines (i.e., averages or sums) the two original ratings to produce the operational score.
Parity	Raters assign non-adjacent scores.	Adjudicator might be an expert or another rater with a similar level of expertise as the original raters.	Solicits the score of a third rater; i.e., adjudicator. Combines the three scores, i.e., the two original raters' scores and the adjudicator's score.
Tertium quid	Raters assign non-adjacent scores	Adjudicator might be an expert or another rater with a similar level of expertise as the original raters.	Solicits the score of an adjudicator. Combines the adjudicator's score with the closest score of the original raters. Discards discrepant rating.
Expert	Raters assign non-adjacent scores	Adjudicator is someone with substantially more expertise than the original raters.	Solicits the score of an expert adjudicator. Adjudicator's score replaces both original scores.
Discussion	Raters assign non-adjacent scores	No adjudicator	Requires that the two original raters re-score the response that received discrepant ratings. Raters mutually review the scoring guide, compare the response to benchmark performances, review the features of the response that support the initial ratings, consider any evidence that challenges the original judgments, and seek to achieve consensus on a final score.

Source: Johnson, Penny & Gordon, 2009, p.242

4.6.4.1. Rater Mean Method

As it has been mentioned previously, rater mean is used when raters assign adjacent scores to the same script. The operational score is computed by averaging the marks of the two original raters. In the BAC English test, rater mean method is usually implemented by the clerical staff after combining and averaging the scores resulting from the first and the second phases of rating.

4.6.4.2. Parity Method

This method involves the incorporation of adjudication or moderation techniques. In case of disagreement between the original raters, an adjudicator (third rater) is involved to carry out a blind review of the disputed paper. In parity method, the final score is computed by combining, then averaging the three marks. What is worth mentioning here is that adjudicators are raters of more expertise than the original raters.

4.6.4.3. Tertium Quid Method

This method derives its name from “the medieval practice in which a deadlock in a debate is resolved by eliciting a decision from a third party in favor of one of the disputants” (Johnson, Penny, & Cordon, 2009, p.243). The incorporation of adjudication in this method is different from the one adopted in parity method. One form requires the adjudicator to carry out a blind review of the disputed scripts. The operational score is produced by averaging the adjudicator’s mark with the closest score. In this case, the discrepant score is eliminated. The second form is implemented when the adjudicator’s mark happens to be in a position in-between the two original scores. This involves "averaging the original scores, doubling the third score; or combining the third score with the higher of the two original scores" (Johnson, Penny, & Johnson, 2000, as cited in Penny & Johnson 2001 p. 222). The two other forms do not call for third correction. In one form, the mediator reviews the previously rated scripts and decides which of the two original ratings is to be retained. In the other form, the third judge reviews the corrected scripts and the scoring guide, and then moves one of the original scores up or down.

4.6.4.4. Expert Judgement Method

As its name implies, this model underlines the important role of the expert rater who is supposed to be "someone with substantial more expertise in scoring" (Penny & Johnson, 2001 p.224). Experts' characteristics include "experience in the scoring of constructed-response items, advanced training in the subject area being scored, familiarity with a wide range of student capabilities, the respect of his or her colleagues, and the ability to communicate clearly" (Wolcott, 1998, as cited in Johnson, Penny, Gordon, Shumate, & Fisher, 2005, p. 5). What is worth mentioning here is that adjudication is not incorporated to moderate the discrepant scores. On the contrary, the scores of expert raters eliminate and replace the two original marks which "implies that the judgment of the expert provides a more accurate estimate of the examinee’s proficiency than do the combined judgments of the original raters" (Johnson, Penny & Gordon, 2009, p.244).

4.6.4.5. Discussion Model

Discussion is another method for resolving rater variability. This method requires the identification of the raters who assigned the discrepant scores. These raters are invited to meet and mutually reexamine and review the scripts, the scoring guide, and the rating scale. Then, they "review the features of the performance that support the initial ratings, consider any evidence that challenges their original judgments, and seek to achieve consensus on a final score" (Penny & Johnson, 2011, p. 224).

Conclusion

In conclusion, reliability can be estimated in terms of stability across repeated measures and of scoring consistency in single administration. Stability and consistency refer to the dependability of test scores over different occasions; over parallel tests; over different parts of the same test; and within and across different raters. The importance of reliability as second quality for evaluating language tests lies in the fact that it provides confidence to test scores which can, in their turn, be generalized to target language contexts beyond the test itself. This means that if our ratings produce unreliable measurement, the interpretations that we provide to test scores will be inconsistent and inappropriate (Gronlund, 1977; Miller, Linn, & Gronlund, 2009). Additionally, reliability in language testing attempts to distinguish between two main features: what is the extent to which test scores are affected by differences in test takers' levels of language ability? And how much variance is related to factors that are not related to the ability being measured? The response to these questions helps us design dependable, consistent and reliable measures.

Chapter Five

Investigating Test Validity

Chapter Five

Investigating Test Validity

Introduction

The conceptualization of validity has been a topic of debate amongst language testers and educational measurement specialists (Cronbach & Meel, 1955; Gronlund; 1987; McNamara & Rover, 2006; Messick, 1989, 1994; Miller, Linn & Gronlund, 2009). The traditional school, for instance, conceives this concept as a property that is relevant to the test itself and nothing else. In other words, it considers a test to be valid to the extent to which it measures what it purports to measure. This trend divides this quality into three main types: content, construct and criterion validities; and tests can be validated with reference to each one these types (Ruch, 1924). Conversely, the modern trend considers validity as a unitary concept comprising several features such as content, criterion, construct, substantive, structural, generalizability, external and consequential aspects which function in an integrated unifying validity framework. According to this trend , what needs to validated is not the test itself nor its scores, but the interpretations, uses and the consequences emerging from these scores (([AERA], [APA], & [NCME], 1999; Messick, 1989, 1995).

5.1. Validity in the Perspective of the Traditional Trend

Most of the traditional definitions to validity lend themselves to Ruch (1924). Ruch (as cited in Fultcher, 2010) defines validity as “the degree to which a test or examination measures what it purports to measure”(p.19). On his part, Lado (1961, as cited in Weir, 2005) inquires ‘‘does a test measure what it is supposed to measure? If it does, it is valid” (p.12). In the same way, Henning (1987) points out that this concept “refers to the appropriateness of a given test or one of its component parts as a measure of what it is purported to measure" (p. 89). According to him, "a test is said to be valid to the extent that it measures what it is supposed to measure" (p. 89). The point of view of this trend is summarized by Heaton (1988) who regards "the validity of a test [as] the extent to which it measures what it is supposed to measure and nothing else" (p. 159).

5.1.2. Types of Evidence in the Traditional Paradigm

As we have mentioned previously, the intent of validity in the traditional paradigm is to validate tests with respect to the purposes for which they have been designed. Consequently, this trend divides validity into three major distinct types: content validity (and/ or face validity), criterion-oriented validity (predictive and concurrent), and construct validity (Davis & Elder, 2005; Fulcher & Davidson, 2007, 2008; McNamara, 2006). Each type is, as illustrated in Table 14, "related to the kind of evidence that would count towards demonstrating that a test was valid" (Fulcher & Davidson, 2007, p. 4).

Table 14: Basic Types of Validity in the Traditional Paradigm

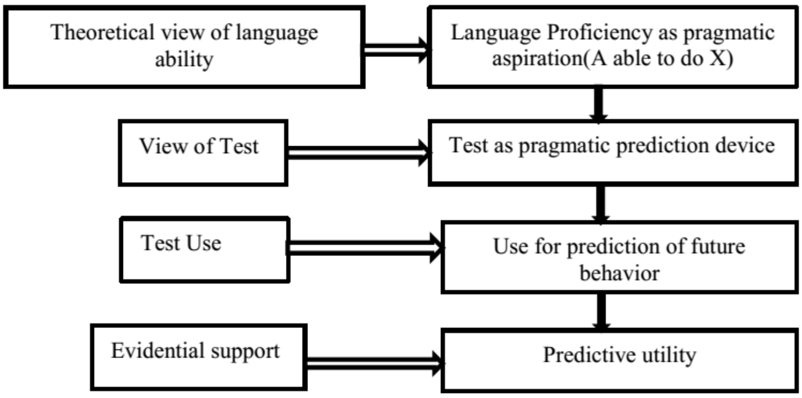
Basic Types of traditional Validity	
TYPE	Question to be Answered
Content validity	How adequately does the test content sample the larger universe of situations it represents?
Criterion-related validities	How well does test performance predict future performance (predictive validity) or estimate present standing (concurrent validity) on some other valued measure called a criterion?
Construct validity	How well can test performance be explained in terms of psychological attributes?

Source: Gronlund, 1977, p. 131

5.1.2.1. Criterion-oriented Validity

Criterion-related validity can be "evaluated by comparing the test scores with one or more external variables (called criteria) considered to provide a direct measure of the characteristic or behavior in question" (Messick, 1990, p. 7 [parentheses in original]). For example, a good score obtained by a teacher trainee or an aviation apprentice can be associated with a highly qualified teacher or pilot (Alderson, 1990). Criterion related validity is usually used to describe two subtypes of validity: predictive and concurrent validities. The former which is established when the test and the criterion are administered at about the same time, "indicates the extent to which the test scores estimate an individual's present standing on the criterion" (Messick, 1990, p. 7). However the latter, as shown in fig 29, concerns the extent to which test scores can predict the examinees' future performance or standing on an occupational position (Weir, 2009).

Fig 29: Predictive Utility


Theoretical view of language ability
Language Proficiency as pragmatic aspiration(A able to do X)
View of Test
Test as pragmatic prediction device
Test Use
Use for prediction of future behavior
Evidential support
Predictive utility

Source: Bachman, 1990, p. 254

5.1.2.2. Content Validity

Before providing a definition to content validity, let us first specify what we mean with test content.The Standards of Educational and Psychological Testing (AERA, APA & NCME, 1999) define the test content as “the themes, wording, and items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring” (p. 11). Concerning content validity, it can be defined as “any attempt to show that the content of the test is a representative sample from the domain that is to be tested” (Fulcher & Davidson, 2007, p.5). Language testers identify two aspects of content validity: content relevance and content representation or coverage (AERA, APA & NCME,1999; Bachman, 2005; Henning, 1987; Mesick, 1989). Content relevance requires the content of the test to be relevant to the construct which is intended to be measured. The other aspect 'content coverage' or 'ecological sampling' as Brunswick (1956) calls it, entails “that all important parts of the construct domain are covered, which is usually described as selecting tasks that sample domain processes in terms of their functional importance (Messick, 1995, p. 746). In case the syllabus happens to be homogeneous, the best technique to ascertain content coverage is random sampling; but if the syllabus is heterogeneous, content representation can be implemented by means of stratified random sampling (Bachman, 1990).

5.1.2.2.1.Face Validity

There is a consensus amongst educational measurement specialists that face validity refers to the extent to which a test appears to measure what it claims to measure based on the intuitive judgment of someone (usually naïve, lay-person, or untrained observer) who lacks the expertise to scrutinize evidence of validity (Alderson et al.,1995; Henning, 1987; Nunnally & Bernstein, 1994; Richards & Schmidt, 2010; Urbina, 2004).

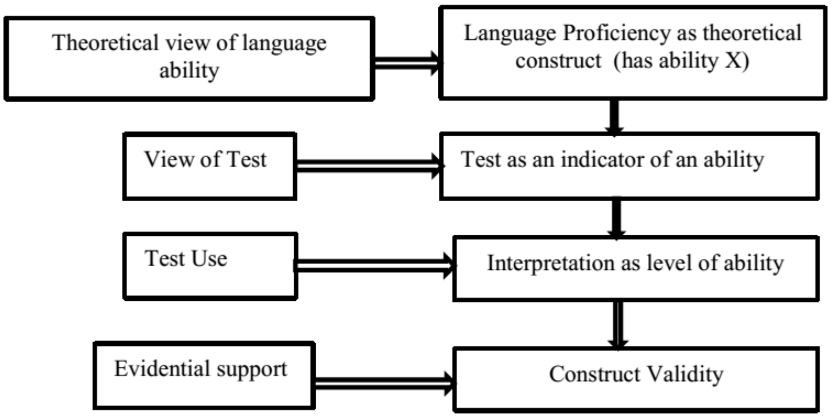
Yet, their divergence is on whether to consider the subjective and superficial impression of what a test claims to test as a part of validly. Despite the fact that Hughes (1989) recognizes that this type of validity is "hardly a scientific concept" (p. 27), he highlights its role in engaging test takers' language knowledge to interact with the test input and underlines that "a test which does not have face validity may not be accepted by candidates [since this] may mean that they do not perform on it in a way that truly reflects their ability” (p. 27). In the same way, though Urbina (2004) thinks of face validity to refer "to the superficial appearance of what a test measures from the perspective of a test taker or any other naive observer" (p. 169), she stresses that test developers need to design tests whose content and skills seem to measure what they purport to measure. This is because "if the content of a test appears to be inappropriate or irrelevant to test takers, their willingness to cooperate with the testing process is likely to be undermined" (p. 169).

Opponents of face validity do not see its efficacy in test validation (Bachman, 1990, 2005, 2013; Cronbach, 1984, 1988; Messick,1989, 1994, 1995). Cronbach (1984, as cited in Bachman, 1990) warns against "adopting a test just because it appears reasonable" (p.286) to the lay man and considers this to be a 'bad practice'. In the same way, the Standards for Educational and Psychological Testing (1974, as cited in Bachman, 1990) maintain that that this “so-called “face” validity, the mere appearance of validity, is not an acceptable basis for interpretive inferences from test scores’(pp. 284-285). Presumably, the “final interment of the term", according to Bachman (1990) was "marked by its total absence from the most recent (1985) edition of the ‘Standards’"(p 285) of Educational and Psychological Testing.

5.1.2.3. Construct Validity

Construct validity investigates the extent to which a test can "be interpreted as a measure of some attribute or quality which is not "operationally defined" (Cronbach & Meel, 1955, p. 283). This definition is, of course, consistent with the trait-based approach which limits the scope of constructs to psychological traits (see Fig 30). However according to the task-based approach conceptualization, this scope can be extended to encompass not only what people have in terms of language knowledge, but to what they can do with language in communicative target situations beyond the test (Bachman, 2007; Messick, 1996; Richards & Schmidt, 2010; Stuart-Hamilton 2007; Tavakoli, 2012). The views of the trait-based and task or context-based approaches of construct validity emphasize that tests should address "both the cognitive and linguistic abilities involved in activities in the language use domain of interest, as well as the context in which these abilities are performed" (Weir, 2005, p. 14).

Fig 30: Trait-based Approaches of Construct Validity


Theoretical view of language ability
Language Proficiency as theoretical construct (has ability X)
View of Test
Test as an indicator of an ability
Test Use
Interpretation as level of ability
Evidential support
Construct Validity

Source: Bachman, 1990, p. 254

In sum, validity in the traditional trend refers to the extent to which a test measures what is claims to measure. This trend breaks validity into three distinct types: content, criterion and construct validities. Content validity investigates the extent to which a test content samples skills, task, themes or items from the construct domain. Criterion validity, which compares the degree of correspondence between the scores obtained on a given test and a criterion score, is subdivided into two classes: predictive and concurrent validities. The former examines how well test scores can predict test takers' future performance; and the latter associates test results with a pretesting instrument that has previously proven to be reliable and valid. The third type 'construct validity' examines the degree to which a test measures a psychological trait.

5.2. Validity as a Unitary Concept.

5.2.1. Historical Overview

The principle of construct validity as an overall process for test score interpretations lends itself to the American Psychologists Association's (APA) Ethical standards of 1953 and to the seminal article of Cronbach and Meehl published in 1955 (APA, 1985: McNamara, 2006; Messick, 1989, 1998). Between 1950 and 1954, the 'APA' set up a committee for the purpose of specifying “what qualities should be investigated before a test is published” (Cronbach & Meehl, 1955, p.283). According to the authors, the main ‘innovation’ of the committee was the coining of the term construct validity (p.283). “In the thirty years since”, as Bachman (1990) points out “construct validity has come to be recognized by the measurement profession as central to the appropriate interpretation of test scores, and provides the basis for the view of validity as a unitary concept” (p. 255). In this context, Messick (1980) argues that “construct validity is indeed the unifying concept that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships” (p.1015). In the mid-eighties, the Standards for Educational and Psychological Testing (APA, 1985) considered validity to be referring "to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores"(p.9). According to Bachman (1990), this means that "the measurement profession has clearly linked validity to the inferences that are made on the basis of test scores" (p.244). In 1989, Messick introduced his new model of construct validity which, in addition to the interpretation and use of test scores, he incorporated factors related to test consequences**.** During the early nineties, the concept of construct validity as an overall evaluative concept continued to gain grounds at the expense of the conventional view. By the end of the decade, the validity pendulum fell completely in the advantage of the unitary trend (Bachman, 2005, 2007, 2013; Kane, 2013; McNamara & Rover, 2006).

5.2.2. Definition of Construct Validity

Several definitions to construct validity as a unitary concept have been proposed in the literature (APA, 1953, 1966; 1974; AERA, APA & NCME, 1999; Cronbach, 1970, 1988; Cronbach & Meehl, 1955; Messick, 1996). Messick (1989 as cited in Messick, 1995), for example, considers construct validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment" (p.741). The author does not restrain his definition to the interpretive purposes of test scores but extends it "to inferences based on any means of observing or documenting consistent behaviors or attributes" (Messick, 1990, p.1). On their part, AERA, APA, and NCME (1999) regard validity as "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests" (p. 11). According to the 'Measurement Profession', "the proposed interpretation refers to the constructs or concepts the test is intended to measure" (p. 11). The unitary trend considers the traditional division of validity into distinct types as 'fragmented and incomplete' (Messick, 1989, 1990, 1994, 1995) because this on the one hand does not account for the way in which the accumulated evidence supports the score interpretation; nor does it describe the effect of the intended and unintended consequences on test takers on the other (AERA, APA, & NCME, 1999; Chappelle, 2012; Chapelle, Enright, & Jamieson, 2008, 2010; Kane, 2012, 2013; Messick, 1989, 1995).

5.2.3. Messick's Model of Construct Validity

As we have mentioned above, Samuel Messick (1989, 1990, 1994, 1995, 1996, 1998) thinks of validity as an overall evaluative concept. According to him, "the essence of unified validity is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable" (Messick, 1995, p.747). In this perspective, he maintains that "both meaning and values are integral to the concept of validity" (p.747). Messick sketches out his conceptualization of construct validity in a four-fold classification framework comprising two columns crossing two horizontal rows (Table 15). The columns represent the function and outcome of testing; and the rows represent the source of justification for the information included in the columns. The first column accounts for score interpretation (meaning); and the second one delineates the purposes for which the test outcome (scores) can be used. In the same way, the source of justification, which is supposed to provide logical support for the trustworthiness of score interpretations and the decisions that we intend to make, is provided by two types of information: evidential based and consequential based information.

Table 15: Facets of Validity as a Progressive Matrix

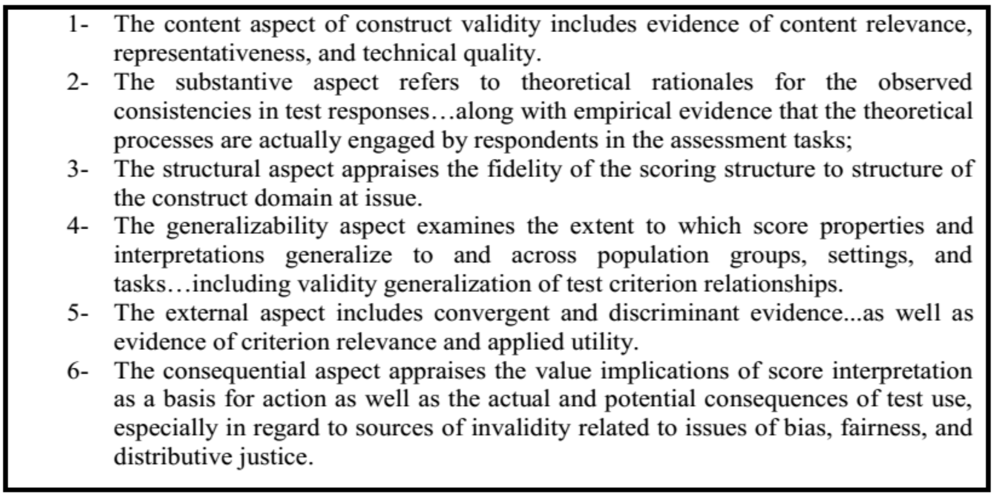
	Test Interpretation	Test Use
Evidential Basis	Construct Validity (CV)	CV + Relevance/Utility (R/U)
Consequential Basis	CV + Value Implications (VI)	CV + R/U + VI + ·Social Consequences

Source : Messick, 1995, p. 746

5.2.3.1. The Source of Justification

As included in Fig 31, the source of justification refers to the extent to which all the accumulated types of evidence give logical support to the score interpretation and uses. Messick (1989, 1994, 1995) emphasizes that speaking of validity as a unitary concept does not imply that we cannot gather information from different sources to justify the score interpretations and uses. In this context, he distinguishes six sources of evidence which include content, substantive, structural, generalizability, external, and consequential aspects. The first five aspects are classified within the evidential basis; while the last aspect forms a part of the consequential basis.

Fig 31: Aspects of Construct Validity



1- The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality.
2- The substantive aspect refers to theoretical rationales for the observed consistencies in test responses…along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks;
3- The structural aspect appraises the fidelity of the scoring structure to structure of the construct domain at issue.
4- The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks…including validity generalization of test criterion relationships.
5- The external aspect includes convergent and discriminant evidence...as well as evidence of criterion relevance and applied utility.
6- The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice.

Source: Messick, 1995, pp. 248-9

5.2.3.1.1. The Evidential Basis of Construct Validity

As we have indicated previously, the evidential basis of construct validity requires the accumulation of five types of information which include evidence based on content, evidence based on response processes (substantive), evidence based on internal structure (structural), evidence based on relations to other variables (external) and evidence based on score generalization (AERA, APA, & NCME, 1999; Messick, 1989, 1996, 1998). The content aspect of construct validity provides evidence about construct representation, content relevance and coverage. The substantive aspect or evidence based on response processes provides "evidences concerning the fit between the construct and the detailed nature of performance or response actually engaged by examinees" (AERA, APA & NCME, 1999, p. 12). The structural aspect or 'structural fidelity' investigates the extent to which the scoring criteria reflect the aspects of the construct to be measured. Evidence based on score generalization provides information about groups and contexts beyond the test to whom or where test scores are to be generalized. The external aspect of construct validity "may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs" (AERA, APA, and NCME, 1999, 13). In other words, using different methods to measure similar constructs can yield high levels of correlation (convergent validity). Conversely, discriminant validity tells us that using similar measures to assess different constructs can yield low level of correlation. The latter "is particularly critical for discounting plausible rival alternatives to the focal construct interpretation" (Messick, 1995, p. 746).

5.2.3.1.2. The Consequential Basis of Construct Validity

The consequential aspect of construct validity "includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term" (Messick, 1995, p. 746). Language tests are commonly administered for the purpose of generating scores (Bachman, 1990). The scores are, then, interpreted as indicators of test takers' levels of language ability. The score interpretations are mostly used as a basis for making decisions about test takers and institutions. The decisions can, for example, include "student selection, certification, classification, tracking, promotion or retention in educational programs, and allocating resources to schools" (Bachman & Purpura, 2008, p. 456). They can also be used for political reasons such as restricting the number of immigrants, depriving minority groups of their social and political rights, or in determining citizenship (McNamara & Roever, 2006; McNamara & Shohamy, 2008; Shohamy, 1996, 2000, 2001). So, in order to use score interpretation as a justification for making decisions, Bachman (2004a) suggests that we need to respond to three questions "What decisions are we going to make on the basis of test scores? How relevant is the ability we are measuring to make these decisions? How useful are the test scores for making these decisions?" (261). The decisions will certainly have consequences on participants and institutions. These consequences fall into two types: intended or beneficial (positive) and unintended or harmful (negative) (Bachman, 2004a, 2005; 2013; Bachman & Purpura, 2008; McNamara, 2006, 2008). Intended consequences result from the intended uses of test scores. According to Bachman & Purpura (2008) “if used as intended, tests will maximize the chances for fair and equitable treatment of individuals and groups in terms of their access to opportunities based on merit” (p.461). Conversely, unintended consequences result from unintended uses of test scores. Inadvertent consequences can deny test takers their right of certification, graduation, entrance to institutions or minimize their chances for joining employment positions. This is why Bachman (2004a) reminds test users that they need to respond to these questions before making any type of decisions:

> (a)Who will be affected by this use of the test scores, and how? (b)What institutions, organizations, agencies, or segments of society, will be affected by this use of the test scores, and how? (c) What are the possible positive consequences of this use of the test scores? How likely is it that these will happen? (d)What are the possible negative consequences of this use of the test scores? How likely is it that these will happen? (Bachman, 2004a, p.261 [parentheses added])

Consequently, In order to minimize the effects of adverse consequences on examinees, language testers advise test users not to consider "using scores from a test for making decisions if questions about score reliability or the validity of interpretations are raised" (Bachman & Purpura, 2008, p. 461).

5.2.4. Sources of Invalidity

Language testers identify two main sources that threaten and distort the validity of test score interpretations and uses. These include construct underrepresentation and construct irrelevant variances (AERA, APA & NCME, 1999; Bachman, 1990, 2004a; Bachman & Palmer, 1996; McNamara & Roever, 2006; Messick, 1989, 1995). The former refers to the extent to which a test:

> fails to capture important aspects of the construct. It implies a narrowed meaning of test scores because the test does not adequately sample some types of content, engage some psychological processes, or elicits some ways of responding that are encompassed by the intended construct (AERA, APA & NCME, 1999, p. 10).

As the definition above implies, construct underrepresentation entails that the assessment is 'too narrow' in that it fails to cover important features relevant to content relatedness, content coverage, or correspondence between test tasks and target language tasks. Additionally, this can extend to display the test inability to fully assess the construct to be measured in terms of psychological trait or from performance-based perspectives.

Concerning construct irrelevant variances, these can affect test scores when "the assessment is too broad, containing excess reliable variance associated with other distinct constructs as well as method variance" (Messick, 1995, p. 742). Messick classifies construct irrelevant variances into two sets: construct irrelevant difficulty and construct irrelevant easiness. In the former, the features of tasks and skills that are external to the construct to be measured make the test input inappropriately difficult for some examinees rather than others. This can, for example, occur in cases when the test content includes some topics that may seem to be offensive to some individuals or groups, or when the administration and scoring procedures are not standardized in all the examination or rating centers. This type of construct-irrelevant variances leads to "scores that are invalidly low for those individuals adversely affected" (Messick, 1994, p.10). Contrariwise, construct-irrelevant easiness may enable some test takers to respond correctly to the tasks because of their familiarity with the test content. This type of variance "leads to scores that are invalidly high for the affected individuals as reflections of the construct under scrutiny" (p.10).

5.3. Test Validation

Language test validation refers to the practical steps that we conduct in order to support or discredit the interpretations provided for the scores obtained in a given testing situation. This is to ensure that the decisions intended to be made as a results of these interpretations; and the consequences that may affect the participants and institutions because of these decisions will be valid (AERA, APA & NCME, 1999; Bachman, 2005, 2013; Chapelle, 2012; Chapelle, Enright & Jamieson, 2008, 2010; Kane, 2013). The process of validation involves a chain of empirical reasoning staring from score meaning analysis and culminating with solid inferences and conclusions. The first step in the train of reasoning or argument involves the examination of test takers scores. The second step requires providing meaning (interpretations) to these scores. For example, if test takers obtain good marks, this will be interpreted that they have a high level of language ability; or they can use the language fluently in non-test target contexts. However, if they obtain low marks, this means that their level of language ability is low. The reasoning from the first step to the second one needs to be supported with solid justifications**.** The process of validation, then, engages in evidence collection**.** If all types of evidence (content/ criterion/ construct) reinforce the score interpretations, the test scores will be considered as valid. If the evidential basis rebuts the solidity of the gathered information, this may invalidate the score interpretations and the resulting decisions.

5.3.1. Definition of validation

Before we provide an explanation to the method through with we can validate score interpretations, let us first review some of the definitions that have been proposed to 'validation' in the literature of language testing. According to AERA, APA & NCME (1999):

> Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use. The conceptual framework points to the kinds of evidence that might be collected to evaluate the proposed interpretation in the light of the purposes of testing (p. 9).

On his part, Messick (1995) considers validation as “an empirical evaluation of the meaning and consequences of measurement. As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use" (p.742 [parentheses in original]). In his book *'Fundamental Considerations in Language Testing'*, Bachman (1990) points out that validation refers to "the process of building a case that test scores support a particular interpretation of ability, and it thus subsumes content relevance and criterion relatedness"(p. 290). In the point of view of Kane (2006 as cited in Kane, 2012b) "to validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the proposed conclusions and decisions" (p. 3).

5.3.2. Arguments in Language Test Validation

Due to the fact that the process of validation in language testing is conducted by the incorporation of arguments and mainly of Stephan Toulmin's model (1958, 2003); and for a better understanding of this process, let us first to distinguish between the terms 'argumentation' and 'argument'. In their book '*Introduction to Reasoning'*, Toulmin, Rieke and Janik (1984) consider the former as: "the whole activity of making claims, challenging them, backing them up by producing reasons, criticizing those reasons, rebutting those criticisms, and so on" (p. 14). However an argument can be defined as:

> a set of assumptions (i.e., information from which conclusions can be drawn), together with a conclusion that can be obtained by one or more reasoning steps (i.e., steps of deduction). The assumptions used are called the support (or, equivalently, the premises) of the argument, and its conclusion (singled out from many possible ones) is called the claim (or, equivalently, the consequent or the conclusion) of the argument. The support of an argument provides the reason (or, equivalently, justification) for the claim of the argument. (Besnard & Hunter, 2008, p.2)[parentheses in original].

5.3.2.1. The Structure of Toulmin's Arguments

Stephan Toulmin (1958, 2003) organizes his model of arguments into six components which include: claims, data, warrants, backings, qualifiers and rebuttals (Bachman, 2005, 2013; Kane, 2013; Mislevy, Russell & Almond, 2003; Toulmin, Rieke &

Janik, 1984). Toulmin (2003) defines the claim (C) as an assertion or "a conclusion whose merits we are seeking to establish" (p. 90), and he refers to the data (D) as "the facts we appeal to as a foundation for the claim" (p.90). The move from the data to the claim can be stated in the form of a hypothesis or a deduction. For example, "('If D, then C')….Or this can profitably be expanded, and made more explicit: 'Data such as D entitle one to draw conclusions, or make claims, such as C', or alternatively 'Given data D, one may take it that C'" (Toulmin, 2003, p. 92). The third component 'the warrant' (W) is used to justify or authorize the chain of inferences or the move from (D) to (C) so as to give legitimacy to the deduction (Bachman, 2004b).

Now if we consider the information provided in Fig 32, the argument goes on as follows:

A: Harry is a British subject (the claim).

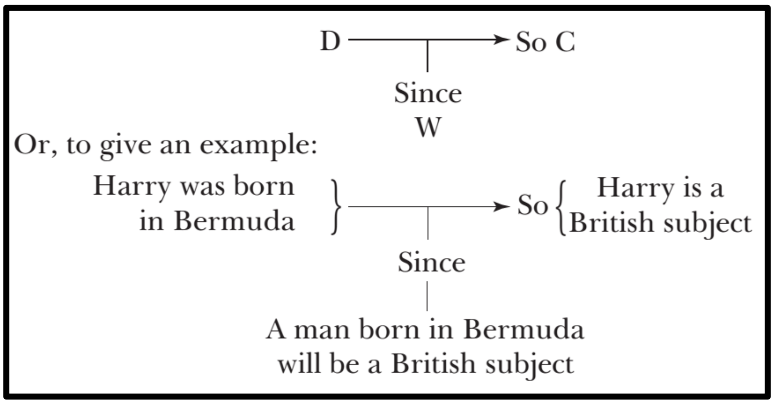
B: How did you know?

A: Given he was born in Bermuda, he becomes a British subject (the datum).

B: On what grounds have you built your assumption?

A: There is a legal decree which implies that people who are born in Bermuda will be granted British citizenship (the warrant).

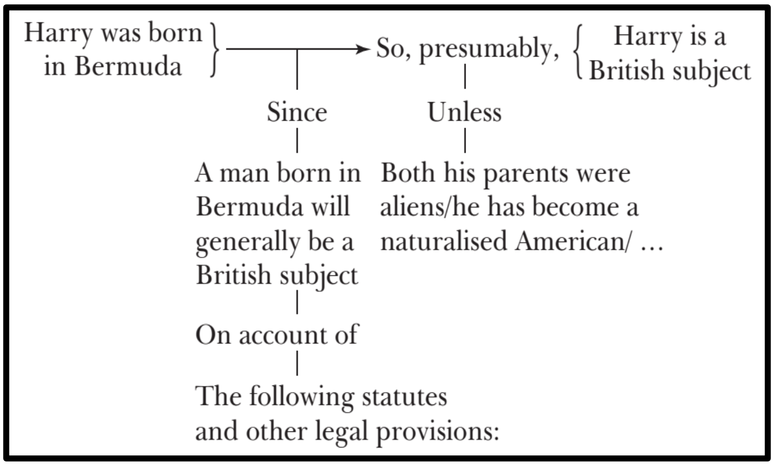
Fig 32: Toulmin's Data, Claims and Warrants



Source: Toulmin, 2003, p. 91

If the questioner assumes that the warrant lacks validity, weightiness or soundness, he may ask for further evidence to believe in the trustworthiness of the justification. In this case, (A) needs to reinforce his warrant with a 'Backing' (B) which consists of "assurances without which warrants themselves would possess neither authority nor currency" (Toulmin, 2003,p. 96). The next component refers to the qualifier (Q). The latter which can take different forms such as 'probably', 'possibly', 'certainly', 'surely', or 'presumably', refers to the degree of force or support that warrants confer on the claims (Hitchcock & Verheij, 2006, Toulmin, 2006). The sixth component of Toulmin's argument is the rebuttal (R). This constituent refers to “the exceptional conditions which might be capable of defeating or rebutting the warranted conclusion” (Toulmin, 2003, p. 94). On the one hand, the rebuttal can provide more credibility to the conclusions or claims; and it can also override them on the other (see Fig 33).

Fig 33: The Role of Backings, Qualifiers and Rebuttals


Harry was born in Bermuda
So, presumably,
Harry is a British subject
Since
Unless
A man born in Bermuda will generally be a British subject
Both his parents were aliens/he has become a naturalised American/ …
On account of
The following statutes and other legal provisions:

Source: Toulmin, 2003, p. 97

As we have mentioned in p. (150), we suppose that (B) has not been convinced with the justification provided by (A); hence, the dialogue in p. (149) will go on like this:

B: I doubt if that (granting citizenship) can really happen.

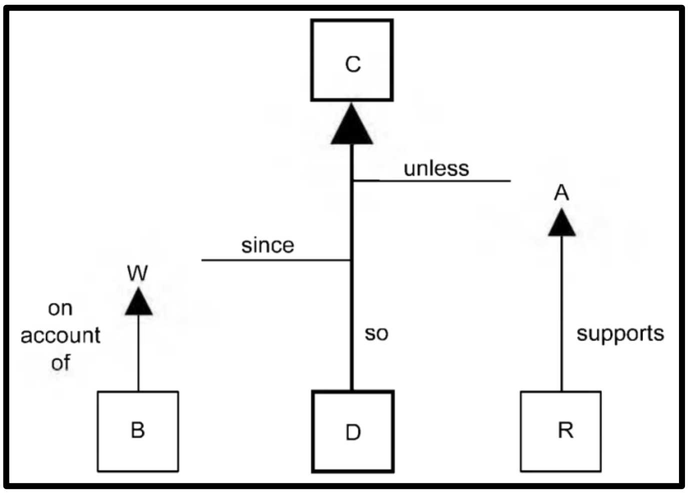
A: Why not? Harry is presumably a British subject according a law passed by the parliament (qualifier and Backing).

Of course, this hypothesis could be overridden if Harry's parents were aliens, or if he were granted American citizenship. In sum, the chain of inferences in Toulmin's argument starts when "reasoning flows from data (D) to claim (C) by justification of a warrant (W), which in turn is supported by backing (B). The inference may need to be qualified by alternative explanations (A), which may have rebuttal evidence (R) to support them" (Mislevy & Riconscente, 2006, p. 70).

5.3.2.2. The Incorporation of Toulmin's Argument in Language Test validation

The researchers who adopted, modified and implemented Toulmin's arguments in language test validation were R.J. Mislevy, L. S. Steinberg and R. G. Almond, and more precisely in their seminal article '*On the Structure of Educational Assessments*' published in *Measurement: Interdisciplinary Research and Perspectives*' (Mislevy, Steinberg & Almond 2003). Since then, their framework has widely been incorporated in language testing (Bachman, 2005, 1013; Chapelle, 2012; Chapelle, Enright & Jamieson, 2008, 2010; Kane, 2006, 2013; Mislevy & Riconscente 2006). As it illustrated in Fig 34, Mislevy et al's (2003) modified framework consists of five components: the claim, the datum, the warrant, the backing and the rebuttal.

Fig 34: Toulmin's Model in Language Assessment

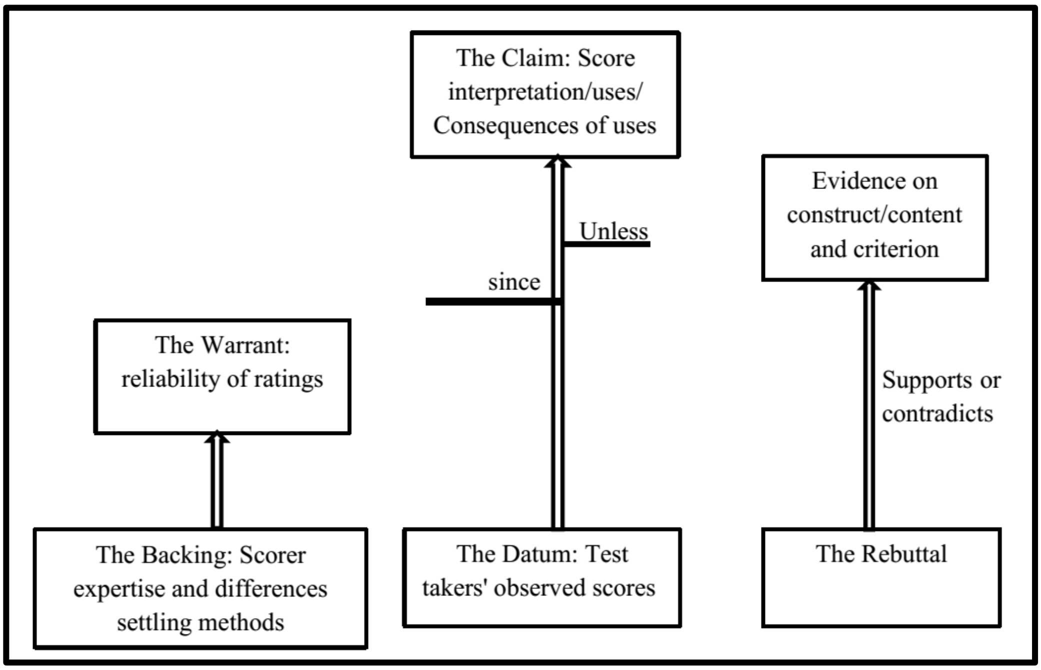


Source: Mislevy et al., 2003, p.11.

The claims refers to the interpretation of what test takers have in terms of language ability and/or to what they can do in terms of their capacity of using language in situations beyond the test. The data, according to the authors, refer to test takers' performance within a testing situation. The warrant is used to justify the interpretations based on learners' responses on the test. For example, a good mark can justify the claim that a given examinee has a high level of language ability and vice versa. Following Toulmin (1958, 2003), Mislevy et al., (2003) emphasize that "warrants themselves require *backing* (B), in the form of theories, research, data, or experience. The substantive foundations of warrants in assessment are our beliefs about the nature of knowledge and how it is evidenced" (p. 12)[italics in original]. If empirical analysis about construct representation, content relevance and coverage or criterion relatedness come to support the warranted reasoning, we can assume that the interpretations are valid; if the warranted chain of inferences is challenged by the available evidence, the interpretations would not be considered valid (Kane, 2013).

The scope of these components has been extended in Kane's interpretive argument (Kane, 2004, 2006, 2008); Bachman's 'assessment utilization argument' (Bachman, 2005, 2013); Bachman and Palmer's justification arguments (Bachman & Palmer, 2010) and Kane's interpretation/use argument (Kane, 2012b, 2013). These assessment arguments, as Fig 35 implies, consider the claim to include the meaning or the interpretation that we provide for test scores, the purposes for which the scores will be used, decision making and the potential consequences that may affect participants and institutions. The datum, in this framework, refers to the scores obtained by test takers on a given test. The chain of inferences from the scores to the claim is warranted by the reliability of the scoring processes. The consistency of scoring can be backed by raters' expertise and methods for settling raters' differences. Since reliability is a necessary condition for validity (Bachman, 1990; Kane, 2012a, 2012b, 2013), once the scoring procedures are proven to be inconsistent; the validity of the interpretation and uses will also be discredited. If the scoring is found to be reliable (warrant/backing); we need to gather more evidence (the evidential basis) to examine whether the test has really measured the construct intended to be measured; and to see whether the test content is relevant to and samples from the syllabus content (AERA, APA & NCME, 1999; Bachman, 2005; Messick, 1989, 1995). If the available evidence (construct representation/ content relevance and coverage/ criterion relatedness) supports the plausibility of the score interpretations and uses, these interpretations and uses will be considered to be valid; if the collected evidence disproves or rebuts the warranted information, the score interpretations and uses will be considered as invalid.

Fig 35: The Structure of Assessment Argument


The Claim: Score interpretation/uses/ Consequences of uses
Unless
since
Evidence on construct/content and criterion
The Warrant: reliability of ratings
Supports or contradicts
The Backing: Scorer expertise and differences settling methods
The Datum: Test takers' observed scores
The Rebuttal

Organized from Toulmin, 2003; Mislevy et al., 2003; Bachman, 2004a, 2005, 2008, 2013; Bachman & Palmer, 2010; Kane, 2004, 2006, 2008, 2012a, 2012b, 2013.

5.4. Relationship between Reliability and Validity

The most fundamental concepts in the evaluation of language tests are reliability and validity (Kane, 2010, 2013; Miller, Linn & Gronlund, 2009; Gronlund, 1987; Messick, 1989; Tavakoli, 2012). Reliability is a requirement of test scores and investigates the extent to which measurement is free from errors. Validity is a quality for test score interpretations and uses (Messick, 1989). In the field of educational and psychological testing, reliability attempts to answer these questions "how much variance in test scores is due to measurement error? [and] How much variance is due to factors other than measurement error?" (Bachman, 1990, p. 240); whereas validity attempts to respond to this question “What specific abilities account for the reliable variance in test scores?” (p. 240). The investigation into the relationship between these two requirements leads us to raise questions like: can there be reliability without validity; or can there be validity without reliability? (Bachman, 1990; Henning, 1987; Lee, 2003; Miller, Linn & Gronlund, 2009; Mislevy, 2004; Moss, 1994). As far as the first question is concerned, psychometricians and language testers agree on the fact that reliability which “is a necessary condition for validity has always been regarded as a fundamental principle in psychometrics” (Lee, 2003, p. 90). This is because unreliable test scores “cannot provide a basis for valid interpretation and use” (Bachman, 1990, p.289). Concerning whether there can be validity without reliability. Henning (1987) responds that ‘yes’ “it is possible for a test to be reliable without being valid for a specified purpose, but it is not possible for a test to be valid without first being reliable” (pp. 89-90). In the same way, Moss (1994) backs this hypothesis; if, according to her, by reliability we mean consistency of scoring (Lee, 2003, Mislevy, 2004). A test can be reliable without being valid only in limited contexts and for specific purposes. This can, for instance, occur when we want to diagnose learners in order to place them at different levels (Spolsky, 1995). Conversely, in the case of achievement tests or in examinations that focus on measuring mental or contextual constructs, validity is considered as the most fundamental concept. This is because if reliable test scores do not reflect the construct that it is intended to be measured, the interpretations and uses will certainly lead to unintended consequences (Messick, 1989). Davies (2012) summarizes the relationship of reliability to validity in these lines “reliability gives form to a test; validity gives it its meaning…the higher a test’s reliability, the greater the possibility for validity, but ‘if one could demonstrate that a measure has good validity, its reliability can be assumed and becomes a secondary issue’ " (p. 38).

Conclusion

The conceptualization of validity has been revisited several times since the last half of the twentieth century (Cronbach & Meehl, 1955). Successive definitions and models have been proposed to identify the meaning and role of the concept. For the traditional paradigm, for instance, validity refers to the degree to which a test measures what it claims to measure. This trend splits the concept into three distinct types: criterion, content and construct validities. The criterion model is implemented to justify selection and placement purposes. The content model is used to measure the extent of authenticity between test tasks and the instructional syllabus tasks. The construct model attempts to examine the degree to which tests measure the traits they claim to measure. Conversely, the unitary trend regards validity as an overall evaluative concept concerned with the examination of the plausibility of test score interpretation, uses and consequences. According to this school, evidence supportive for score interpretation and uses can be collected by means of an integrated process involving criterion (convergent and discriminant), content, construct, substantive, structural, external, generalizable and consequential considerations.

The empirical phase of validity is conducted by the implementation of validation arguments and more specifically by the incorporation of Toulmin's framework (1958, 2003) comprising the datum, the claim, the warrant, the backing, the qualifier, and the rebuttal. The 'datum' refers to test takers' scores on the test. The 'claim' summarizes the testers' interpretations of these scores and the purposes for which they will be used. The 'warrant' justifies the chain of inferences that testers make from the datum to the claim. The 'backing' gives more force to the warrant. The 'qualifier' displays the degree of force of warrants; and the 'rebuttal' may support, weaken, or reject the credibility of the score interpretations.

Chapter Six: Field Study

Validating the Score Interpretations of EL-Oued Technology Streams' BAC English Tests

Chapter Six: Field Study

Validating the Score Interpretations of EL-Oued Technology Streams' BAC English Tests

Introduction

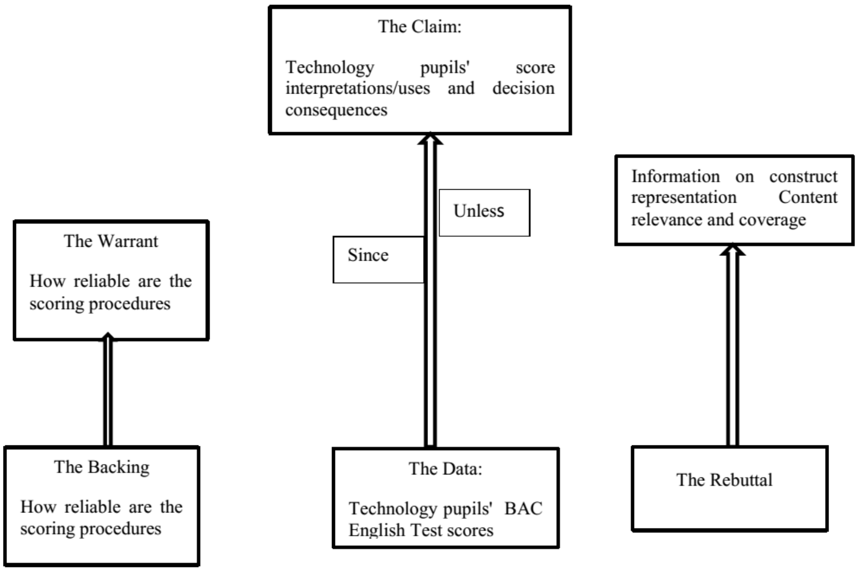
Chapter six 'field work' focuses on the analysis of the data that we previously collected by means of the questionnaire, the interview and the documentary sources. The data included in the first two instruments seek to verify hypothesis one which assumes that the scoring practices in the BAC English rating centers are not reliable. On its part, the information in the documentary sources attempt to test hypothesis two, three and four which postulate that the BAC English tests in technology streams lack four aspects of construct validity as a unitary concept: construct representation, content relevance, domain coverage and criterion relatedness.

The results of the analysis will be incorporated in the validity arguments that we intend to build for the purpose of reinforcing or discrediting the interpretations provided for technology pupils' scores from 2001 to 2006; and the purposes for which these scores have been used. The argument will include these constituents: the datum (technology pupils' observed scores), the claim (the score interpretations), the warrant (the scoring processes), the warrant (scoring expertise and mediation methods), and the rebuttal (the BAC English tests' topical content).

6.1. Components of the Validity Argument in Technology Streams

In the same way as language testers and educational measurement specialists, the validity argument that we will implement in evaluating the credibility of technology pupils' score interpretations is the one proposed by Toulmin (1958, 2003) and modified by Mislevy et al., (2003). The structure of this argument, as Fig 36 illustrates, includes the following constituents:

Fig 36: Structure of Toulmin's Argument


The Claim:
Technology pupils' score interpretations/uses and decision consequences
Unless
Since
Information on construct representation Content relevance and coverage
The Warrant
How reliable are the scoring procedures
The Backing
How reliable are the scoring procedures
The Data:
Technology pupils' BAC English Test scores
The Rebuttal

Adapted to language assessment by Mislevy, Steinberg, & Almond, 2003, p.11.

6.1.1. The Datum: It refers to Eloued Technology pupils' BAC English test scores in seven sessions (2001-2006). What is worth mentioning here is that in 2001 two BAC sessions have been organised: the first in June and the second in September (see appendix B).

6.1.2. The Claim: The score interpretations, decisions and consequences of uses.

6.1.3. The Warrant: Information gathered by means of the questionnaire and the interview about the reliability of the scoring procedures.

6.1.4. The Backing: Information from the questionnaire and the interview about raters' expertise and methods for settling their differences.

6.1.5. The Rebuttal: Evidence gathered from documentary sources (BAC English tests from 2001 to 2006 (see appendix B) and technology streams' third year syllabus about construct representation, content relevance and coverage.

6.2. Describing Test Takers' Scores

According to Gronlund (1977) test scores can be described with reference to two types of measures: the average score (the central tendency) and the spread of scores (measures of variability). Concerning the first type, Gronlund points out that "statisticians frown on the use of the term "average"…because there are a number of different types of average. [Thus,] it is more precise to use the term that denotes the particular average being used" (121). Statisticians identify three types of average: the median, the mean and the mode (Ebel and Frisbee, 1991; Miller, Linn & Gronlund, 2009).

6.2.1. The Mode

The mode, which is the most frequently occurring score, can have more than one value. The mode can be determined by examining the score with the highest frequency, or by "find[ing] the score with the largest number of test takers" (Bachman, 2004a, p. 55).

6.2.2. The Median

The median (the counting average) can be determined by organizing the scores in a given order of size (from top to bottom, or the other way around) and counting up or down to the midpoint of the list; and the median will be the score above which and below which the half of the marks is found. If the list contains an even number of scores, the median will computed by averaging the two middle scores (Ebel and Frisbee, 1991; Miller, Linn & Gronlund, 2009; Tavakoli, 2012).

6.2.3. The Mean

The mean or the arithmetic average is the most common used measure of central tendency. This measure can be determined by adding up all of the scores obtained by the examinees on a given test, and then dividing the sum by the total number of the scores. The mean can be computed by using the following formula:

$$\overline{X} = \frac{\Sigma X}{N}$$

Where $\overline{x}$ (X-bar) is the mean

∑: This represents the summation sign.

N: refers to the total number of the scores

∑X: The sum of the obtained scores

Which implies: $\overline{X} = \frac{\text{Sum of all scores}}{\text{Number of scores}}$

6.2.4. The Frequency Distribution

The frequency distribution refers to a table or a diagram which displays the number of occurrences (frequencies) "of values of any given variable. For QUALITATIVE VARIABLEs this is the number of times each of the categories occurs whereas for QUANTITATIVE VARIABLEs this is the number of times each different score (or range of scores) occurs" (Tavakoli, 2012, p. 236 [Capitalization in original]). The frequency distribution is a two-column list. The first column includes all the scores obtained by test takers organized from highest to lowest; and the other column (the frequency column) shows the frequency of occurrences for each score (Ebel and Frisbee, 1991; Miller, Linn & Gronlund, 2009, Tavakoli, 2012). However, the grouped frequency distribution "lists frequencies for class intervals rather than individual scores. The data are grouped in intervals of equal range and each frequency represents the number of data values in one of the intervals" (Tavakoli, 2012, p. 236).

Concerning the measures of variability, these include the range and the standard deviation. The former refers to "the interval between the highest and lowest scores" (Gronlund, 1977, p. 121). As for the 'standard deviation', this measure is composed of two terms: standard and deviation. The latter "refers to the difference between an individual score in a DISTRIBUTION and the average score for the distribution" (Tavakoli, 2012, p. 615) [Capitalization in original]. The term standard means typical, "therefore, a *SD* is the typical, or average, deviation between individual scores in a distribution and the MEAN for the distribution" (p.615[Italics and capitalization in original]). In this context, the deviation score can be thought of the extent to which an individual score deviates from the mean of that distribution.

Gronlund (1977) and Miller, Linn and Gronlund (2009) state that the simplest method for describing and interpreting test scores, especially when the number of examinees is not large is to implement the range and the median. The first step is to arrange the set of scores in order of size. Then we can count up or down until we locate the midpoint of the list of scores (see Table A. 2)**.** The range of scores can be determined by subtracting the lowest score from the highest one.

As far as this research is concerned, there are six technical schools in the 'wilaya' of Eloued; apart from 'Djemaa' school which contains one technology specialty 'civil engineering', each of the other schools contains two specialties: electrical and mechanical engineering. Concerning the scores obtained by technology pupils in seven BAC sessions (2001-2006), these were provided to us in two forms: detailed and abridged lists. In

'Guémar' technical school, we were provided free access to the pupils' BAC score records. In this file, every single mark of the pupils from 1998 until 2006 is documented. Conversely, in the other schools, the information concerning this issue is scarce and not of much details in that it is limited to categorizing the pupils into two sets: the pupils who got marks above average and those who were ranked below average in English (Guèmar Technical School, 1998-2006; Orientation Centre of Eloued, 2001-2006).The other point that we would like to mention is that the analysis of the pupils' marks is not an end in itself. Our concern is to use these marks as the datum upon which the validation argumentation will be conducted.

6.3. Analysis of Eloued Technology Pupils' Scores from 2001 to 2006

6.3.1. Analysis of Electrical Engineering Scores from 2001 to 2006

The frequency distribution of electrical engineering streams' scores in 2001 implies that the most recurrently score (the mode) was (3). Additionally, the scores which fall in intervals 0-4 count 21 and the ones in 4-8 count 15. In other words, the scores in intervals 0-4 and 4-8 form a percentage of 80%. The counting average resulting from this session was (4) and the arithmetic average was 4.9 (see Table A 2). The students who got marks above average in 2001 count 5 out of 45 with a success rate of 11.11%.

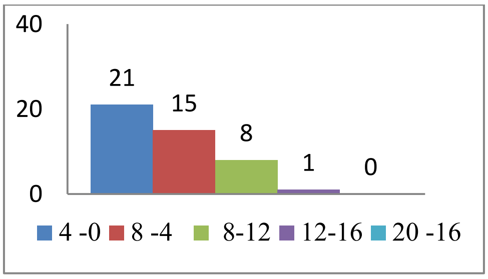
Table 16: Frequency Distribution of 2001 BAC English Test Scores.

Test Scores (X)	Frequency (f)		Test Scores (X)	Frequency (f)
12	1		5	2
11	1		4.5	2
10.5	1		4	2
10	2		3.5	3
9	2		**3**	**7**
8.5	1		2.5	3
8	1		2	3
7	4		1.5	5
6	4		1	1
5.5	1			

Table 17: Grouped Score Frequency Distribution of the 2001 Sessions

Score Interval	Midpoints	Frequency
0-4	2	21
4-8	4	15
8-12	10	8
12-16	14	1
16-20	16	0

Graph 1: Histogram of 2001 Grouped Score Frequency Distribution



In 2002, as Table 18 implies, the score 5 has reoccurred for 10 times. The majority of the scores (28) assemble in interval 4-8. The median of the scores during this session was 5 and the arithmetic average (the mean) was after its rounding 4.9. In this session, no test taker was able to obtain a score equal or above average.

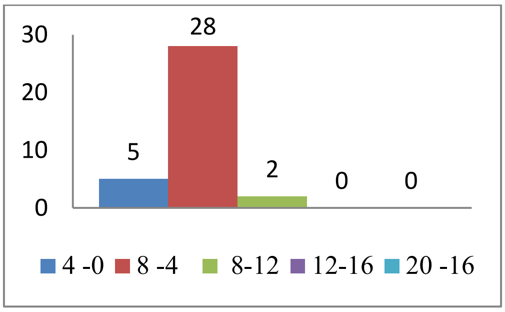
Table 18: Frequency Distribution of 2002 BAC English Test Scores.

Test Scores (X)	Frequency (f)
8.5	1
8	1
7.5	1
7	1
6.5	1
6	1
5.5	4
5	**10**
4.5	5
4	5
3.5	2
3	1
2.5	1
00.5	1

Table 19: Grouped Score Frequency Distribution of the 2002 Session

Class interval	Midpoints	Frequency
0-4	2	5
4-8	6	28
8-12	10	2
12-16	14	0
16-20	18	0

Graph 2: Histogram of 2002 Grouped Score Frequency Distribution



In 2003, the distribution of scores formed a bimodal frequency; in that each of the scores 6.5 and 5 has reoccurred for 4 times. Additionally, the largest number of the scores (24 scores) gather in interval 4-8 which represents a percentage of 85.71 % of the whole number of the marks. The median of the obtained scores was 6 and the mean was 5.8. Again in this session, no student was able to get a score equal or above average which implies that the rate of success in the BAC English test was 00%.

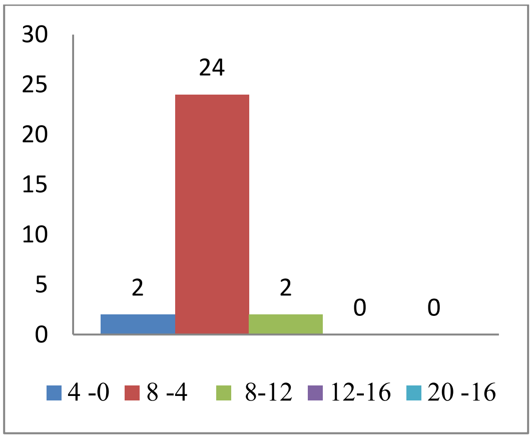
Table 20: Frequency Distribution of 2003 BAC English Test Scores.

Test Scores (X)	Frequency (f)
9.5	2
7.5	3
7	3
6.5	**4**
6	3
5.5	2
5	**4**
4	3
4	2
2.5	1
1	1

Table 21: Grouped Score Frequency Distribution of The 2003 Session

Class interval	Midpoints	Frequency
0-4	2	2
4-8	6	24
8-12	10	2
12-16	14	0
16-20	18	0

Graph 3: Histogram of 2003 Grouped Score Frequency Distribution



In 2004, the distribution tells us that the score (6) was the mode which has reoccurred for 4 times. Moreover, 4 scores group in interval 0-4, and the rest of the scores in interval 4-8. The median of the distribution was 3, and the computed mean was 3.6. In the same way as the previous session, the rate of success in English was 00%.

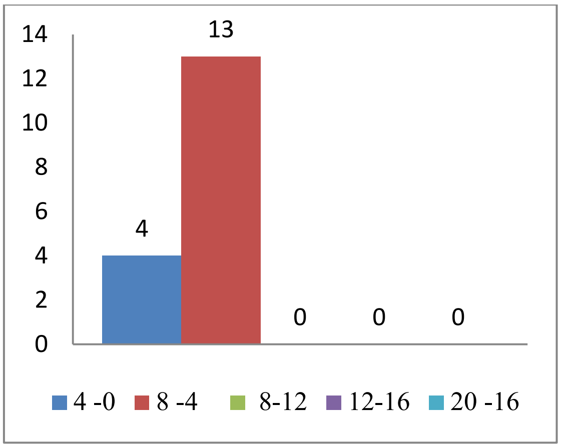
Table 22: Frequency Distribution of 2004 BAC English Test Scores.

Test Scores (X)	Frequency (f)
7	2
6.5	1
6	**4**
5.5	2
5	1
4	3
3.5	2
3	1

Table 23: Grouped Score Frequency Distribution of the 2004 Session

Class interval	Midpoints	Frequency
0-4	2	4
4-8	6	13
8-12	10	0
12-16	14	0
16-20	18	0

Graph 4. Histogram of 2004 Score Grouped Frequency Distribution



In 2005, the most frequently reoccurred score was (3) which has been repeated for 6 times. Concerning the grouped frequency distribution, 20 scores are included in interval 0-4; 10 scores in interval 4-8; and 1 score in interval 8-12. The median of the distribution was 3, and the computed mean was 3.6. The rate of success in this session was 00%.

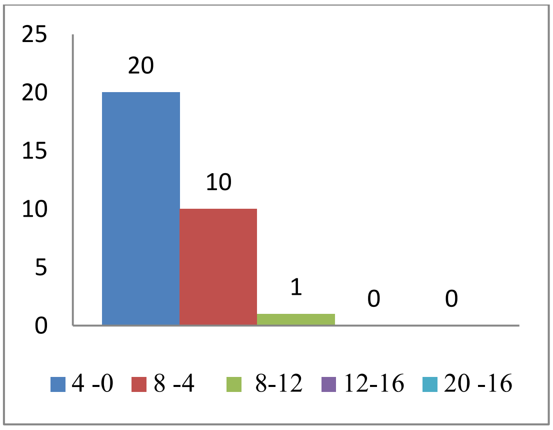
Table 24: Frequency Distribution of 2005 BAC English Test Scores.

Test Scores (X)	Frequency (f)
10.5	1
6.5	1
6	1
5.5	2
5	1
4.5	1
4	3
3.5	5
3	**6**
2.5	4
2	4
1.5	2

Table 25: Grouped Score Frequency Distribution Of The 2005 Session

Class interval	Midpoints	Frequency
0-4	2	20
4-8	6	10
8-12	10	1
12-16	14	0
16-20	18	0

Graph 5: Histogram of 2005 Grouped Score Frequency Distribution



In 2006, the mode was (9) with 5 occurrences. The grouping of the scores in this session witnessed some improvement in that 17 scores are assembled in intervals 0-4 and 4-8; and 12 scores are accumulated in intervals 8-12 and 12-4. Equally important, the rate of success in the BAC English test rose to 19%.

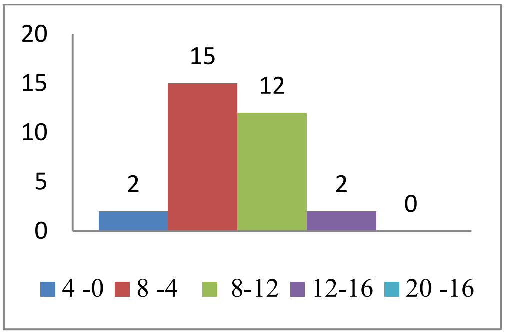
Table 26: Frequency Distribution of 2006 BAC English Test Scores.

Test Scores (X)	Frequency (f)
12.5	1
12	1
11.5	2
10.5	1
10	1
9	**5**
8.5	2
8	1
7.5	3
7	1
6.5	3
6	1
5.5	2
5	1
4.5	4
2.5	1
2	1

Table 27: Grouped Score Frequency Distribution of the 2006 Session

Class interval	Midpoints	Frequency
0-4	2	2
4-8	6	15
8-12	10	12
12-16	14	2
16-20	18	0

Graph 6: Histogram of 2006 Grouped Score Frequency Distribution



6.3.2 Analysis of Mechanical Engineering Scores from 2001 to 2006

The frequency distribution of mechanical engineering scores in 2006 suggests that the mode was 5.5 with five occurrences. Concerning the grouping of the scores, it can be described as follows: in interval 0-4, we can count 8 scores; in 4-8, there are 16 scores; 6 marks in 8-12; 7 marks in 12-16; and 1 score in interval 16-20. The median of the obtained scores was 6 and the mean was 7.3 (see Table A. 2). In this session, the rate of success in the BAC English test reached 31.57%.

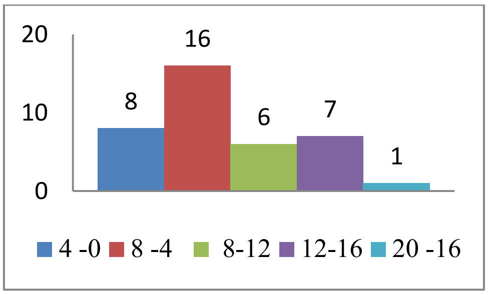
Table 28: Frequency Distribution of BAC English Test Scores for 2001 Sessions.

Test Scores (X)	Frequency (f)		Test Scores (X)	Frequency (f)
16	1		6.5	2
14	1		6	3
13.5	2		**5.5**	**5**
13	1		5	3
12.5	1		4	1
12	2		3.5	1
11.5	1		3	1
11	2		2.5	4
10.5	1		2	1
8	2		0.5	1
7.5	1			

Table 29: Grouped Score Frequency Distribution of the 2001 Sessions

Class interval	Midpoints	Frequency
0-4	2	8
4-8	6	16
8-12	10	6
12-16	14	7
16-20	18	1

Graph 7: Histogram of 2001 Grouped Score Frequency Distribution



In 2002, as Table 30 implies, the most reoccurring score was 5.5 with 9 occurrences. Additionally, the scores are arranged as follows: 4 scores fell in interval 0-4; 33 scores in 4-8 and 1 score in 8-12 which implies that 97.36% of the scores have assembled between 0 and 8. The median was 5 and the mean 4.9. In this session, the rate of success in the BAC English test was 00%.

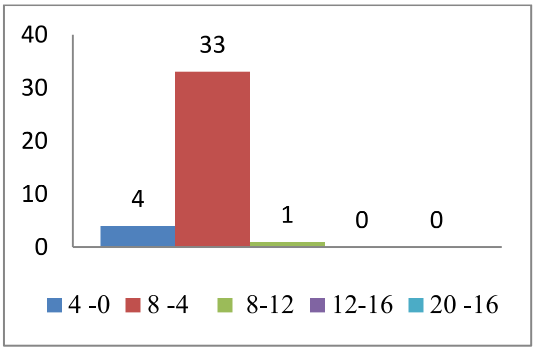
Table 30: Frequency Distribution of 2002 BAC English Test Scores..

Test Scores (X)	Frequency (f)
8.5	1
7.5	1
7	1
6.5	2
6	1
5.5	**9**
5	8
4.5	6
4	5
3.5	1
3	2
0.5	1

Table 31: Grouped Score Frequency Distribution of the 2002 Session

Class interval	Midpoints	Frequency
0-4	2	4
4-8	6	33
8-12	10	1
12-16	14	0
16-20	18	0

Graph 8: Histogram of 2002 Grouped Score Frequency Distribution



In 2003, the highly reoccurred score was 5.5 with 6 frequencies. As for the condensation of the marks, we can see 5 scores in interval 0-4; 27 scores in 4-8; 3 in 8-12: and 1 score in interval 12-16. The median of the obtained scores was 5, and the mean was 4.9. In this session, the rate of success was 8.33%.

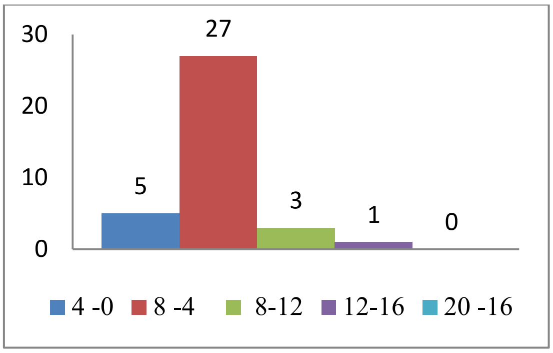
Table 32: Frequency Distribution of 2003 BAC English Test Scores.

Test Scores (X)	Frequency (f)
12	1
10.5	2
9	1
7	4
6.5	2
6	5
5.5	**6**
5	2
4.5	3
4	5
3.5	2
3	2

Table 33: Grouped Score Frequency Distribution of the 2003 Session

Class interval	Midpoints	Frequency
0-4	2	5
4-8	6	27
8-12	10	3
12-16	14	1
16-20	18	0

Graph 9: Histogram of 2003 Grouped Score Frequency Distribution



In 2004, the mode was 6 which reoccurred for 6 times. The scores have assembled as follow: 5 in interval 0-4; 20 in 4-8; and 1 in interval 8-12. The median was 5.5; and the mean was 5.2. Once again in this session, the rate of success was 00%.

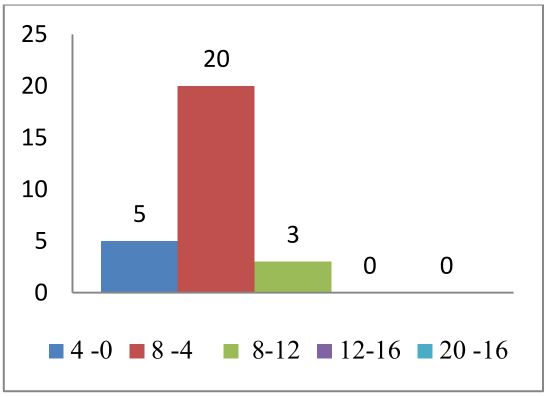
Table 34: Frequency Distribution of 2004 BAC English Test Scores

Test Scores (X)	Frequency (f)
9.5	1
8	2
7.5	1
6.5	4
6	**5**
5.5	4
5	2
4.5	2
4	2
2.5	3
1.5	1
0.5	1

Table 35: Grouped Score Frequency Distribution of the 2004 Session

Class interval	Midpoints	Frequency
0-4	2	5
4-8	6	20
8-12	10	3
12-16	14	0
16-20	18	0

Graph 10: Histogram of 2004 Grouped Score Frequency Distribution



In 2005, Table 36 implies that we are in the context of a multi-modal situation, in that each of the following scores: 8, 7, 5, and 4 has reoccurred for three times. Concerning the score grouping, interval 0-4 includes 10 frequencies; and in 4-8, we can count 19 scores. This means that 90.6% of the scores fall in interval 0-8. The median was 5; and the mean was 5.3. Once again in this session, all the scores of the BAC English test were below average.

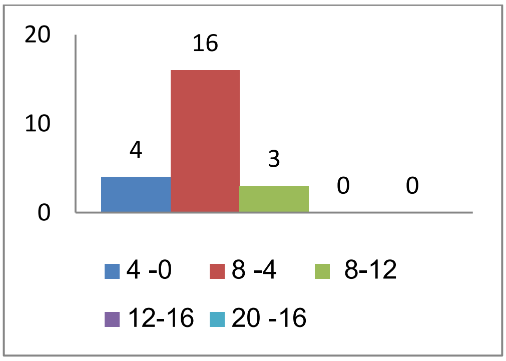
Table 36: Frequency Distribution of 2005 BAC English Test Scores.

Test Scores (X)	Frequency (f)
8	**3**
7.5	1
7	**3**
6.5	2
5.5	2
5	**3**
4.5	2
4	**3**
3.5	1
3	1
2	1
1.5	1

Table 37: Grouped Score Frequency Distribution Of The 2005 Session

Class interval	Midpoints	Frequency
0-4	2	4
4-8	6	16
8-12	10	3
12-16	14	0
16-20	18	0

Graph 11: Histogram of 2005 Grouped Score Frequency Distribution



In 2006, the most frequently occurring score was 4.5 with nine frequencies. As for the grouping of scores, it can be arranged as follows: 10 scores fall in interval 0-4: 19 marks in interval 4-8; and 3 marks in 8-12. The median was 4.5; and the mean was 5.5. The rate of success was 00%.

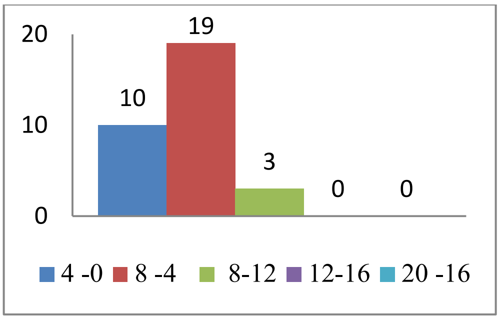
Table 38: The Frequency Distribution of 2006 BAC English Test Scores

Test Scores (X)	Frequency (f)
9.5	1
8.5	2
7	1
6	2
5.5	2
5	4
4.5	**9**
4	1
3.5	4
3	4
2.5	1
2	1

Table 39: Grouped Score Frequency Distribution of the 2006 Session

Class interval	Midpoints	Frequency
0-4	2	10
4-8	6	19
8-12	10	3
12-16	14	0
16-20	18	0

Graph 12: Histogram of 2006 Grouped Score Frequency Distribution



As we have mentioned previously, the data concerning technology pupils' results at the level of the 'wilaya' of Eloued were not available to us in detailed forms. In the same way, this information enabled us to have an overall view concerning the number of the pupils who succeeded in the BAC English test and those who did not (see appendix A).

In 2001, for example, out of 395 pupils, 129 got marks equal or above average with a rate of success of 32.65%. In 2002, the whole number of pupils in the 'wilaya' failed to get a pass mark in this test. In 2003, out of 129 test takers only 13 succeeded in this test forming a rate of 10.07% of the whole number of the pupils. Once again in 2004, the rate of success was 00%. In the same way, in 2005, no test taker was able to attain a pass score in English. In 2006, out of 329 pupils only 21 were able to get a score equal or above 10.

6.2.5.3. The Claim: Score Interpretations, Uses and Consequences

Language testers and measurement institutions emphasize that score validation should address three main criteria: score meaning, uses of the scores and intended and adverse consequences which may affect test takers whether in the short or in the long term ([AERA], [APA], & [NCME], 1999; Bachman, 1990, 2005, 2012, Kane, 2013; Messick, 1989, 1996; McNamara, 1996, 2006; Miller, Linn & Gronlund, 2009).

6.2.5.3.1. Score Interpretation

The scores obtained by technology streams in seven BAC sessions suggest that apart from June 2001 session, these pupils have low level of language ability, which does not allow them to use this language whether in real target domains, or for pursuing further studies where English is the leading language.

6.2.5.3.2. Uses of Scores

According to Bachman (2005): "the fundamental use of language tests is to make decisions" (p. 5). Additionally, in large scale assessment, a single test score can be used to determine the future of test takers whether in their academic or occupational life (Bachman & Purpura, 2008; Davies, 2008; Shohamy, 2008).In Algeria, the scores obtained in the BAC exam are used for making inferences about test takers language abilities, and for certification, placement, selection, or prognostic decisions (Ministry of Education, 1998, 2000, 2004). The uses of these scores will certainly have consequences on the stakeholders. If the test is used for the purpose it was designed for, the decisions will yield intended (beneficial) consequences; otherwise, we can speak of adverse or unintended consequences. At this point of the research, we cannot say that the consequences affecting students or teachers, are intended or unintended unless we support these findings with an empirical study by means of data and evidence collection (the questionnaire, the interview and documentary sources).

6.2.5.3.3. Consequences Affecting the Pupils

Low scores in English can lead to decreasing the rate of success in the BAC exam as a whole. This of course can have other consequences on test takers such as denying them certification, limiting their opportunities to join higher education institutions, or English language departments, minimizing their chances for occupational positions; or even expulsion from formal education.

6.2.5.3.4. Consequences Affecting the Teaching Staff

Low scores can affect teachers in different ways, for instance, their "self-esteem, reputation, and even career progression may be affected" (Wall, 2012, p. 79). In El-Oued, the Orientation Centre publishes a yearly evaluation record measuring teachers' contribution to the improvement of test takers' level of language ability (2001-2006). This document often specifies the outcome of the teachers of English in technology streams in the BAC English test as of 00 % (Orientation Centre of Eloued, 2001-2006).

6.2.5.4. Warrant and Backings

As we have indicated previously, these components of Toulmin's validity argument (warrants and backings) tend to legitimize the chain of inferences that we intend to make from the datum to the claim. As far as this research is concerned, the warrant refers to the data that we have gathered by means of the questionnaire and the interview about the extent of scoring consistencies. This information will be reinforced by the Backing (rater expertise and mediation methods) which tends to examine whether "the judgements or scores [are] reliable and…[whether] their properties and relationships [are] generalizable across the contents and contexts of use" (Messick, 1996, p. 246).

6.2.5.5. The Rebuttal

The rebuttal in this argument refers to the information or evidence that we have collected by means of documentary sources about construct representation, construct irrelevant variances, criterion relatedness, content relevance and domain coverage. (Bachman, 1990; Bachman and Palmer, 1996 Messick, 1989). This will enable us to answer Messick's (1996) question "What evidence is there that our scores mean what we interpret them to mean?" (Messick, 1996, 247). In other words, if the collected evidence supports the information included in the claim, the score interpretations will be considered valid. If the collected evidence rebuts the information in the claim, this can invalidate the score interpretations and the purposes for which they have been used.

6.3 Analysis of the Information Gathered by Means of the Questionnaire

6.3.1. Description of the Questionnaire

Goode and Hatt (1952) define the questionnaire as "a device for securing answers to questions by using a form which the respondent fills in himself" (p.137). Singh (2006) explains that this device consists of factual questions which are "designed for securing information about certain conditions or practices, of which recipient is presumed to have knowledge"(p. 191).

6.3.2. Structure of the Questionnaire

This questionnaire consists of twenty-nine (29) highly structured items composed of multiple-choice and dichotomous questions. These items are organized into eight sections each of which highlights a given aspect of the rating process in the BAC English test such as raters' appointment, rater training, inter-rater and intra rater reliability, the rating procedures, rating scales, methods for solving raters' discrepancies, future perspectives for the incorporation of automated scoring as well as test tryout.

There are some reasons which led us to focus on closed items. The first of these is related to the number of respondents themselves; or to what Cohen, Manion and Morrison (2007) refer to as the 'simple rule of thumb' which states that "the larger the size of the sample, the more structured, closed and numerical the questionnaire may have to be, and the smaller the size of the sample, the less structured, more open and word-based the questionnaire may be"(p.320). Additionally, we know that the respondents who would assemble for rating test takers' BAC English tests might not have enough time to respond to open-ended questions because of their concentration on scoring rather than on responding to questions. More importantly, structured questions allow comparisons to be made across groups of raters and ensure a high proportion of questionnaires to be returned.

In brief, closed questions "are quick to complete and straightforward to code (e.g. for computer analysis), and do not discriminate unduly on the basis of how articulate respondents are" (p. 32 [parentheses in original]).

The main aim of this questionnaire is to verify hypothesis one which assumes that the process of scoring the BAC English tests may not be reliable. Equally important, seeing that the scoring practices are almost the same, the data which we have collected by means of this tool will not be limited to verifying the scoring practices during one specific rating session, but it aims to examine the extent of rater reliability in the BAC exam rating centers as a whole.

6.3.3. Piloting the Questionnaire

Questionnaire piloting refers to the small-scale of trials that researchers administer to a representative sample of the target population before the main investigation is conducted (Blaxter, Hughes & Tight 2006; Cohen, et al., 2007). The main purpose of this process is to "assess the adequacy of the research design and of the instruments to be used for data collection [and]…to devise a set of codes or response categories for each question" (Wilson & Sapsford, 2006, p. 103). The drafts of the questionnaire, which were piloted in fifteen (15) secondary schools in the 'wilaya' of Eloued, were administered to 35 teachers of different levels of expertise in scoring. At the same time, we were committed to ensure an equal representation of both genders. Piloting allowed us to check the validity and the practicability of the questions and to check the time taken to complete the questionnaire. Moreover, it enabled us to gain feedback about the clarity, readability, and order of items and sections.

6.3.4. Population and Sampling

The respondents who are composed of secondary school teachers appointed by the educational authorities to participate in rating June 2013 BAC English test session in 'Guémar Technical school' in the 'wilaya of Eloued' include, according to the chief examiner in the same center, sixty-three raters (63): thirty-three (33) females and thirty (30) male raters. Most of these respondents participated in the sessions that had been held from 2001 to 2006. In order to ensure a high level of validity, the questionnaire was administered to the whole number of respondents.

6.3.5. Administration of the Questionnaire

Research methodologists identify two types of questionnaires: mailed and self-administered questionnaires. The first type is sent by post or emailed to respondents and the second type can be administered by the researcher himself, or a by a person(s) who represent(s) him. We can also speak of group questionnaires which can "be administered to groups of people who have gathered together for any purpose" (Goode & Hatt 1952, p. 170). Because of the security measures which limit the access of outsiders to large scale rating centers, this questionnaire was not administered by the researcher himself. Instead, it was administered by one member of the rating team who volunteered to do so. What is worth mentioning here is that we had several daily debriefings with a large number of respondents after the working hours to discuss different points in the questionnaire. However, as it has been planned, the questionnaire was returned in four weeks' time; the period in which the rating process was drawing to its end. As Table 40 implies, out of sixty-three raters, forty-eight of them returned the questionnaire (26 females and 22 male raters).

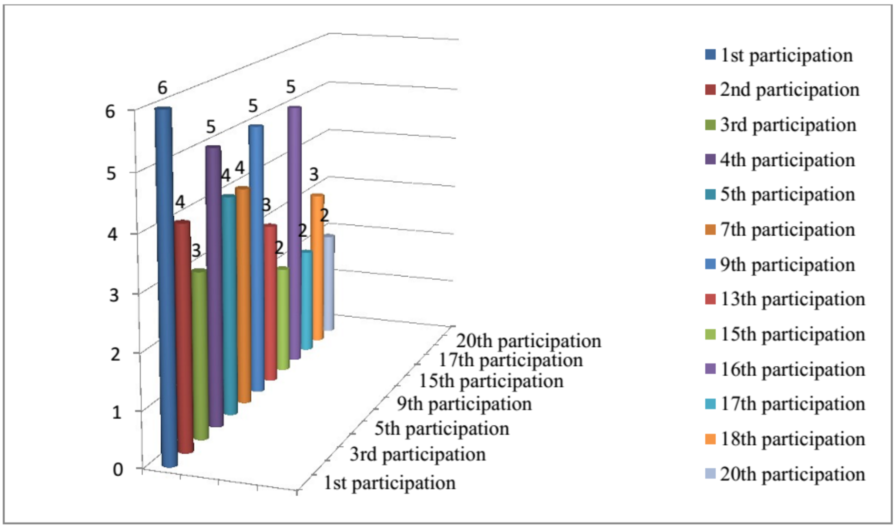
Table 40: Proportion of Questionnaire Returns

	Gender			
	Males	Females	Total number	Percentage
Number of respondents	30	33	63	100%
Questionnaire returns	22	26	48	76%
The subjects who did not return the questionnaire	08	07	15	24%

6.3.6. Respondents' Level of Expertise in Scoring

As Graph 13 indicates, the level of expertise in rating the BAC English test varies between two extremities. There are respondents who have participated in scoring this type of tests for 20 times, and there are others whose participation in the 2013 session was the first. We have, for example, thirteen (13) respondents whose participations range from one (01) to three (03) sessions; and other thirteen (13) ones who have participated from four to seven sessions. Additionally, there are ten respondents whose expertise extends from nine to fifteen rating sessions; and finally, there are twelve who rank at the top of the list with an expertise ranging from sixteen to twenty sessions.

Graph 13: Expertise in Rating



Additionally, as Table 41 indicates, twenty-two raters participated in the sessions from 2001 to 2006; and four respondents participated in 2002, 2004 and 2006 sessions.

Table 41: Raters Participating from 2001 to 2006 Scoring Sessions

Number of Raters	Number of participations	Participated in					
		2001	2002	2003	2004	2005	2006
02	20	×	×	×	×	×	×
03	18	×	×	×	×	×	×
02	17	×	×	×	×	×	×
05	16	×	×	×	×	×	×
02	15	×	×	×	×	×	×
03	13	×	×	×	×	×	×
05	09	×	×	×	×	×	×
04	07		×		×		×
04	05						
05	04						
03	03						
04	02						
06	01						

6.3.7. Ethical Issues

Due to the fact that items in questionnaires can represent "an intrusion into the life of the respondent, be it in terms of time taken to complete the instrument, the level of threat or sensitivity of the questions, or the possible invasion of privacy" (Cohen, al, 2007, p. 317), we ensured the respondents that the information we were seeking to gather would exclusively be used for research purposes. At the same time, we guaranteed the anonymity of subjects and the confidentiality of the information they provided. Besides, we attempted, as far possible, to avoid any type of offensive or intrusive questions.

6.3.8. Data Analysis

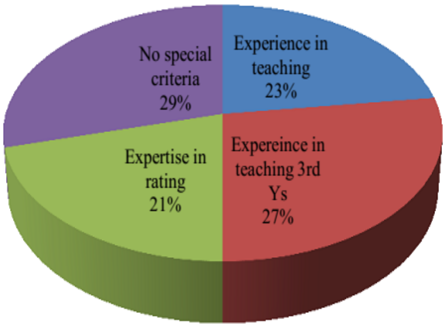
6.3.8.1. Section 1. The Qualities of Raters

Item 1. In your point of view, on what criteria do the educational authorities appoint teachers for the rating process?

Table 42: The Educational Authorities' Criteria for the Appointment of Raters

Their experience in teaching	11
Their experience in teaching the third year level	13
Their expertise in rating	10
There are no requirements in the appointment of raters	14

Graph 14: The Educational Authorities' Criteria for the Appointment of Raters



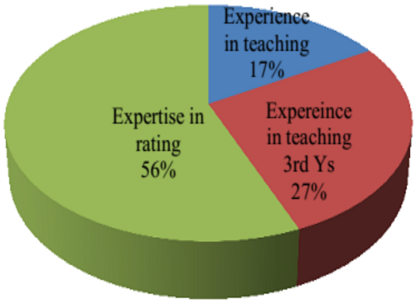
There was too much divergence in the points of view of respondents concerning the criteria upon which the educational authorities appoint teachers for the scoring process. 29% of them think that this issue is arbitrary and not built upon any specific criteria. However, 27% link it to experience in teaching examination levels; and 23% of them see that raters are chosen because of their experience in the field of teaching in general. Surprisingly, only 21% of the subjects relate the choice of raters to the extent of their expertise in scoring.

Item 2. Suppose that you are responsible for the selection of raters, on what criteria will you base your choice?

Table 43: Respondents' Criteria for Raters' Selection

Experience in teaching	08
Experience in teaching the third year level	13
Expertise in rating	27
Other factors	00

Graph 15: Respondents' Criteria for Raters' Selection



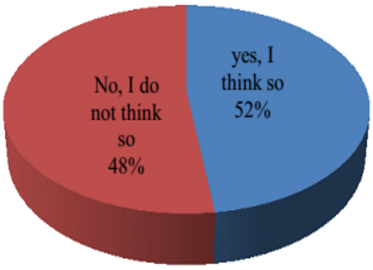
In item one, we wanted to know the respondents' perceptions of the decisions made by the 'academies' concerning the selection of raters and whether these decisions stand on logical grounds, such as rating expertise or positive record in previous scoring sessions. Item two seeks to see the judges' opinions about the standards which should normally be taken into consideration during the choice of scorers. 56% of respondents think that expertise in rating should rank at the top of list; while 27% consider experience in teaching examination levels as the first criterion; however a minority of 17% relate the choice to expertise in teaching.

Item 3. Do you think that raters' educational or cultural background can affect their scoring behavior?

Table 44: The Impact of Raters' Background on the Scoring Behavior

Yes I think so	25
No, I do not think so	23

Graph 16: The Impact of Raters' Background on the Scoring Behavior



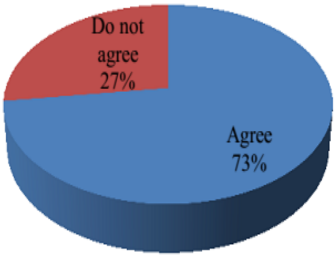
Responses to item three manifest great divergence between the views of the subjects on whether the educational or cultural background of raters can affect the consistency of their scoring. A slim majority of 52% see that this factor can influence the consistency of their rating against 48% who think that this feature does not have any impact on the scores they assign to test takers.

Item 4. Do you think that raters' judgment in general can bear elements of subjectivity?

Table 45: Respondents' Views on Raters' Subjective Judgments.

agree	34
Do not agree	14

Graph 17: Respondents' Views on Raters' Subjective Judgments.



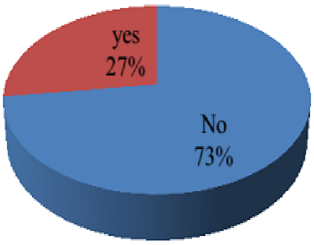
In response to item four, 73% of respondents think that raters' judgments can bear elements of subjectivity; whereas 27% of them do not share the same point of view. This means that the authorities responsible for the scoring process need to take this issue into consideration; especially by providing the means for mediating any probable rater inconsistencies.

Item 5. According to you, do experienced and novice raters employ the same scoring strategies?

Table 46: Scoring Strategies of Expert and Novice Raters

No, they do not	37
Yes they do	11

Graph 18: Scoring Strategies Employed by Expert and Novice Raters

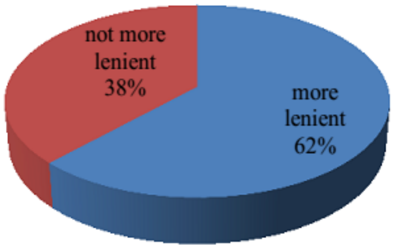


- If no, are novice raters significantly more lenient in their judgment than expert raters?

Table 47: Respondents' Views Regarding Scorers' Leniency

More lenient	23
Not more lenient	14

Graph 19: Respondents' Views Regarding Scorers' Leniency



77% of the subjects think that expert and novice raters employ different strategies during their scoring; while 23% of them do not share the same point of view. Now, twenty-three (23) subjects out of the thirty-seven (37) who think that raters do not use similar strategies informed us that differences in rating are caused by leniency on the part of novice raters.

6.3.8.2. Section Two: The Rating Process

Item 6. Operational scoring starts…………

Table 48: The Beginning of Live Scoring

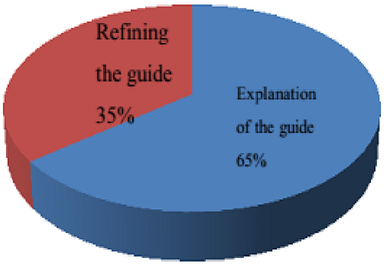
As soon as raters meet	00
In the second session of the first day	00
On the second day	48

- If operational scoring is delayed to the second session or to the second day, what is the first session devoted to?

Table 49: Works in the First Session

Explanation and analysis of the scoring guide	31
Refining the scoring guide	17
Drafting a new scoring guide	00

Graph 20: Works in the First Session



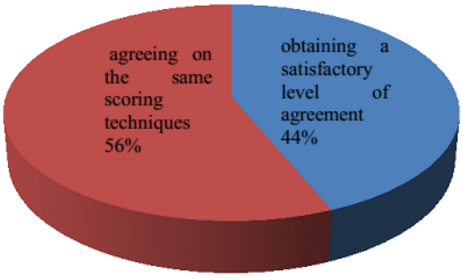
The whole number of respondents answered that operational scoring starts on the second day of their meeting. When we wanted to know what the works on the first day are devoted to, 65% said that the first meeting focuses on the explanation and analysis of the scoring guide while 35% them consider the works to focus on the refinement of the guide, but none of them told us that the discussion results in drafting a new guide.

Item 7. Discussion in the first session aims at……

Table 50: Purpose of Discussion in the First Meeting

Obtaining a satisfactory level of agreement	21
Agreeing on the same scoring techniques	27
Other purposes	00

Graph 21: Purpose of Discussion in the first Meeting



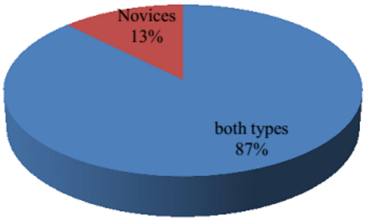
According to 56 % of the respondents, the purpose of the discussion that raters engage in on the first day enables them to agree on the same scoring procedures ; while 44 % think that this allows them to obtain a satisfactory level of agreement. Both opinions imply that the chief examiners do not allow live scoring to start unless raters come to consensus about the directions included in the guide.

Item 8. In your point of view, the scoring guide is indispensible to….

Table 51: The Type of Raters that Mostly Need the Scoring Guide

Novice raters	06
Expert raters	00
Both types	42

Graph 22: The Type of Raters that Mostly Need the Scoring Guide



When asked whether the scoring guide falls in the advantage of novices, experts or of both types, the respondents' answers came as follows: 87% of them think that it is useful for both types of raters; conversely only 13% limit its efficacy to novices. Responses to this item imply that the use of scoring guides in the rating process should not be related to the degree of raters' expertise.

Item 9. In the pre-scoring session, sample scripts are…………

Table 52: Pre-Scoring of Sample Scripts

blindly single-rated by the chief examiner	00
blindly double-scored by pairs of raters	00
scored collectively by all the participants	48

This item attempts to see how the sample scripts are corrected in the standardization session. This is because it is in this session that raters learn how to comply with the guide and how to stay in close agreement with one another. Additionally, training in this session can help them overcome the difficulties that they may encounter during live rating. So, when we wanted to know whether the sample scripts are scored by the chief examiner; blindly double-scored by pairs of raters; or scored collectively by all the participants, the whole number of respondent told us that these papers are corrected collectively during a general session.

Item 10. In the pre-scoring session, the sample papers represent the ……….

Table 53: The Type of Scripts Chosen for the Pre-scoring Process

problematic scripts	00
consensus scripts	00
randomly-chosen scripts	48

The purpose of this question is to see how the sample scripts are chosen for the pre-scoring session, because if one limits his/her training to one type of scripts, problems may rise from the other types. Let us speak, for example, about the problematic scripts which fall into three types: off-task scripts, memorized scripts, and incomplete tasks (see pp. 120-121). Now, if the scripts are randomly chosen, we may come up with one type of responses and miss the opportunity of training raters on the other types. So, in order "to anticipate as far as possible the kinds of problems that might occur with a given prompt, [and] to reduce the possibility that different raters will approach problematic scripts differently and thus introduce unwanted errors into the scoring procedures" (Weigle, 2002, pp. 131-132), chief examiners need to train raters by means of the four types of scripts.

Item 11. Once live scoring is under way, do you discuss with table leaders or the chief examiner the difficulties that might encounter you during your correction of test takers' papers?

Table 54: Communication between Raters and Table Leaders.

Certainly	48
Not necessarily	00

Responses to this item suggest that the role of the chief examiner or table leaders is not limited to the standardization session, but it extends to helping raters overcome the difficulties that they may encounter during the whole process of rating.

6.3.8.3. Section Three: Rater Training

Item 12. Have you attended a seminar, a colloquium, or a meeting about rating?

Table 55: Respondents' Participations in Rater Training Gatherings

Yes, I have	00
No, I have not	48

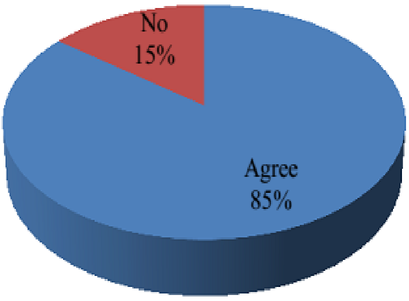
Language testers emphasize that the role of training in reinforcing the consistency of scoring within and across raters (intra-rater and inter-rater reliability) is of great importance. Additionally, these testers point out that reliable scoring is unlikely to be assigned by unqualified raters (McNamara & Roever, 1996; Weigle, 2002). Surprisingly, the whole number of respondents, whatever the extent of their expertise was, told us that they have been introduced to the rating process without any type of training. This reminds us of Spolsky (1979) who comments on the rating practices in the pre-scientific stage where "no special expertise is required, if a person knows how to teach, it is to be assumed that he can judge the proficiency of his students" (p. 7).

Item 13. Do you think that introducing raters to the assessment without any type of training can affect the consistency of their scoring?

Table 56: The Incorporation of Untrained Raters into the Scoring Process.

agree	41
do not agree	07

Graph 23: The Incorporation of Untrained Raters into the Scoring Process.

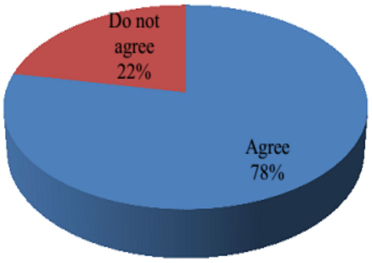


- If so, training sessions can determine whether a rater will participate satisfactorily in the scoring process?

Table 57: The Role of Training Sessions in the Improvement of Raters' Behavior

Agree	32
Do not agree	09

Graph 24: The Role of Training Sessions in the Improvement of Raters' Behavior



Responses to this item have come to reinforce language testers' conclusions about the importance of training for reliable scoring in that 85% of them think that the lack of training can affect the consistency of their scoring; against 15% who think that this issue does not affect the credibility of their ratings. Now, out of the forty-one subjects who believe in the efficacy of training, 32 respondents think that this practice helps identify the raters who can participate satisfactorily in the scoring process from those who may show significant variations. Furthermore, in our point of view this process enables the educational authorities to invite the discrepant raters for additional training sessions before their participation in live scoring.

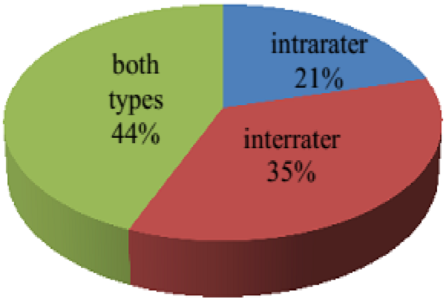
6.3.8.4. Section Four: Rater Reliability

Item 14. According to you, rater consistency can be understood of ....

Table 58: Respondents' Conception of Rater Consistency

intra-rater reliability	10
inter-rater reliability	17
both types of reliability	21

Graph 25: Respondents' Conception of Rater Consistency



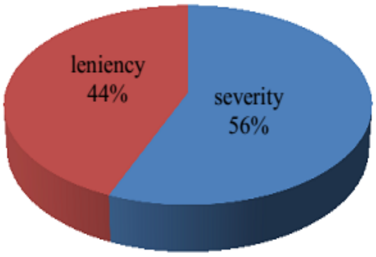
Responses to this item enabled us to see how raters conceive the quality of reliability. 44% of respondents consider it to mean consistency across and within raters; 35% relate the concept to the consistency between raters; while 21% take it as a matter of stability within raters themselves. The answers of respondents suggest that the implementation of reliability yields consistent scores which can reflect the construct to be measured. Suppose for example that the collected evidence in an empirical study has come to validate a given test with respect to construct representation, content relevance and coverage as well as criterion relatedness, but the scoring of this test has been found to be unreliable which may affect the validity of interpretations. This is because "a test score that is not reliable, therefore, cannot be valid" (Bachman, 1990, p. 25)

Item 15. According to you, variability between raters could be understood in terms of………..

Table 59: Respondents' Conception of Rater Variability

severity	27
leniency	21

Graph 26: Respondents' Conception of Rater Variability



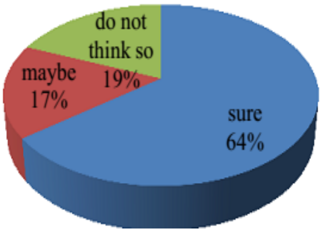
Answers to this item demonstrate great disparity between raters' opinions concerning the reasons that lead to variability in scoring. 56% of them think that wide discrepancies are caused by raters' severity; while 44% think that leniency is the cause of the problem.

Item 16. Can judges' severity or leniency be modified by training?

Table 60: The Role of Training in Modifying Raters' Behavior

Sure	31
Maybe	08
Do not think so	09

Graph 27: The Role of Training in Modifying Raters' Behavior



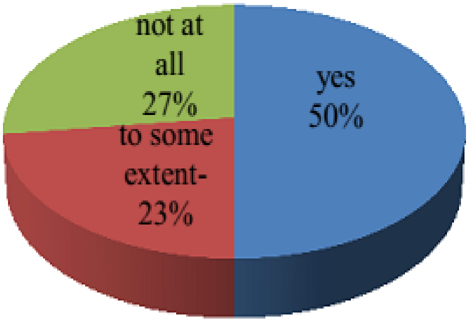
When we wanted to know whether training can help in narrowing the gap between severe and lenient raters, 64% of respondents expressed their certainty of this relationship. Additionally, although with a lesser extent, 17% of them think that training may result in modifying the scoring behavior. However 19% of the subjects disagreed completely with this idea. Now if we add 64% to 17% of respondents, this gives us 83% of the subjects who believe, to a certain extent, that training can contribute to the consistency of scoring.

Item 17. Can the consistency of your scoring be affected by the succession of the number papers that you are supposed to correct each day?

Table 61: The Impact of Script Sequencing on Intra-Rater Reliability

Yes	24
Yes, to some extent	11
No, not at all	13

Graph 28: The Impact of Script Sequencing on Intra-Rater Reliability



The main purpose of this item is to see whether intra-rater consistency can be affected by the number of papers that judges are required to score each day. 50% of respondents belief that the succession of ratings does affect the reliability of the scores they assign to test takers; in the same way but with a lesser degree of certainty, 27% of them share the same point of view; however a minority of 20% think that the sequencing of papers has no effect on the consistency of their ratings. Now, if we add 50% to 27% of raters we will have a percentage of 77% whose opinions coincide with those of language testers who emphasize that "in any rating situation, effects due to sequencing may introduce inconsistency into either the rating criteria themselves or the way in which they are applied " (Bachman, 1990, p. 179).

6.3.8.5. Section Five: Methods for Solving Raters' Discrepancies

Item 18. In the BAC exam, scripts are…

Table 62: Procedures of Script Rating

.

blindly single-rated	00
blindly double-rated	48

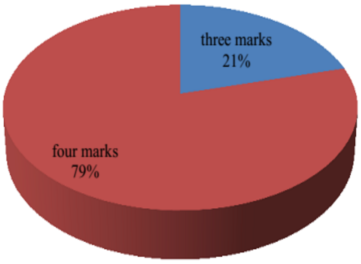
In response to the question whether scripts are blindly single-rated or double-rated, 100% of the answers have come to confirm that scoring takes place at two phases. During the first phase, a given number of raters correct the anonymous scripts; and in the second phase, the same scripts will be rated by different judges. In the BAC exam, blind double scoring is one of the most efficient methods of reinforcing intra-rater and inter rater consistency of the scores obtained on the basis of a single administration. This is because if one of the raters assigns inconsistent marks in the first phase; the discrepancy can be adjusted by the rater who will correct the same scripts in the second phase.

Item 19. How much tolerance for discrepancies between raters is allowed in the BAC exam?

Table 63: The Extent of Tolerance for Raters' Discrepancies

One mark	00
Two marks	00
Three marks	10
Four marks	38

Graph 29: The Extent of Tolerance for Rater Differences



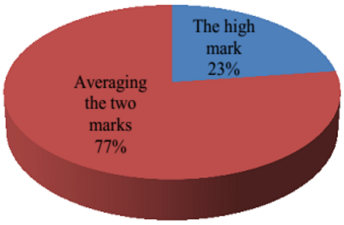
In item 18, we wanted to know the procedures of adjusting inconsistencies (within raters) due to the sequencing of correction. This item seeks to examine the extent of variability between raters which test designers consider as tolerated agreement. 79% of respondents told us that adjacent agreement can extend to four (04) marks, whereas 21% of them think that the disparity is limited to three (03) marks. The reason of this divergence between the points of view of the respondents is related to the previous participations of raters themselves in the scoring process. In other words, we have found that the 21% of the respondents who thought that the tolerated difference between raters is 3, was composed of novice raters.

Item 20. In the case of adjacent agreement, how will the final score be computed?

Table 64: Score Reporting in Adjacent Cases

We consider the high mark	11
The low and the high marks are averaged	37
Other solutions	00

Graph 30: Score Reporting in Adjacent Cases



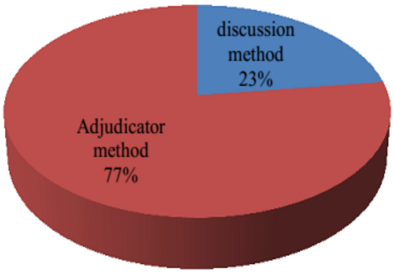
Measurement specialists point out that adjacent agreement should be settled by rater mean method which does not call for the involvement of a third rater (adjudicator). According to this method, the composite score is computed by combining and averaging the adjacent scores. When we asked the raters about this issue, 77% of them told us that the two scores are averaged; while 23% of them think that the high mark will be considered as the final mark.

Item 21. What happens in the case of disagreement between the first and the second raters?

Table 65: Settling Raters' Disagreement

The two raters discuss the issue and assign a consensus score	11
A third rater is brought in to resolve the discrepancy	37

Graph 31: Settling Raters' Disagreement

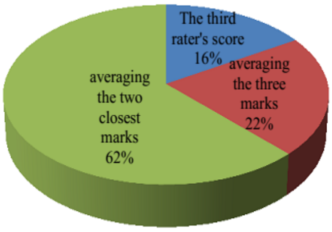


- If a third rater is brought in, how the final score will be computed

Table 66: Methods for Solving Raters' Discrepancies

Considering the expert score	06
Averaging the three scores	08
Averaging the two closest scores	23

Graph 32: Methods for Solving Rater Discrepancies



As we have pointed out in Chapter IV, measurement specialists identify two main methods for resolving rater discrepancies: discussion method and arbitration methods (parity/ tertium quid and expert methods). When we asked the respondents about the method which is usually used in the BAC rating centers, 77% of them informed us that an adjudicator will be brought in to settle the differences; while 23% of them think that the two original raters are invited to discuss the reason of their discrepancy and then agree on a consensus score.

Now the question was directed to the thirty-seven (37) respondents who think that discrepancies are resolved by means of mediation methods. 62% of them told us that the adjudicator's score is averaged with the closest mark (Tertium quid). 22% think that differences are solved by means of parity method which involves the combination of the original scores with the adjudicator's mark; and 16% think that the variability is settled by means of the expert method in which the judge's mark replaces the original ratings. The respondents' answers imply that inconsistent scores are always adjusted by one method or the other.

Item 22. Does the chief examiner communicate to discrepant raters the amount of variability which they might have done?

Table 67: Informing Raters of their Discrepancies

yes	00
No	48

The whole number of respondents informed us that the chief examiner did not communicate to them the amount of variability which they might have done. This means that resolving discrepancies by means of discussion method is not considered in the BAC exam rating centers. In our opinion, the identification of discrepant raters and the evaluation of their scoring records can serve two purposes: judges' rating behavior can be modified by training sessions; or by not considering these scorers for future rating sessions.

6.3.8.6. Section Six: Rating Scales

Item 23. Does the scoring guide include a rating scale?

Table 68: The Availability of Rating Scales in Subjective Scoring

A: Yes	00
B: No	48

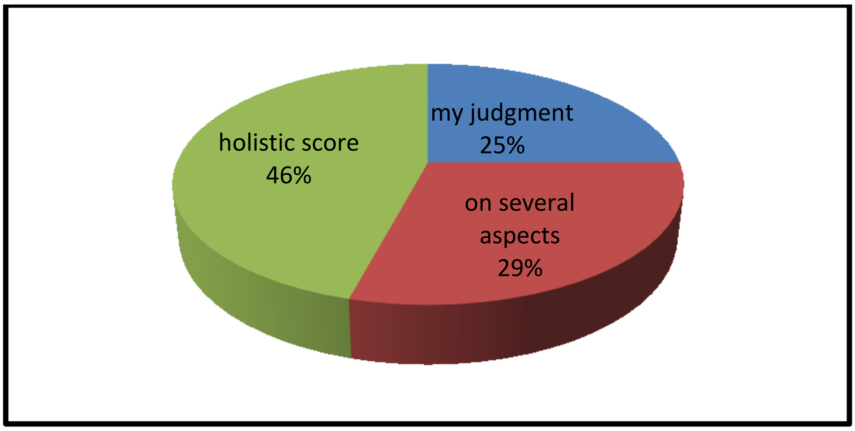
This item attempts to examine the criteria upon which raters measure test takers' written performance. The whole number of respondents informed us that the guide does not include such a scale. As we have seen in Chapter IV, language testers identify three types of rating scales: primary traits, holistic and analytic scales. Each one of these scales is used for a particular type of scoring. The respondents' answers imply that none of these scales is used to guide them in correcting written expression tasks. Language testers question the credibility of subjective scoring if scorers are not provided with rating scales because "there is no feasible way to 'objectify' the subjective procedures" (Bachman 1990, p. 76) unless a rating scale is used to guide them.

Item 24. In the lack of rating scales, how do you score the writing tasks?

Table 69: Techniques of Scoring Writing Tasks

Depend on my own judgment	12
Rate the script on several aspects	14
Read the script and assign a holistic score	22
Other techniques	00

Graph 33: Techniques of Scoring Writing Tasks



Language testers emphasize that scoring the 'writing tasks' should be guided by a given type of rating scales; and if raters do not use these tools, then, on what criteria, as Bachman (1990) asks, 'do they base their scoring?' Item (24) attempts to respond to this question, in that 46% of respondents told us that they read the script and assign a holistic score; 29% of them read the task and assign several scores; however 25% told us that they evaluate tasks according to their own judgment. In fact, not only do 25% of raters correct written performance according to their own perception; but the other two types do the same thing as well; since all of them do not use rating scales.

Item 25. If two raters assign the scores included in Table 70 to the same script, will their ratings be considered identical or variable?

Table 70: Raters' Correction of the Same Script

Exam Sections	Rater 1	Rater 2
Reading	06/08	05/08
Mastery of Language	05/08	02/08
Written Expression	00/04	04/04
Final Score	11/20	11/20

Table 71: Raters' Views Concerning Composite Scores

Identical	48
variable	00

In items 19, 20 and 21 respectively, we talked about the computation of scores in cases of adjacent agreement and discrepancies. This item attempts to identify the drawback of the mediation methods specifically on their focus on composite scores. So, when we asked our respondents to comment on the scores included in Table 70, all of them told us that such ratings will be considered identical. Identical scores are, of course, considered more reliable than adjacent scores since they do not call for rater mean method adjustment.

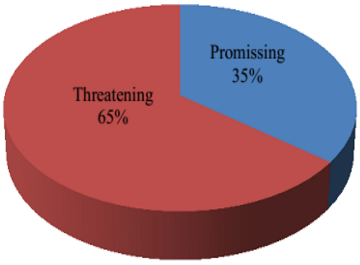
6.3.8.7. Section Seven: The Incorporation of Automated Scoring Systems

Item 26. What is your point of view on the incorporation of automated scoring in the BAC English tests?

Table 72: Raters' views on Automated Scoring

Promising	17
Threatening	31

Graph 34: Raters' views on Automated Scoring

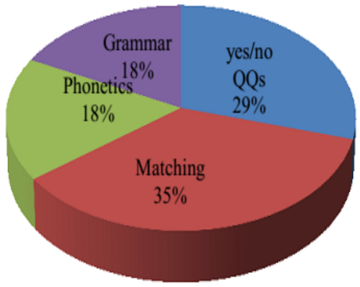


- If promising, which tasks can, in your opinion, better be scored by the computer?

Table 73: Raters' Views Concerning Computerized Scoring of Specific Tasks

Yes-no questions	05
Matching activities	06
Phonetics	03
Grammar	03
Other tasks	00

Graph 35: Raters' Views Concerning Computerized Scoring of Specific Tasks



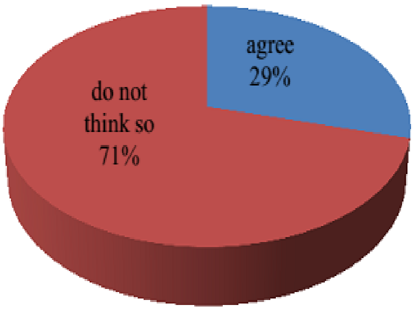
The purpose of this item is to see how raters' conceptualize the incorporation of machine scoring in the BAC English tests. 65% of the respondents think that its use is threatening, and 35% consider it promising. When we asked what tasks can better be scored automatically, the ones who believe in the efficacy of computer scoring responded as follows: 33% of them think that machines can rate matching activities; 28% think that these can score yes/no questions; 22% of the answers see that we can involve these devices in scoring phonetics; however in the point of view of 17% of respondents, computers can better be involved in correcting grammar.

Item 27. Do you think that computerized scoring can soon be operational in the BAC Exam?

Table 74: The Future Incorporation of Computerized Scoring in the BAC English Rating Centers.

Yes, I think so	14
I do not think so	34

Graph 36: The Future Incorporation of Computerized Scoring



This item attempts to see raters' views concerning the incorporation of computerized scoring in the BAC examination in the near future. 71% of respondents see that it is not possible to implement such technology, at least in the near future. Conversely, 29% of them consider automated scoring can soon be operational. In Chapter I, we have seen that the incorporation of automation in objective scoring dates back to the mid-thirties in the USA. If these practices are implemented in the BAC exam rating centers, ratings such as the ones provided in Table 65 may not occur.

6.4.8.8. Section Eight: Test tryout

Item 28. Has the Ministry of Education piloted a draft sample of the BAC English test in your school?

- If so, how often has that happened?

Table 75: Test tryout

Yes,	00
No	48

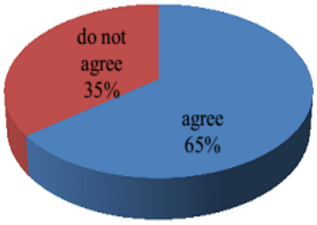
Despite the fact that this item may seem irrelevant and out of place in a questionnaire devoted to the scoring procedures in the BAC exam, our objective was to seize this opportunity to get some information about the BAC English test tryout and pretesting. So, in response to the question whether the ONEC has administered a draft sample of the BAC English test in their schools, the answers of respondents were all negative. In addition, the ONEC itself has confirmed that in test development, information concerning item difficulty, discrimination indices and test takers' levels of language ability is all provided by expert teachers (Echorouk Online, 2009). However, language testers emphasize that information concerning this issue can only be collected by means of test tryout.

Item 29. Do you think that test tryout can provide more efficient evidence on item difficultly and discrimination indices than the information provided by teachers' expertise?

Table 76: The Role of Test Tryout in Information Collection

Agree	31
Do not agree	17
Do not know	00

Graph 37: The Role of Test Tryout in Information Collection



In response to whether test tryout can provide more efficient evidence on item difficultly and discrimination indices than the information provided by teachers' expertise, 65% of the subjects' answers were positive, against 35% of their colleagues who still do not see the efficacy of item piloting in gathering useful information about test items. The opinion of the majority of respondents coincides with that of language testers who emphasize that the issue of 'item facility values' and 'discrimination features' cannot be obtained by expertise in test development but from item tryout ([AERA], [APA], & [NCME], 1999; Alderson et al., 1995; Livingston, 2006).

Discussion of Results

The data that we have collected by means of the questionnaire gave us an overall view on the scoring practices in the BAC exam rating centers. The first of these refers to the appointment of raters which is generally based on some type of expertise whether in teaching, teaching examination levels, or in previous participations in rating. This, of course, does not exclude the hypothesis that inexperienced raters are also invited to the rating process since this is the only opportunity provided to them to attend large scale scoring. Regarding rater training, our informants whether experienced or novices informed us that they have all been introduced to the assessment without any type of training.

Concerning operational scoring, it is always preceded by a standardization session to ensure a uniform interpretation of the scoring guide and of all the assessment practices. In this session, sample scripts are selected on random basis and corrected by all the raters. Differences amongst raters regarding the scoring of items are resolved by means of discussion methods. The purpose of training in the session is to anticipate any type of difficulties that may encounter raters during live scoring. Additionally, this is one of the procedures which can reinforce consistency within raters (intra-rater reliability).

Live scoring is organized into three phases. In phase one, all the scripts are blindly scored by individual raters. In phase two, the same scripts are re-rated by different judges. The scoring process in both phases is overseen by the chief examiner and room leaders. In case the correction results in adjacent scores, the clerical staff will settle the adjacencies by averaging the two marks; but if two raters assign discrepant scores, their variability will be settled in the third phase of scoring by one of the adjudication methods.

In the BAC English test, both types of scoring (objective and subjective) are implemented. In objective scoring, raters judge the correctness of items against predetermined criteria available in the scoring guide. However in subjective scoring, no type of scales is provided to raters to evaluate tasks such as the written expression section. In this case, raters have no choice but to evaluate these tasks according to their own judgments.

Concerning the implementation of automation in scoring, not only did we find the majority of respondents still suspicious about its use, but they consider it threatening as well. However, empirical researches estimating inter-rater reliability of equally trained and expert raters scoring the same product and using the same rating criteria, have demonstrated that human scorers have always fallen in one type or the other of variability. The issue of human raters' variability which has long been recognized, is illustrated by Edgeworth (1888, as cited in Bejar, Williamson & Mislevy , 2006) "let a number of equally competent critics independently assign a mark to the (script)...even supposing that the examiners have agreed beforehand as to ... the scale of excellence to be adopted, there will occur a certain divergence between the verdicts of competent examiners "(p. 51**)**.

In sum, despite the shortcomings that we have signaled above concerning raters and ratings, we found the scoring practices in the BAC English rating centers, to a large extent, consistent and reliable.

6.4. Analysis of the interview

The interview is one form of data gathering "in which a researcher and participant engage in a conversation focused on questions related to a research study.…Its main function is to provide a framework in which respondents can express their own thoughts in their own words" (Tavakoli, 2012, 294). During this procedure, the interviewer attempts to elicit information from (an)other person(s), the interviewee(s). The interview can be conducted in a face-to-face way (personal interview); by means of telephone (telephone interview) or other technology programs such as the Skype, the Paltalk, or the Facebook . Unlike the questionnaire which is based on predetermined and more structured items, the interview can include both structured and unstructured items. Additionally, much more flexibility can be allowed in the wording and type of questions.

Due to the fact that the majority of raters do not attend the mediation stage of scoring; in addition, their responsibility in this process is limited to scoring individual scripts, we felt the need to supplement data from the chief examiner who oversees this process from its initial until its final phases. Furthermore, this procedure will enable us to compare and contrast the responses of the interviewee with the ones that we have previously collected by means of the questionnaire.

6.4.1. Structure of the Interview

This interview consists of 49 questions: 27 open-ended and 22 closed questions. The items which attempted to cover all the aspects that we have previously raised in the questionnaire include standardization meetings, the division of raters into groups, the appointment of team leaders, live scoring, the procedures implemented to monitor raters' discrepancies, post scoring procedures such as the analysis of students marks as well as the chief examiner's opinion on the implementation of automated scoring in the BAC exam.

6.4.2. Description of the Interviewee

Our informant is the chief examiner of the rating committee in charge of correcting the BAC English tests in 2013 in Eloued Rating Center. His expertise extends for more than twenty participations as a rater; and five times as a chief examiner during 2007, 2008, 2009, 2010; and 2013 BAC sessions.

6.4.3. Analysis of the Interviewee's Responses

6.4.3.1. Quality of Raters

The introductory items in the interview attempted to examine the number, gender, level of expertise of raters and the criteria upon which they are selected for the scoring process. Our respondent informed us that the number of raters in this session (2013) has reached (63) raters: thirty-three (33) females and thirty (30) male raters exclusively selected by the 'Directions de l'Education'. Their appointment is, according to him, arbitrary and does not stand on any specific criteria. In addition, the interviewee accentuated the impact of expertise and believes that it should form at least two thirds of the whole number of raters.

6.4.3.2. The Standardization Session

In the same way as the responses in the questionnaire, the chief examiner informed us that the works on the first day were fully devoted to the explanation of the scoring guide so that it could be interpreted and implemented uniformly. In the standardization session, sample scripts are randomly taken from the batches, scored in an open session; if the scoring manifests differences amongst the judgments of raters; the latter will discuss the issue until they agree on a consensus score.

6.4.3.3. Live Scoring

As soon as live scoring is underway, raters are split, according to their level of expertise, into a number teams. The room or team leaders are selected out of the most expert raters. Their role is to ensure standard interpretations of the scoring guide; to protect the security of scoring, to ascertain that raters do not communicate information relevant to the scripts they are correcting with their colleagues; and to help novice raters in their application of the rating criteria.

6.4.3.4. Types of Scoring

Our informant told us that the type of scoring depends on the characteristics of the activities themselves in that some items call for objective scoring; while the correction of the writing tasks, or constructed responses calls for subjective scoring. Seeing that subjective scoring requires the use of rating scales, when we wanted to know the criteria upon which raters evaluate 'written expression' tasks without using these instruments; the chief examiner informed us that the judges read the tasks and assign one composite score according to their own judgments.

6.4.3.5. Adjusting Raters’ Variability

According to the chief examiner, there are two methods for adjusting raters' differences. If two raters assign adjacent scores to the same script (adjacencies can be four points apart), the difference can be settled by 'rater mean' method. That is, the two scores will be combined and averaged. However, if two raters assign discrepant scores to the same paper; the discrepancy will be resolved by one option of the 'Tertium Quid' method which recommends that the adjudicator's score needs to be combined and averaged with the closest of the original scores. However, when we wanted to know on what grounds adjudicators (raters of more expertise than their colleagues) are selected for the mediation phase of scoring (la troisième correction), the chief examiner told us that these are appointed simply because they live in the vicinity of rating centers.

6.4.3.6. Evaluation of Discrepant Raters' Records

Our informant told us that the results of scoring in 2013 manifested around 160 discrepant ratings of the same scripts. Despite the fact that the discrepant raters can be identified, the ONEC seems not to be interested in documenting the number of their incongruent scores; nor in evaluating their scoring record. The evaluation of raters' discrepant records, enables test users either to invite these raters for additional training sessions; or simply not to consider their participations in future rating sessions.

6.4.3.7. Post Scoring Procedures

Questions in this section attempted to see whether the process of scoring is concluded with adjudication and score reporting, or whether it extends to the analysis of the scores obtained by test takers; or to the evaluation of raters' discrepant records in scoring (post scoring gatherings). Our respondent told us that he has never been invited to such meetings; and as far as he knows, none of these procedures is implemented in the BAC exam. Language testers emphasize that the analysis of test takers' marks enables test designers and users to examine the extent to which the interpretation of the scores are reliable and valid.

6.4.3.8. Automated Scoring

The last section concerns the chief examiners' points of view regarding the implementation of automated scoring. The responsible considers this issue threatening because it requires more outsiders (computer technicians) to be involved into the field. Moreover, the incorporation of machines in correction can slow down the rating process in that instead of human rating; we will be faced with two types of scorers; human raters and automated correction. For this reason, the interviewee did not think that the use of computers in scoring would be workable, at least in the near future.

Discussion of Results

As we have mentioned in the introduction, the main purpose of this interview is to match its information with the data that we have formerly collected by means of the questionnaire. The analysis of this information enabled us to examine the rating process and procedures from the point of view of the chief examiner who has overseen the rating process in this session starting from the pre-rating until the adjudication phase. This analysis led us to reinforce the conclusion that we have drawn concerning the consistency of scoring in the BAC English tests. This implies that reliability of scoring in these large scale tests is implemented by the incorporations of different procedures such as rater expertise, standardization training sessions, blind double correction, rater mean adjustments, and arbitration methods.

Briefly speaking, despite the shortcomings that we have identified during the analysis of the questionnaire, such as the disparity between the techniques implemented in the training session (resolving discrepancies by means of discussion method) and the ones incorporated in live scoring (adjudication methods), or the lack of training and assessment literacy on the part of raters, the information that we have collected by means of the interview has come to reinforce the conclusion that the process of scoring in the BAC exam is, to a large extent, consistent and reliable. This means that the interpretations and uses provided for technology pupils' scores from 2001 to 2006 are supported with the information gathered by the questionnaire (the warrant), and reinforced by the data provided by our interviewee (the backing).

6.5. Evidence Collection and Analysis

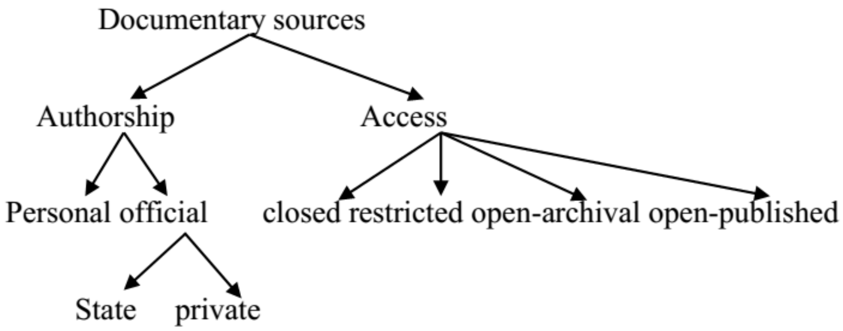
In scientific research, interviews, questionnaires, observations and experiments or tests are considered as the main tools for data gathering (Cohen et al., 2007; Goode & Hutt, 1952; Lee McKay, 2008). Nonetheless, these instruments do not always provide us with all the information that we need to test the hypotheses. This is why in certain cases of study; we feel the need to supplement data from other existing documentary “sources whether in writing, figures or electronic form” (Finnegan, 2006, p. 139). Seeing that validity is "***inferred*** from available evidence (not measured)" (Gronlund, 1977, p. 132 [italics, emphasis and parentheses in original]), our interest in this section is to gather evidence so as to support or challenge the interpretations and uses of test scores emerging from technology pupils' results in Eloued during seven BAC English sessions (2001-2006).

6.5.1. Typology of Documentary Sources

Documentary sources can, as illustrated in Fig 37, be organized according to two main categories: 'authorship' and 'access' (Finnegan, 2006, Sapsford & Jupp, 2006; Scott, 1990). Authorship, which specifies the origin of documents, can be subdivided into 'personal' and 'official' sources. The former include, for example, diaries, autobiographies, or personal notes; while the latter can be found in 'bureaucracies'. We can also distinguish two types of official documents: state (governmental) and private (nongovernmental) files. The other criterion, which we can use in the organization of documentary sources, refers to 'access' or "the availability of documents to individuals other than the authors" (Jupp, 2006, p.277). Scott (1990, as cited in Jupp, 2006) identifies four types of access: closed, restricted, open-archival, and open-published:

> ‘Closed’ documents are available only to a limited number of insiders, usually those who produce them; ‘restricted’ documents are available on an occasional basis provided permission has been granted; ‘openarchival’ documents are those documents which are stored in archives and are available to those who know of them and know how to access them; ‘open-published’ documents are the most accessible of all and are in general circulation (277).

Fig 37: Typology of Documentary Sources of Data

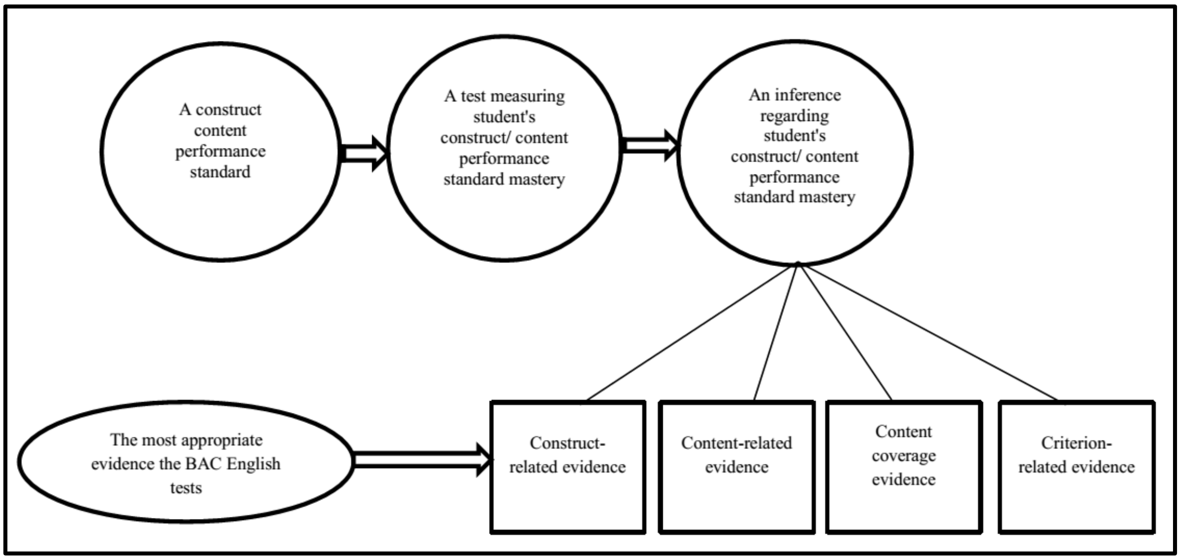


Adapted from Jupp, 2006, p. 277

In the case of this research, the documentary data or the evidential sources that we intend to collect concern official open-archival, and open-published information consisting of technology pupils' BAC English tests from 2001 to 2006 (see appendix B), the scores obtained by engineering pupils in the 'wilaya' of Eloued during seven BAC sessions (see appendix A) (Guémar Technical School, 2001-2006; Eloued Orientation Centre, 2001-2006) as well as the official syllabi designed for these specialities from 2001 to 2006 (Ministry of Education, 1998, 2000, 2001, 2004).

Demonstrating the validity of test score interpretations, uses, and consequences requires the collection of three types of evidence in relation to construct representation, content relevance, and content coverage (AERA, APA & NCME, 1999; Bachman, 2007; Messick, 1995; Popham, 2003, 2004, 2009). The first type of evidence informs us whether the BAC English tests from 2001 to 2006 have measured the constructs intended to be measured. The second type enables us to see whether the content of these tests mirrors the content of the official syllabus. The third type examines the extent of sampling from the content domain. In case the target domain is homogenous, the standard procedure that we normally follow is random sampling; however if the domain is designed around heterogeneous constituents, the technique we will implement refers to 'stratified random sampling' (AERA, APA & NCME, 1999; Bachman, 1990; Cronbach & Meel, 1955; Messick, 1989, 1955). Each type is, as it is shown in Fig 38, usually collected by some sort of investigation or analytic effort contributes to the conclusion that a test is yielding data that will support valid inferences (Popham, 2003).

Fig 38: Evidence for Validating Test Score Interpretations, Uses and Consequences


A construct content performance standard
A test measuring student's construct/ content performance standard mastery
An inference regarding student's construct/ content performance standard mastery
The most appropriate evidence the BAC English tests
Construct-related evidence
Content-related evidence
Content coverage evidence
Criterion-related evidence

Modified from Popham 2003, p 50

As far as this research is concerned, the analysis of the data collected from the documentary sources will be used to verify hypotheses two, three, and four. These hypotheses respectively assume: (1) that technology specialties' BAC English tests from 2001 to 2006 did not measure the constructs that test designers intended to measure; (2) that the content of these tests did not represent the content of the official syllabuses; and

(3); that the content of the tests failed to sample from the different themes of the content domain. In the same way, this type of evidence will, as we have mentioned in the introduction of this chapter, be used to support or to disproof the claims about technology streams' score interpretation, uses and consequences.

6.5.2. Defining the Construct to Measured

As we have mentioned in chapter II, three main approaches have been identified for the definition of constructs. For example, we can speak of trait-based constructs when we want to measure what test takers have in terms of language competence. Performance or task-based approaches focus on what test takers can do in situations beyond the test itself. Moreover, constructs can also be defined in terms of interaction between underlying abilities and the external contexts (Purpura, 2004; Chapelle, 1999, 2012; Chapelle, Enright & Jamieson, 2008, 2010).

The conceptualization of constructs has also been delineated in terms of their relationship with topical knowledge (the content domain) (Bachman & Palmer, 1996). In the first context, topical (thematic) knowledge is completely excluded from the definition of constructs. This occurs in situations where specific knowledge is not of significant importance to learners. In the second context, topical knowledge is incorporated within the delineation of constructs, especially when the former constitutes an integral part of the program of study. This, for example, occurs in thematic-based syllabi, such as the content syllabus designed for the third year pupils in Algeria (Ministry of education, 1998). In the third context, topical knowledge is in itself defined as a construct. This is appropriate in cases "where language-for-specific-purposes ability is defined as topical knowledge and language knowledge. The construct in these cases involves a discipline specific component of learning points and is usually determined in conjunction with a subject-matter specialist" (Purpura, 2004, p. 160).

6.5.3. Defining the Construct to be measured in Technology Streams

In Chapters I and II, we have stated that applied linguists and language testers do not conceptualize the concept of ESP constructs in the same way. Some of them argue that ESP testing is not built around a theoretical definition of the language ability to be tested; and it is the degree of specificity in content that distinguishes a general language test from an ESP test (Basturkmen, 2006, 2010; Davies & Elder, 2004; Widdowson, 2001, 2003). However, other linguists, such as Alderson and Bachman (2000) and Douglas (2000, 2006, 2013) maintain the fact that ESP tests are based on a theoretical description of specific purpose language ability, which according to Douglas (2000) "results from the interaction between specific purpose background knowledge and language ability, by means of strategic competence engaged by specific purpose input in the form of test method characteristics" (p.40).

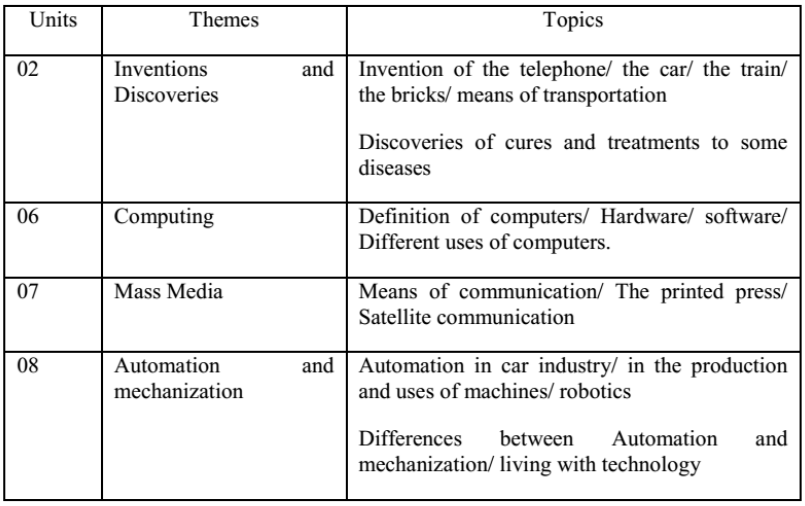
Now whether we consider ESP testing as a practical field or as emerging form a theoretical description of specific language ability, the main consideration relevant to these tests is that thematic knowledge, specific purpose background knowledge and specific purpose input (test content) are all combined to constitute the construct to be tested in ESP classes (Douglas, 2000, 2013; Purpura, 2004).

In the process of evidence gathering, Bachman (1990) reminds us that "if we cannot examine an actual copy of the test, we would generally like to see a table of specifications, example items, or at least a listing of the content areas covered, and the number of items, or relative importance of each area"(p. 244). As far as this research is concerned, the copies of the test, and the listing of the content areas of the syllabus are all available to us (Ministry of Education, 1998, 2000, 2004; ONEC, 2001-2006). In other words, in order to demonstrate the availability of construct representation, content relevance, and content coverage, we need to examine the areas covered in the official program of study and match them to copies of actual BAC English tests.

6.5.4. Analysis of Technology Streams' Instructional Syllabus

The Ministry of Education (1992) emphasizes that the syllabus designed for technology specialties at the first and second year levels “has been restricted to selected functions in relation with E.S.P and their related structures” (p 5). However in the third year level, "it was thought useful to build the syllabus around themes" (Ministry of education, 1998, p 11). As it was stated by the Ministry of education (1998, 2001), technology and technical streams share the same content domain which is built around four main themes (see table 72): 'inventions and discoveries', 'computing', 'mass media' and 'automation and mechanization'. The first unit includes themes accounting for the history of inventions and discoveries and their impact on modern life. The second unit describes computers in terms of hardware, software, and uses. The third theme 'mass media' identifies the contribution of the means of communication in getting us well-informed; and finally in 'automation and mechanization', the themes focus on describing machine industrialization and robotics.

Table 77: Technology and Technical Streams' Syllabus



Units	Themes	Topics
02	Inventions and Discoveries	Invention of the telephone/ the car/ the train/ the bricks/ means of transportation Discoveries of cures and treatments to some diseases
06	Computing	Definition of computers/ Hardware/ software/ Different uses of computers.
07	Mass Media	Means of communication/ The printed press/ Satellite communication
08	Automation and mechanization	Automation in car industry/ in the production and uses of machines/ robotics Differences between Automation and mechanization/ living with technology

Source: Ministry of Education, 1998

6.5.5. Analysis of Technology Streams' BAC English Tests

Technology specialties' BAC English tests relevant to this study include seven copies. Two copies of 2001 (June and September) sessions; and five other copies relevant to the sessions that had been held from 2002 to 2006 (see appendix B). This study takes the year 2001 as a starting point for its analysis because it is in this year that the syllabus witnessed radical changes (Ministry of Education, 2000).

After we have examined the BAC English test of June 2001 'the Use and Misuse of Science', which described the advantages and disadvantages of inventions and discoveries (Unit 2), we found that this theme and topic constitute an integral component of the pupils' official syllabus. However despite its representation of the construct and relevance to the content, its deficiency lies in the failure to sample from the other units of the domain, which can, according to Messick (1996), threaten the validity test score interpretations:

> [I]t is not sufficient merely to select tasks that are relevant to the construct domain. In addition, the assessment should assemble tasks that are representative of the domain in some sense. The intent is to ensure that all-important parts of the construct domain are covered, which is usually described as selecting tasks that sample domain processes in terms of their functional importance. Both the content relevance and representativeness of assessment tasks are traditionally appraised by expert professional judgment, documentation of which serves to address the content aspect of construct validity (p. 249)

In September 2001 session, the test was intended to measure information relevant to biological knowledge such as ways of prolonging life or delaying ageing. The topic of this test emerges from Unit 11 'Great Challenges to Mankind', which does not form a part of technology specialties' syllabus. In the same way as September session, in 2002 the test failed to measure the defined construct in that it measured topics from 'Unit 11' describing pollution and ecological problems. In 2003, the focus of the test was on physical fitness, weight control, dietary and physical exercises. These topics constitute a part of Unit 01 'Modern Life in English Speaking Countries'. This unit includes other topics such as consuming habits, education, family life, holidays and recreation, lifestyle, sport, youth, and so on. In the same way as 2003 session, the 2004 test concentrated on describing a topic from Unit 01 the 'Origin of Soccer' and how football has been played along the history. Once again, in 2005 test designers introduced a topic from 'Unit 01' measuring knowledge about holidays and recreation in the USA. In 2006, the test sampled from (Unit 11) to measure information about the Ozone layer. The summary of the constructs and content domains measured in technology specialties from 2001 to 2006 are illustrated in Table 78.

Table 78: Technology Specialties' Test Constructs and Content from 2001-2006

Session	Unit	Theme	Topic of the test
June 2001	02	Inventions and Discoveries	The Uses and Misuses of Science
Sept 2001	11	Great Challenges to Mankind	Prolonging life and Delaying Ageing
2002	11	Great Challenges to Mankind	Pollution of Oceans
2003	01	Modern Life in English Speaking Countries	Sport/ getting fit
2004	01	Modern Life in English Speaking Countries	The Origin of Soccer
2005	01	Modern Life in English Speaking Countries	Holidays and recreation in English Speaking Countries
2006	11	Great Challenges to Mankind	The Ozone Layer

Source: ONEC, 2001-2006

These constructs form, as Table 79 implies an integral part of natural and exact sciences specialties with whom, technology pupils have shared the same test from 2001 to 2006.

Table 79: Literary and Scientific Streams' Syllabus

Unit	Theme	Topics
01	Modern Life in English Speaking Countries	Youth and their Problems/ Family Life/Education / Sport Consuming Habits/ Democracy.
05	Trade and Development	Trade Relationship/ Market Research/ The Developed and the Developing Countries/ Work and unemployment
07	Mass Media	Means of communication/ The printed and the broadcasted press / Satellite communication
09	Human Rights and Racial Problems	UN Declaration of Human Rights/ Individual Liberties/ Apartheid and Racism/ Immigration
11	Great Challenges to Mankind	Ecology and Environment/ Pollution./ Overpopulation/ Starvation .Social Evils/ Natural Disasters / Wars/ The Space Race / Health

Source: Ministry of Education, 1998, p. 89

6.5.6. Analysis **of Technical Streams' BAC English Tests**

The discrepancy of technology streams' English tests with the content domain which they have studied from 2001 to 2006 led us to examine the BAC English tests of the specialties (technical streams) with whom they have shared the same syllabus (see Appendix D). The 2001 test described mass media and their contribution to shaping the American public opinion. In 2002, the test included information on the role of internet as the largest communication network. The theme of 2003 session sampled its content from automation and mechanization, and illustrated how machines have replaced craftsmen in the manufacture of goods. In 2004, the test introduced the advantages and disadvantage of automation. The 2005 session introduced the benefits of the different means of transportation. In 2006, the theme of the test was 'automated industries'. As Table 80 implies, unlike technology specialties whose tests failed to measure the constructs that have been supposed to be measured, technical and scientific pupils' BAC English tests have successfully assessed the intended constructs.

Table 80: Content of Technical Streams' BAC English Tests 2001-2006

Session	Unit	Theme	Topic of the test
2001	07	Mass Media	The printed and the Broadcasted press
2002	02	Inventions and Discoveries	The internet
2003	08	Automation	The manufacture of goods
2004	08	Automation	Automation in society
2005	02	Inventions and Discoveries	The means of transportation
2006	08	Automation	The impact of automation on our way of life

Source: ONEC, 2001-2006

In conclusion, validity is not a property of the test itself, nor is it a property for the resulting scores; rather, it is a quality of the interpretations and uses that we propose for the obtained scores. Now, if the interpretations and uses are supported by empirical evidence, such as construct representation, content relevance, and coverage (Bachman, 1990), they will be "considered to have high validity (or for short, to be valid), [however, if the] interpretations or uses that are not adequately supported, or worse, are contradicted by the available evidence are taken to have low validity (or for short, to be invalid)" (Kane, 2013, p. 3[parentheses in original]).

As we have seen previously, the evidence that we have collected from technology pupils' test copies from 2001 to 2006, and the syllabus they have learnt, suggests that except for June 2001 session, their BAC English tests failed to ensure three main qualities: construct representation, content relevance and content coverage. In ESP testing, the construct to be tested consists of thematic (topical) knowledge; and these tests did not measure this construct. In addition, when we compared the test copies to the areas covered in their syllabus, two other drawbacks have been identified. The first concerns content irrelevance to the content domain; and the second is related to the deficiency of these tests to sample from the various components of the syllabus. This is because "demonstrating that a test is relevant to and covers a given area of content or ability is…a necessary part of validation" (Bachman, 1990, p. 244).

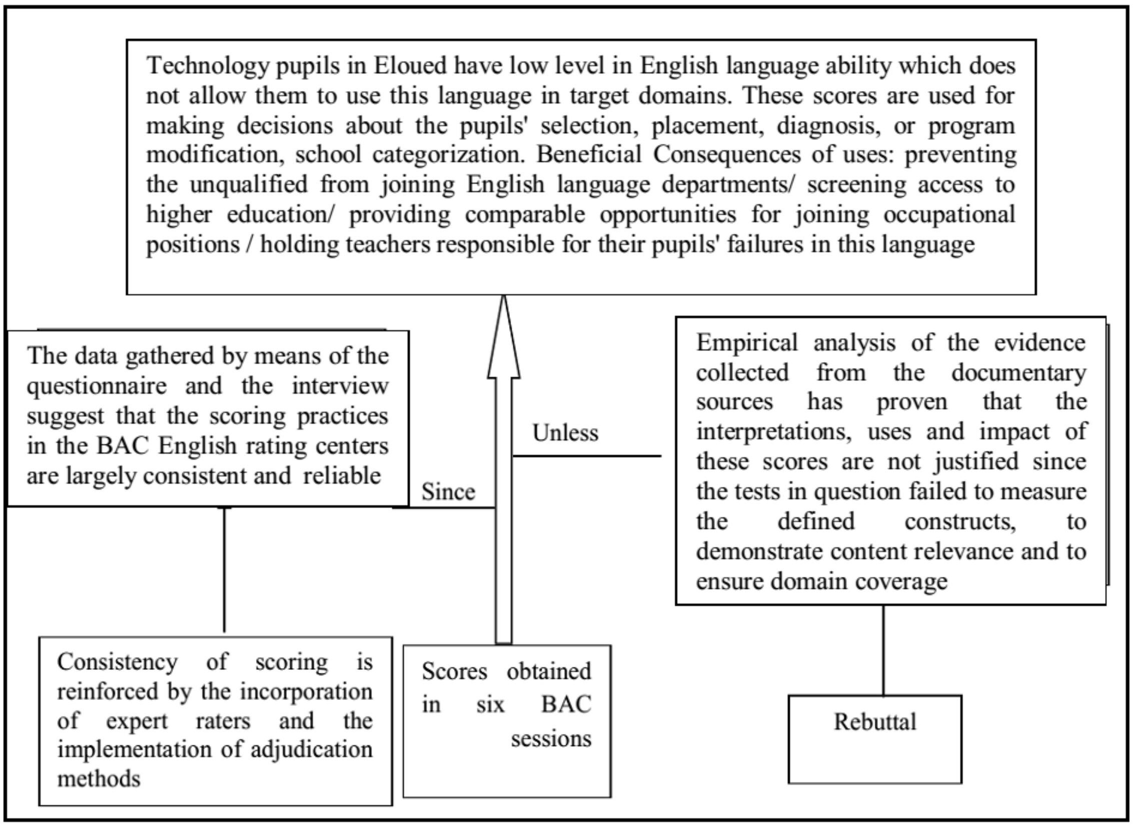
In short, our analysis of the data that we have gathered from technology streams' BAC English tests and the content of the official syllabus served us to verify and confirm hypotheses two, three, and four. This has enabled us to state that technology pupils' BAC English tests (September, 2001-2006) did not assess the constructs that these tests have been designed to measure; that these tests have failed to represent the content of the official syllabus; and that they have also failed to sample from the different constituents of the domain in question.

Based on this analysis, we can confirm that the BAC English tests designed for technology streams from (September, 2001-2006) lacked four main aspects of construct validity: construct representation, content relevance, content coverage and criterion relatedness (for the scores obtained from invalid tests cannot be used as a predictor for examinees' language performance in non-test contexts). This, of course, will distort the efficiency of the other aspects since validity operates as an overall unifying concept. In other words, the score interpretations, the purposes for which the interpretations have been used and the emerging decisions and consequences have all proven to be unjustifiable; and thus considered to be invalid.

6.5.7. Summary of the Validity Argument

As we have pointed out in the introduction of this chapter, language testers emphasize that the best technique which can be employed in validating the meaning of test scores is by the incorporation of Toulmin's arguments (see Fig 39).

Fig 39: The Incorporation Toulmin's Argument in Validating the Interpretations of the Scores Obtained by Technology Streams



Technology pupils in Eloued have low level in English language ability which does not allow them to use this language in target domains. These scores are used for making decisions about the pupils' selection, placement, diagnosis, or program modification, school categorization. Beneficial Consequences of uses: preventing the unqualified from joining English language departments/ screening access to higher education/ providing comparable opportunities for joining occupational positions / holding teachers responsible for their pupils' failures in this language

The data gathered by means of the questionnaire and the interview suggest that the scoring practices in the BAC English rating centers are largely consistent and reliable

Empirical analysis of the evidence collected from the documentary sources has proven that the interpretations, uses and impact of these scores are not justified since the tests in question failed to measure the defined constructs, to demonstrate content relevance and to ensure domain coverage

Consistency of scoring is reinforced by the incorporation of expert raters and the implementation of adjudication methods

Scores obtained in six BAC sessions

6.5.7.1. The Claim:

The claim in this validity argument is multi-componential: it accounts for the interpretations provided to technology pupils' BAC English test scores; the purposes for which these scores have been used; the decisions based on score uses and the probable consequences resulting from these decisions.

6.5.7.1.1. Score Interpretation: Technology streams' BAC English scores in seven sessions imply that these pupils have low level of language ability to the point that they will not be able to use English in target language situations whether for academic or for occupational purposes.

6.5.7.1.2. Score Uses: These interpretations are used as a basis for making decisions about pupils' placement, classification, diagnosis, prognostic, prediction, or selection

6.5.7.1.3. Consequences Affecting the Pupils: The consequences include increasing the rate of failure in the BAC exam, denying the pupils' certification, limiting their opportunities to join higher education institutions, or English language departments, minimizing their chances for occupational positions; or even expulsion from formal education.

6.5.7.1.4. Consequences Affecting the Teaching Staff

Low scores can affect teachers in different ways, for instance, their 'self-esteem, reputation can be affected and their career progression can slow down (Wall, 2012, p. 79). In El-Oued, the Orientation Center publishes a yearly evaluation record measuring teachers' contribution to the improvement of test takers' levels of language ability (2001-2006). This document often specifies teachers' output in technology BAC English tests as of 00% (Orientation Center of Eloued, 2001-2006)

6.5.7.2. The Warrant

As we have mentioned previously, the warrant allows us to justify the logical chain of inferences that we intend to make starting from the datum (pupils' scores) to the conclusion or the claims. The information that we have gathered by means of the questionnaire and the interview concerning the scoring procedures, such as standardization meetings, the use of the scoring guide, blind double scoring implies that this process is largely consistent and reliable.

6.5.7.3. The Backing:

Toulmin (2003) points out that "logic is concerned with the soundness of the claims we make—with the solidity of the grounds we produce to support them, the firmness of the backing we provide for them—or, to change the metaphor, with the sort of case we present in defence of our claims" (p.7). In certain cases, the warrants that we provide in order to support the claims can be challenged especially if they do not stand on solid grounds. Consequently, these warrants need to be reinforced with backings. In this validity argument, the backing that we provide in defense of the claims concerns the procedures implemented for adjusting adjacent scores and settling discrepancies (rater mean and adjudication or mediation methods). In addition to the level of raters' expertise, these methods tend to solve any type of differences resulting from intra-rater or inter-rater inconsistencies.

6.5.7.4. The Rebuttal

This component refers to the "circumstances in which the general authority of the warrant would have to be set aside … [or to] the exceptional conditions which might be capable of defeating or rebutting the warranted conclusion" (Toulmin, et al., 2003, p.94.). If the collected evidence supports the warranted logical chain of inferences from the datum to the claim, we can say, that the interpretation, uses and consequences of technology pupils' scores in seven sessions are valid. Conversely, if we demonstrate that the gathered evidence does not support this chain of inferences, this will be taken as sign of invalidity of the meaning that we have provided for these scores.

Now, after we have matched the evidence that we gathered by means of the test copies (ONEC, 2001-2006) to the official syllabus and the content domain designed for technology specialties (Ministry of education, 1998, 2000), we found that apart from June 2001session, the test in the other sessions (September, 2001-2006) did not measure the construct intended to be tested; nor did it mirror the content included in the official program of study. In addition to content irrelevance, these tests failed to sample from the whole components of the official syllabus. This implies that the interpretations, uses, and consequences emerging from technology streams test scores from September 2001 to 2006 are invalid.

In conclusion, the chain of logical inferences (validity argument) starting from the datum (technology pupils' scores in Eloued 2001-06) to the claim (score interpretation, uses and consequences) supported by the warrant (reliable scoring procedures); and reinforced by the backing (adjudication methods and expert raters) have been defeated (invalidated) by the rebuttal (construct and content underrepresentation, content irrelevance and criterion 'unrelatedness').

Conclusion of Field Study

In this chapter, we have conducted a field study for the purpose of validating technology pupils' score interpretations, uses and consequences in seven BAC sessions: (June and September 2001-2006). The validation process was implemented by means of Toulmin's (2003) argumentation model. This required us the collection of data by means of three instruments: the questionnaire, the interview and from documentary sources.

We have started this process with the analysis of technology pupils' BAC English test scores from 2001 to 2006 as the datum (D) of the validity argument. Based on these data, the scores interpretations (the claim (C)) imply that these pupils have low levels of language ability preventing them from engaging in communication in real target language domains. The logical chain of inferences between the scores (D) and their meaning (C) have been supported by the information that we have gathered by the questionnaire and the interview (warrant) and reinforced by the rating expertise and methods for settling differences (Backing) suggesting that scoring in the BAC exam rating centers is consistent and reliable.

However, the reliability of scoring has been challenged and defeated by the findings that we have reached as a result of the empirical documentary analysis (the Rebuttal). After we have matched technology streams' BAC English tests to the official syllabi designed for these specialties we found that expect for June 2001, not only have the other tests been invalid, but they have been used for unintended purposes leading to negative consequences affecting all the stakeholders as well.

Findings Implications and Recommendations

Chapter Seven

Implications, Findings and Recommendations

Introduction

This chapter discusses the main results of the research, provides pedagogical implications for the institutions and individuals responsible for designing large scale assessment and proposes some recommendations intended to improve the process of English language Testing in the BAC examination.

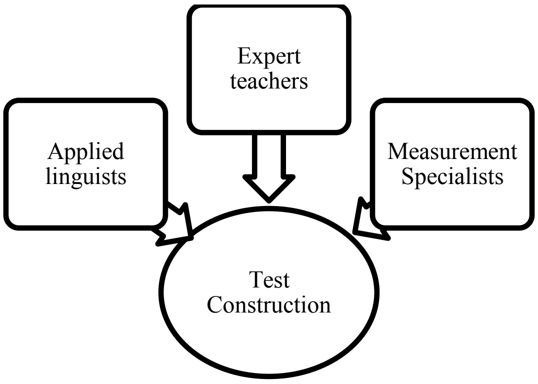
In this perspective, the chapter describes the BAC English developmental phases starting from its initial conceptualization until live test delivery. It lists and discusses the main findings relevant to the test makers, its constructional constituents, item writing and compilation and the mechanisms used for measuring test scores consistency and inferring its validity. Then, it concludes with some suggestions highlighting the main criteria for the selection of the BAC English test writers, raters and adjudicators, the appropriate test architectural layers and sequential stages; and the techniques implemented for item facility value and indiscrimination indices, reinforcing inter rater and intra rater reliabilities and argumentation frameworks for validating the score interpretations, the purposes for which the scores are intended to used and the impact which might affect all the stakeholders.

7.1. Implications and Findings

7.1.1 Implications for Test Constructors

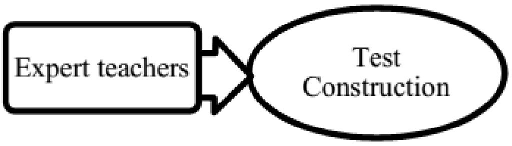
Reviewing the literature relevant to language test construction, the specialists in the field emphasize that test development requires three types of expertise (see Fig 40): applied linguists, psychometricians and teachers (Alderson, 2001; McNamara, 2011). The role of applied linguists is to provide explanations of how the components of theories of language interact to create and interpret discourse. At the same time, they can tell us how to define the constructs to be measured, and how to specify them for testing contexts. Psychometricians or measurement specialists can, on their part, draw the broad lines for the rating procedures and ensure "the fidelity of the scoring structure to structure of the construct domain" (Messick, 1996, p.248) so that numbers (scores) can be interpreted as real indicators of test takers' language ability. Furthermore, applied linguists and testers need to be informed by the persons who work on the ground, the teachers:

Fig 40: The Tripartite Test Constructors



As it has been confirmed by the ONEC (2012), the design of the BAC English tests has always been the exclusive business of secondary school inspectors and teachers. This reminds us of Sposlky's division of the history of language testing (Spolsky, 1979) and more specifically of the pre-scientific period which "still holds sway in some parts of the world" (p. 6) including our country where language tests are the exclusive business of teachers "or, in more formal situations, of language teachers promoted or specially appointed as examiners. No special expertise is required, if a person knows how to teach, it is to be assumed that he can judge the proficiency of his students" (pp. 6&7). The test design practices in the prescientific period as well as in the BAC English test can be illustrated in Fig 41.

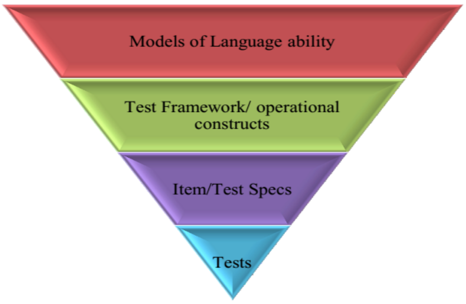
Fig 41: Constructors of the BAC English Test



7.1.2. Implications on Constructional Layers

The literature relevant to language testing identifies three hierarchical layers for test construction: models of language ability, test frameworks and specifications. The first layer provides conceptual understanding of how the components of language ability interact and generate language use. The frameworks select the constructs from the models and operationalize them for particular testing situations. Generated by the information in the frameworks, the specifications tell us how to write items and how to compile them into comprehensive tests. The hierarchical layers generating the construction of language tests are illustrated in Fig 42.

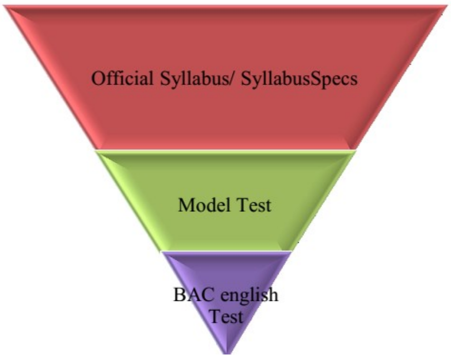
Fig 42: The Hierarchical Layers Generating Test Construction


Models of Language ability
Test Framework/ operational constructs
Item/Test Specs
Tests

Modeled after Fulcher and Davidson, 2007, 103

In the BAC English test, two main components (layers) generate the construction of tests: official syllabuses/ and or syllabus specs and model tests designed by expert teachers (see Fig 43). Language testers emphasize that we cannot design tests directly from syllabuses, or syllabus specifications because the former specify what students will be taught in terms of instructional domain and the latter tell teachers and learners what tests will contain in terms of tasks. The syllabus specs are meant to the persons "who wish to prepare for the test…or to publishers who wish to produce materials related to the test" (Alderson, et al, 1995, p. 9). However, the test specification is designed for test developers and users to tell them "what the test tests and how it tests it" (p.9).

Fig 43: Hierarchical Layers Generating the BAC English Test Construction

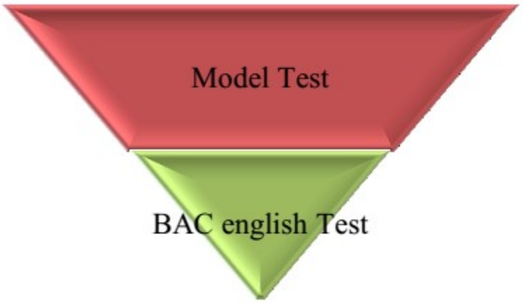


Modeled after Fulcher and Davidson, 2009,p. 127

In addition to the official syllabus and syllabus specifications, the construction of language tests in the BAC examination is built upon a 'model test'. The latter is used by test developers as a reference for item design. However, model tests in themselves are not informative because they lack the minimal instructions, which tell us how to design measures (see chapter III test task characteristics).

In technology streams, we hypothesize that the construction of the BAC English tests has exclusively derived its content and tasks from model tests and not from syllabus specs nor from syllabus content (see Fig 44). The justification that we have used to support this judgment responds to the following question: if technology streams' BAC English tests were written from the programs of study relevant to these specialties; then, why have these tests continuously failed to represent the content of this syllabus?

Fig 44: Hierarchical Layers Generating Technology Streams' BAC English Test



In conclusion, we can say that English language test construction in the BAC examination has not been informed by constructional layers recommended by language testers, which include theories of language abilities describing the constructs of language use; frameworks operationalizing these constructs for particular testing situations and generating the test specification; and test blueprints telling us how to write items and how to accumulate them into complete tests.

7.1.3. Implications on Theme Selection

The examination of the syllabus, which has been intended to technical and technology streams on the one hand; and the one designed for scientific and exact sciences on the other, suggests that the choice of themes for the BAC exam tests is not justifiable. We included technical and exact sciences in this assumption because the empirical analysis of the information collected from the documentary sources has demonstrated that technology streams share the same syllabus with technical branches; at the same time, they share the BAC English tests with the scientific specialties (see appendices B & D).

Returning to topic selection, which we have considered arbitrary and not warranted, let us, for example, consider the themes incorporated in technology and scientific specialties' tests from 2001 to 2006. In September 2001, the theme was chosen from 'Unit 2'; in 2003, 2004 and 2005, the topics of the test derived their content from 'Unit 01'; and in September 2001, in 2002 and in 2006 the test was built around themes from 'Unit 11'. Themes from unit 05 'Trade and Development', unit 07 'Mass Media' and unit 09 'Human Rights and Racial Problems' have completely been discounted.

Similarly, this assumption can be can be extended to the tests written for technical specialties. In 2001, the topic of the test was selected from Unit 07 'Mass Media'; in 2002, 2005, the themes were chosen from Unit 02 'Inventions and Discoveries'; and in 2003, 2004 and 2006, the test content emerged from Unit 08 'Automation and Mechanization'. Unit 06 'Computing' was discounted exception for 'activity one' in 'Section Two' of 2002 session, which required test takers to supply punctuation and capitalization for the following sentence: "the computer is an electronic devise that works at enormous speed it processes data following a given programme now people can use it to receive messages and information" (Technical Streams' BAC English Test, 2002, p.2).

7.1.4. Implications on the Stages of Test Development

In chapter III, we have identified Bachman and Palmer’s (1996) three stages for language test development, which include design, operationalization and administration. We have described how the design stage states the guiding purpose, which in its turn, delineates the scope of the construct(s) to be tested and draws the broad lines for data collection about the characteristics of test takers, test tasks and target language use tasks (learners' needs). We have also explained how operationalization uses the information collected in the first stage for item and test writing. In the third stage, we have seen the significant role of pretesting and try-out in collecting information about test takers' levels of language ability, test taking strategies, item facility values, discrimination indices, as well as about administration procedures.

According to the ONEC (2012, 2013), the process of English language test development from its initial conceptualisation to the production of one or more operational tests goes over three phases. By phases, we mean ‘time periods’ and not interrelated stages, such as the ones identified above in Bachman and Palmer's model (1996).

a- In phase one, teams of expert teachers meet at the local level to draft one or more versions of the test. This work should, as the Ministry of Education strongly recommends, derive its content and design from a model test, and learners’ fields of study (ONEC , 2012, 2013).

b- In the second phase, the ONEC calls local test developers for a central meeting in Algiers where the drafts of the test will rigorously be re-examined, revised, or even modified so as to be certain that items would be of medium difficulty, error free and fall within test takers' varied levels of language ability.

c- The last phase concerns live administration

Matching this process to the one sketched out by the specialists in the field, we can conclude that the BAC English test has failed to comply with the scientific standards of test development. First, this test did not inform itself from a design stage. We have seen that this stage focuses on the identification of the components that should be incorporated in any test such as the statement of the purpose of the testing situation; the delineation of the constructs to be measured; and the description of test takers' and test tasks characteristics.

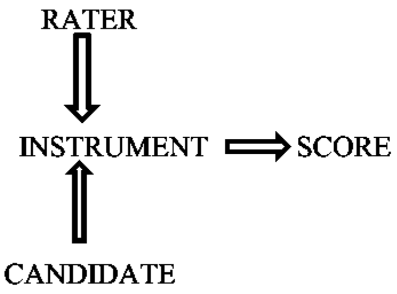
According to the ONEC (2010, 2013), items in this test are designed after a model test written by expert teachers. However, language testers strongly recommend against writing tests whether from other tests or from syllabus specifications. The latter can inform us only about what a test will contain in terms of content, or skills; and model tests cannot generate other tests. It is the test specification that tells developers what the test will test and how to test it. Additionally, it seems that test developers in the BAC do not employ any plan of test usefulness enabling them to conduct a pre-evaluation process to be certain of the test validity.

The other deficiency is that pretesting and tryout have not been implemented in the test developmental process (see Items 28 and 29 in the questionnaire). Information such as item facility value, discrimination indices, test taking strategies as well as information about examinees' levels of language abilities cannot be specified by the expertise of test writers, but can be obtained only as a result of field pretesting and tryout (Alderson, et al, 1995; Bachman & Palmer, 1998; Livingston, 2006; Nevo, 1998).

7.1.5. Implications about the Scoring Process

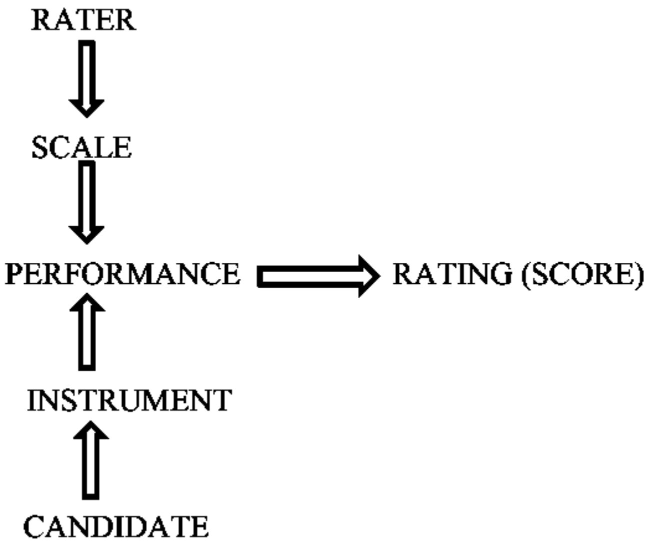
The information that we have collected by means of the questionnaire and the interview suggests that the testing procedures are, largely, consistent and reliable. A standardization meeting to ensure that the interpretation of the scoring guide will be uniform precedes live scoring. Additionally, the incorporation of expert raters, the implementation of anonymous double rating, and the adjustment of adjacent and discrepant scores reinforces the concept of score dependability and fairness. Nevertheless, there are three concerns that need to be raised. The first is relevant to the incongruity between the method by which chief examiners train raters to stay in agreement with their colleagues, and the one used in live scoring. In the pre-scoring sessions, rater variability is settled by means of discussion method; but in operational scoring, adjacent ratings and discrepancies are respectively resolved by the incorporation of 'rater mean' and 'Tertium Quid' methods. The second concern is related to the choice of sample scripts in order to ensure homogeneous understanding of the scoring guide. According to the respondents of the questionnaire, these scripts are picked out randomly. As we have mentioned in Chapter IV, scripts fall into two types: consensus and problematic scripts. The latter can be subdivided into off-task scripts, memorized scripts and incomplete scripts. Radom selection can result only in the choice of one of these types; which may eventually deprive raters of being trained by means of the other types. The third point concerns the measurement of 'written expression'. Language testers identify two types of scoring: objective and subjective scoring. In the former, examinees' responses are "determined entirely by predetermined criteria so that no judgment is required on the part of scorers" (Bachman, 1990, p. 76). In the latter, "the scorer must make a judgment about the correctness of the response based on her subjective interpretation of the scoring criteria" (p. 76). The information we gathered from our subjects as well as from the BAC English test-scoring guide (see appendices C and E) ascertain that no type of rating scale is used in scoring the written expression tasks. This means that every individual rater interprets the scoring criteria according his/her own judgment (see Fig 45). Language testers emphasize that in subjective ratings "there is no feasible way to ‘objectify' the scoring procedure" (p. 76) unless we use rating scales (see Fig 46).

Fig 45: Subjective Scoring of the BAC English Tests



Modified from McNamara, 1996, p. 121

Fig 46: Techniques Implemented in Subjective Scoring



Source: McNamara, 1996, p. 121

7.1.6. Implications on Test Construct and Content

In Chapter II, we reviewed the definitions suggested for ESP constructs. We have seen that these constructs result from the interaction between "specific purpose background knowledge and language ability by means of strategic competence engaged by specific purpose input in the form of test method characteristics" (Douglas, 2000, p. 40). This implies that the delineation of ESP constructs in technology streams should demonstrate itself from the extent of interaction between the components of the language ability in question and the pupils' fields of specialism, such as mechanics, electricity, or architecture by means of strategic competence; and the extent to which these constructs are engaged by the specificity of the test content.

. In secondary education, three specialties of technology streams can be identified: civil, electrical and mechanical engineering (Ministry of education, 1998). Civil engineering "is concerned with making bridges, roads, airports, etc. Mechanical engineering deals with the design and manufacture of tools and machines. Electrical engineering is about the generation and distribution of electricity and its many applications."(Glendinning & Glendinning, 1995, p. 11). Additionally, within each specialty, other sub-branches can be delineated. The examination of technology streams' test copies (see appendix B) and their official syllabus implies that apart from June 2001, none of these constructs has been measured. So, when the test "fails to capture important aspects of the construct, it implies a narrowed meaning of test scores because the test does not adequately sample some types of content, engage some psychological process, or elicits some ways of responding that are encompassed by the intended construct" (AERA, APA & NCME, 1999, p. 10).

7.1.7. Implications on ESP Syllabuses

In their definition of LSP, Basturkmen and Elder (2004) state that this approach is used to "refer to the teaching and research of language in relation to the communicative needs of speakers of a second language in facing a particular workplace, academic, or professional context" (p. 73). The authors stress that in LSP classes, learners are categorized according to their subject of specialism where "courses usually focus on the specific language needs of fairly homogeneous groups…in regard to one particular context referred to as the target situation" (p. 73). The issue of 'specificity of tasks' has also been raised by Douglas (2000) when he inquires whether it is good enough for us to design one syllabus for engineering specialties; or it would be better and more convenient to write a specific syllabus for each subspecialty. Take for example mechanical engineering specialty, which can, according to the author, be subdivided into "combustion science, dynamics, fluid mechanics, metrology, micro-electromechanical systems, nanostructures, tribology, and thermal engineering" (p.48). It is only when the test is highly specific, that it can engage test takers' specific background knowledge to interact with the test input (Douglas, 2000, 2001, 2013).

Returning to the syllabus designed for technology streams in secondary education which includes the following themes: 'invention and discoveries', 'computing', 'mass media' and 'automation and mechanization' (Ministry of Education, 1998). These streams are subdivided into three specialties: civil, electrical, and mechanical engineering each of which requires a specific syllabus. If we accept, for instance, that this syllabus includes some topics about mechanics and electronics; what about the pupils of civil engineering whose specific background knowledge has fully been discounted whether in the syllabus, or in the BAC English test input.

7.1.8. Bias Associated to Test Content

Language testers identify two main sources of bias that can affect test scores: construct underrepresentation and construct irrelevant variances. The former refers to the test failure to adequately measure the construct which test developers intend to test. In the second, "the test scores may be systematically influenced to some extent by components that are not part of the construct "(AERA, APA & NCME, 1999, p. 10). The inspection of technology streams' test copies, and matching them to their program of study has demonstrated that much of the bias in their scores is related either to the inclusion of themes that extraneous to what they have previously studied; or to the inability of the test to select topics on the basis of equal representation.

7.1.9. Implications on Score Reporting

A score report can be defined as "a form of communication [which has] a sender, message, medium, intent, and audience. The sender…is the sponsoring agency or institution…, the message deals with the content of the score report and the medium is the score report format" (Ryan, 2006, p. 677). In Algerian secondary education, the sender refers to the body responsible for test development, rating, and score reporting represented by the ONEC. The messages refer to the initial lists comprising test takers' names and the pass/fail composite scores that they have obtained in the BAC exam; and a more detailed report consisting of the scores obtained by these test takers in the different subjects they have been tested in.

As far as English in technology streams is concerned, these reports do not comply with the minimum requirements set by the educational and psychological assessment institutions, which imply that scores should not be released to test takers unless their interpretations "describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of scores, and how the scores will be used [and] unless the validity…and reliability of such scores have been established" (AERA, APA, & NCME, 1999, p. 65). If the BAC English test developers had pre-evaluated technology streams' tests against Bachman and Palmer's six componential usefulness plan, such type of construction deficiencies might not have occurred.

7.1.10. Implications on Score Interpretations, Uses, and Consequences

"Tests are commonly administered in the expectation that some benefit will be realized from the intended use of the scores, [in case] all examinees have comparable opportunity to demonstrate the abilities or attributes to be measured" (AERA, APA, & NCME, 1999, pp. 16& 175). Some of the benefits of the BAC exam score uses include providing equitable opportunity for all test takers to join university departments or professional positions; at the same time, preventing the unqualified from joining such institutions or positions. However, the empirical study that we had conducted revealed that the BAC English tests in technology streams (2001-2006) did not provide opportunity for these pupils to demonstrate their standing, as their colleagues in the other streams, on the constructs intended to be measured. Consequently, the uses of test scores in these specialties brought about adverse consequences affecting these pupils, such as denying them from certification, preventing them from joining the university specialty of their preferences; limiting their opportunities for occupational positions; or even expulsion from schooling.

Summary of Findings

The findings that this research has reached can be listed as follows:

1- Test developers in the BAC exam are exclusively composed of secondary school teachers, ignoring the role of linguists in describing the abilities and the constructs to be tested; and that of psychometricians in telling how these abilities and constructs can be measured.

2- Despite the fact that language testers emphasize that every language test has a theory behind it, which outlines the constructs that generate the specification to feed test items (Alderson, 2000; Bachman, 2005; Fulcher, 2010), the process of test development in the BAC exam is still generated from model tests.

3- The developmental process in these tests did not go through a recognized set of sequential stages such as design, operationalization, and tryout.

4- The validation study that we had conducted did not support the interpretations of technology pupils' scores in that:

   a) The test failed to measure the defined constructs.

   b) The test content failed to demonstrate its relatedness to the content of the domain of study.

   c) This test failed to base its selection on equal representation of the domain of instruction.

5- Despite the fact that the scoring process has been found largely consistent, its deficiencies lie in the lack of rater training and rating scales in subjective scoring. Additionally, we have also found the method used for settling score variability in the standardization session discrepant with the ones implemented in live scoring (Discussion method versus Rater Mean and Tertium Quid methods).

6- The failure of the test to measure the construct that it has been intended to be measured implies that the meaning of the scores has been interpreted inappropriately, and cannot be considered as an indicator of technology pupils' level of language ability, or as a sign of their inability to communicate in nontest situations.

7- Scores emerging from these tests have been used for unintended purposes of selection placement, certification, criterion identification, program improvement, teacher evaluation and so on.

8- Due to the fact that the score uses have been based on inappropriate interpretations, unintended consequences have affected these pupils, such as increasing the rate of failure in the BAC exam, limiting their chances to join the higher education specialties of their preference, denying them certification, or even minimizing their opportunities for professional occupations.

9- We have also found that score report of the BAC exam has failed to abide itself by the standards of Educational and Psychological Testing (AERA, APA, & NCME, 1999) and more specifically by standard 5.10 which states that:

> When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what the scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used (p. 65).

Or by standard 5.12 which states that "scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established" (p. 65).

10- We have also found that theme selection for the test was not justifiable: test developers did not document the reason why certain themes have completely been discounted.

11- Concerning the design of LSP syllabuses in secondary education, language testers emphasize that highly specific syllabuses should be designed for homogenous groups of leaners (Basturkmen, 2006, 2010; Basturkmen & Elder, 2004; Douglas, 2000; 2013; Dudley- Evans & Waters, 1987). In Algerian secondary education, this rule has been reversed to design one homogenous syllabus for heterogeneous groups of learners.

7.2. Recommendations

The evaluation of English language testing in secondary education led us to conclude that this process has underestimated the ability levels of the pupils studying in technology streams. Contrary to the principles of fairness which require "that all examinees [should] be given comparable opportunity to demonstrate their standing on the construct(s) the test is intended to measure" (AERA, APA, & NCME, 1999, p. 74), this test did not afford this opportunity to technology streams. So, in order to provide equitable treatment to all examinees whatever their specialty is, this research proposes the following recommendations.

7.2.1. Test Developers

McNamara (2011) points out that language testers "typically enter the field from one of these sides: either statistics and measurement or language and linguistics, rarely both. Yet the best language tests are those that are richly informed by the best practice in both areas" (p. 435). Measurement specialists can tell us about technical qualities such as reliability and validity; and linguists can inform us about language learning and acquisition. Additionally, these two types of specialists need to be informed by the persons who are working on the ground, the teachers (Alderson, 2001).

Following the guidelines stated above, we recommend the ONEC to set up a joint committee of test writers comprising of specialists in the field such as university lecturers in the sciences of language, educational psychologists, and statisticians as well as teachers of high quality of expertise in developing BAC tests and in teaching examination levels.

7.2.2. Test Architecture

Before they engage in operational test writing, test developers need to respond to four questions: (1) what theory of language will uphold the test? (2) What components of this theory will the test measure? (3) What frameworks will specify the constructs to be tested? For example, the delineation should specify whether the constructs are trait-based, task-based, interaction-based, and topical-based and so on; and finally, (4) what specifications would generate the test, and its items?

7.2.3. Test Development

7.2.3.1.The Design Stage

Test writing should follow clear and explicit developmental stages in that the operational item writing needs to be preceded by the design stage, followed by test tryout, and pre-testing. In the design stage, test developers need to respond to six questions: (a) how will the problem be stated? (b) How will the constructs be delineated? (c) How will the tasks in the target language domain be defined and constrained? (d) How will the test tasks be described? (e) How will the plan of test usefulness be implemented? And (f) how will the required and available human and material resources be used?

7.2.3.2.The Operational Stage

We would like to emphasize that tests should not be written from syllabus specifications, nor from syllabuses or model tests because these instruments do not tell test developers what the test tests and how it tests it. The only documents that tell so are test blueprints or specifications. Syllabus specifications can only tell us what tests may contain in terms of content or tasks; and model tests do not specify how items can be written.

Consequently, we recommend test designers to write test items from item or task specifications and to use a test blueprint when they want to compile these items into a comprehensive test (Bachman, 1991; Bachman & Palmer, 1996). We also remind test developers in the BAC exam that Bachman and Palmer (1996) propose two strategies for item writing. These can be implemented either by modifying the TLU tasks described in the 'Design Stage', and incorporating them in the test; or by creating new ones from an item specification, which describes the following features: the characteristics of the setting, characteristics of the test rubrics, characteristics of the input, characteristics of the expected response and the relationship between input and response, as well as the specification of the scoring method. Once the items are pre-evaluated against Bachman and Palmer' (1996) six componential usefulness plan, the tasks can be compiled into a complete test. However, the compilation should not be done in a haphazard way but from a test blueprint.

7.2.4. Test Instructions

We also recommend that tests should include clear instructions. The latter enable test takers to understand the nature of the testing procedures, the way to respond to questions and how their responses will be scored. Bachman and Palmer (1996) stress that clear and effective instructions can "assure test takers that the test is relevant, appropriate, and fair" (p. 182). According to the authors, there are four main essential components for clear and effective instructions: (1) the statement of the purpose for which the test is intended; (2) the delineation of the abilities to be measured; (3) providing examples of task solution; and (4) an explicit scoring method.

7.2.5. Test Tryout

A draft of the BAC English test should be piloted in a sample of Algerian secondary schools before its operational administration. This will enable test developers to gather information about item facility values, discrimination indices, test taking strategies and administering procedures. Test writers should not downplay the role of this phase (pretesting) because "the literature concerning language tests suggests that the examiners' assumptions regarding what they test and their expectations from the respondents often do not match the actual processes which the respondents undergo during testing" (Nevo, 1989,

p. 20). What is worth mentioning here is that the 'BAC Blanc' is not informative for the features that we have mentioned above; its role is purely predictive. This examination can only provide us with some expectations about the rate of success in the live BAC exam and not for gathering information about the criteria we have previously identified.

7.2.6. **Live Administration of the Test**

Before the test is administered to test takers, "the test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to asses should be clearly described" (AERA, APA, & NCME, 1999, p. 17).

7.2.7. **The Scoring Procedures**

Despite the fact that the information that we have collected by means of the questionnaire suggests, to a large extent, that the rating process in the BAC exam is reliable, there are some concerns that need to be raised, such as the quality of raters, the disparity between the methods used for adjusting scoring variability in the standardization session and their implementation in live scoring, the quality of adjudicators and the absence of rating scales in subjective scoring.

Concerning the first point, we suggest that the qualification of raters and their expertise in rating should be documented so that we ensure that ratings will be conducted by the most competent teachers. As for the second concern, the information we have gathered from the respondents demonstrates wide disparity between the methods used for settling raters' differences in the standardization session and in live scoring. In the pre-scoring meeting, chief examiners train raters to solve their variability by means of discussion methods; whereas in operational assessment, this issue is settled by the involvement of a third rater (adjudicator). For an efficient scoring process, we suggest the implementation of adjudication methods in the pretesting session as well in live rating.

7.2.8. **Choice of Sample Scripts in the Prescoring Session**

We recommend chief examiners and raters not to select sample scripts for the pre-scoring session on random basis because this will not be of great value for the anticipation of the scoring difficulties. Following the directions of language testers (McNamara, 1996; Weigle, 1994, 2002), we suggest that these scripts would first be divided into two batches: consensus scripts and problematic scripts. Then, the latter will be rearranged into three types: off-task scripts, memorized scripts and incomplete scripts. This enables raters to anticipate the difficulties which may rise from any category of scripts.

7.2.9. Rating Scales

In the BAC English test, the only instrument that guides raters in scoring scripts is the scoring guide. This instrument specifies to raters the way they assign objective scores. Its main deficiency is the lack of rating scales to score the written expression section (see appendices C and E). In this context, we remind the writers of this test that the specialists in the field (language testers) do not see any other instrument appropriate for subjective scoring (written tasks) apart from rating scales whether primary trait, holistic or analytic.

7.2.10. Evaluation of Raters' Scoring Records

We suggest that the record of the discrepant raters will be documented and evaluated so that they may be invited for more training sessions; or if their scoring is not standardized by training, the 'academies' may not consider their invitation for future rating sessions.

7.2.11. The appointment of Adjudicators

The last recommendation relevant to the scoring process concerns the choice of adjudicators. The latter refer to raters of high level of expertise in scoring. When their colleagues assign discrepant scores to the same script, adjudicators are usually involved to settle this variability. In the BAC English test, the scoring process is conducted at three phases. In the first phase, the scripts are blindly rated; in the second phase, the same scripts are blindly scored by different raters. If the result of the scoring in the two phases manifests great varieties between the first and second raters, adjudicators will be invited to settle the variability. The point is that in the BAC English test, adjudicators are not chosen because of their expertise but out of the ones who live in the vicinity of rating centres. This is because the raters who live in distant places are usually freed as soon as the second phase of scoring comes to its end (see appendix G). As a result, we strongly recommend the ONEC and the chief examiners to select adjudicators from the most experienced raters.

7.2.12. Rater Training

One the one hand, there is much emphasis amongst language testers that "reliable ability measures are unlikely to be achieved from untrained raters" (Weigle, 1994, as cited in McNamara & Rover, 2006, p. 124). On the other hand, all the respondents in the questionnaire in addition to the chief examiner (see appendices F and G) informed us that they have never attended any meeting that is devoted to training raters on the scoring practices. We find this problematic and recommend the Ministry of Education to reinforce the literacy of assessment and scoring practices especially at the level of test writers and raters.

7.2.13. **Scores Interpretation, Uses and Consequences**

We recommend the ONEC to set up a committee of test validators whose role will focus on the evaluation of the interpretations of the released scores and the validation of the score uses and of the consequences resulting from these uses. If the empirical study asserts that the test is affected with construct irrelevant variances, construct or content underrepresentation, test validators should inform the ONEC that the interpretations it has provided for the scores are invalid; and cannot be used as indicators of test takers' language ability. At the same time, the information concerning the unintended uses and adverse consequences should be documented so that such kind of problems will not be experienced again. Furthermore, both test developers and users can be held responsible and accountable for the development and delivery of invalid tests and the misuse of invalid scores.

Finally, we recommend that the 'Standards for Educational and Psychological Testing' jointly edited by American Educational Research Association [AERA], American Psychological Association [APA] and National Council on Measurement in Education [NCME] in 1999 would be considered as one of the documents that guide the development of English language testing in the ONEC.

Conclusion

Chapter VII laid out the results of the research, provided pedagogical implications for test makers, users, validators and syllabus designers, and proposed a set of recommendations to the institution responsible for test design.

The findings of the research revealed some deficiencies related to test development and its rating. Concerning the first point, we found that the ONEC selects test designers exclusively out of teachers and inspectors, discounting the specialists in the sciences of language and psychometricians. The BAC tests are generated from other tests or syllabus specs ignoring the role of language theories, frameworks and specs; items' difficulty and discrimination indices are measured according to teachers' expertise not as a result of empirical investigation. The examinees' test taking strategies and their varied levels of language ability are not considered during test writing; and the complication of tests is not preceded by a pre-evaluation phase, nor is it followed by piloting and or test tryout.

As far as the second point is concerned, we listed these drawbacks. Concerning individual raters, the research revealed that they lack two main criteria: assessment literacy and scoring training. Additionally, the choice of adjudicators for the mediation phase is not warranted; the discrepant raters' records are no documented. As for the rating practices, we found the method used for settling raters' differences (discussion method) discrepant with the one used in operational scoring (Tertium Quid) methods. Additionally, the choice of sample scripts for the pre-ratings is arbitrary and not justified. Subjective scoring is conducted without any type of rating scales, and last but not least, the score release does not explain what constructs it has measured and how the scores will be interpreted and used.

The chapter concluded with a list of recommendations to the institutions and the individuals responsible for test design and the scoring procedures for the purpose of improving the process of English language testing in the Baccalaureate examination.

General Conclusion

The Ministry of Education in Algeria set several objectives for the teaching of English in secondary schools. In literary streams, for instance, the intent has been to enable the pupils to use this language for general communicative purposes. In scientific specialties (natural and exact sciences), more focus has been laid on the reading and writing skills since these pupils will, according to the Ministry, need English to read scientific publications and to write scientific reports. In engineering branches, the intended purpose has been to enable learners to use this language for specific purposes and in specific and constrained contexts (Ministry of Education, 1995, 1989, 1995, 2000, 2004).

Judging whether these objectives have really been attained calls for testing and assessment. The scores released by the 'Office National des Examens et Concours' during seven BAC sessions (2001-2006) imply that most of these objectives have not been accomplished, at least in technology specialties. Seeing that engineering pupils study at the same schools, use the same manuals and syllabuses, and are almost taught by the same type of teachers as their colleagues in the other branches, this study hypothesized that the cause of the problem might lie in the BAC English tests themselves.

In order to conduct this analysis, we have formulated four hypotheses. The first one attempted to examine the relationship between the pupils' underachievement and the scoring procedures in the BAC exam rating centers. The second one attempted to see whether the tests in technology specialties have measured the constructs that test developers intended to assess (construct representation). The third hypothesis examined the extent to which the test content is relevant to the thematic knowledge included in the official syllabus (content relevance). Hypothesis four focused on measuring the extent to which the BAC testers have implemented the concept of 'ecological sampling' or content coverage of the themes and tasks from the different constituents of the instructional domain.

Before we started verifying the hypotheses, we reviewed the related literature so that the research conclusions and findings could stand on empirical grounds. The literature review outlined five main areas relevant to test construction and evaluation. Tracing the historical development of language testing, for example, allowed us to identify to which period language testing in Algeria and mainly in the BAC exam is relevant. Reviewing the architectural and developmental constituents of test design helped us examine whether these components have been used for generating test construction in technology streams. The literature relevant to reliability and validity enabled us to measure the extent of scoring consistencies in the BAC examination rating centers, and to conduct a validation process for the purpose of reinforcing or discrediting the score interpretations provided the ONEC.

In order to verify the research hypotheses, we conducted a field study. As the descriptive and analytical methods imply, we used three instruments for gathering the relevant data: the questionnaire, the interview and the documentary sources. The questionnaire was administered to a population of 63 raters which represented the whole number of teachers participating in 2013 session in Eloued rating center. Since the majority of raters do not attend the mediation stage of scoring, we decided to interview the chief examiner of the same committee who had supervised this process from its initial until its final phases. In cases of "both logical analysis and empirical investigation" (Bachman, 1990, p. 256) relevant to construct and content validation, the questionnaire and the interview cannot provide us with all the required data. Consequently, we resorted to evidence collection by means of documentary sources such as papers of technology pupils' BAC English tests, their instructional syllabus and the scores they have obtained in seven BAC sessions (Bachman & Palmer, 1996; Messick, 1990, 1995).

We used the information gathered by the questionnaire and the interview in testing hypothesis one; and the data we got from the documentary sources to test hypotheses two, three and four. Hypothesis one has proven to be false, in that the data we gathered from the respondents have demonstrated that the scoring process and procedures in the BAC exam rating centers are,to a large extent, reliable and consistent. Conversely, the data that we have analyzed from the documentary sources have confirmed that hypotheses two, three, and four are true which implies that apart from June 2001 session, in the other sessions (September 2001-2006) not only have the BAC English tests in technology streams failed to measure the defined constructs; but they have not succeeded to demonstrate two other criteria as well: content relevance and content coverage.

We incorporated the results of the data analysis as the main components of the validity argument intended to support or contradict the interpretations provided for technology streams' test score interpretations, uses and consequences. The scores obtained by these pupils from 2001 to 2006 have been used as the argument data (D) which were interpreted as an indicator of their low level in English, and of their inability to use this language in specific target domains (the claim 'C'). The logical inference from the data to the claims was supported with the consistency of the scoring procedures (the warrant 'W') and reinforced by raters' expertise and adjudication methods (Backing 'B'). Nevertheless, the evidence we gathered from the documentary sources has rebutted the information included in the claim in that the score interpretations, the purposes for which the interpretations have been used as well as the consequences resulting from these decisions (uses) have all proven to be invalid.

Raising the questions of the relationship between validity and reliability, for instance, can there be validity without reliability; or can there be reliability without validity? We know that "while validity is the most important quality of test use, reliability is a necessary condition for validity;in the sense that test scores that are not reliable cannot provide a basis for valid interpretation and use" (Bachman, 1990, p. 279). At the same time, this research has responded that there can be reliability without validity on condition tests will not be concerned with the measurement of a defined construct or a specific content (Bachman, 1990, 1991; Henning, 1987; Moss, 1994). However, when the empirical study demonstrates, such as in the case of technology streams' BAC English tests, that the interpretations and uses of test scores are not valid since the test did not measure what it has been purported to measure, reliability will not be of great significance.

The findings of this study have revealed several deficiencies relevant to test writers, test developmental layers and stages, the scoring procedures in rating centers as well as score uses, score reporting and analysis. Against the directions of the specialists in the field who emphasize that test writers should be composed of linguists, psychometricians and teachers, test writing in the BAC examination has exclusively been the business of teachers. This is because expertise in teaching does not necessarily imply expertise in test writing. This study has also revealed that the BAC English tests have been generated from other tests written by expert teachers or inspectors ignoring the fact that test construction needs to be informed from models of language ability, test frameworks and specifications. Additionally, the lack of expertise and assessment literacy on the part of test developers has affected the dependability of item selection in that the concepts of 'item facility value' and 'discrimination indices' have been implemented according to teachers' own judgments and not as a result of empirical analysis. As for the score uses, this research has proved that they have not been used for the purposes for which they have been intended, and this resulted in adverse consequences affecting the pupils and their teachers. Concerning score reporting, this study has demonstrated that the ONEC constrains this process to score release without describing and justifying the link between these scores on the one hand and the test construct, content, uses or consequences (intended and unintended) on the other hand.

The research work has concluded with a list of recommendations aimed at improving the process of English language testing in the Baccalaureate examination. These involve recommendations for appropriate selection of test writing teams, guidelines for incorporating relevant test constructional layers and developmental stages, procedures for reinforcing the consistency of inter-rater and intra-rater reliability, provision of effective methods for score reporting as well as arguments for validating the score interpretations and the purposes for which the scores are intended to be used.

Concerning test writers, we commended the ONEC to set up teams comprising three types of expertise: university lecturers specialized in the sciences of language, specialists in educational measurement as well secondary school teachers with long expertise in teaching examination levels. The task of the first type will focus on delimiting the constructs intended to be tested. Psychologists will provide techniques for delineating how these traits can be measured. However, linguists and psychologists need to be informed by the persons who are working on the ground, the teachers who experience teaching and test development in examination levels on a on a day-to-day basis.

As far as test construction is concerned, we recommended test designers in the BAC exam against writing tests from other tests or from syllabus specifications because the former are not informative and cannot generate other tests, and the latter can only tell us what tests will contain in matter of content or tasks (Alderson *et al.* 1995). Instead, we proposed that test design should stand on three hierarchical layers: models of language ability describing what it means to know and to use a foreign language, test framework specifying the constructs to be measured, and test specifications telling how to write items and how to compile them into comprehensive tests.

Regarding test development, we suggested the incorporation of Bachman and Palmer's (1996) three-stage model including design, operationalization and administration stages. The first stage will focus on gathering information about the characteristics of test tasks and test takers. Based on the information gathered in the previous stage, the second stage delineates the operational steps for writing items and compiling them into a test. We also strongly recommended that the BAC English tests should not be administered to test takers unless they have been tried out in a representative sample of secondary schools. This enables us to gather valuable information about test takers' levels of language ability and their test taking strategies; about item facility value and discrimination indices; and last but not least, it helps us anticipate the difficulties that may rise during live testing.

Concerning the consistency of scoring in the BAC exam rating centers, we recommended the spread of assessment literacy amongst teachers through training, standardizing raters' interpretations of the scoring guide in the prescoring session, the use of rating scales in subjective scoring, the implementation of the most appropriate mediation methods for settling raters' discrepancies as well as norms for the choice of adjudicators.

On the subject of score reporting and interpretations, we recommended that when test scores are released to test takers, the ONEC should provide an interpretation describing what constructs and contents have been measured, what the resulted scores mean, where and how the score interpretations will be used, and how the consequences of the score uses can be beneficial to test takers and institutions; and in case the interpretations are proven to lack validity, test writers can be held accountable for the delivery of invalid tests.

As a summary, this research attempted to identify some of the factors responsible for technology pupils' underachievement in the BAC English tests from 2001 to 2006. In order to examine the plausibility of the interpretations provided for these pupils' scores, we conducted an empirical study verifying four hypotheses by means of the descriptive method instruments: the questionnaire, the interview and documentary sources. However, despite the fact that the respondents' answers suggest that the scoring practices in the BAC exam rating centers are largely consistent and reliable, evidence from documentary sources revealed several deficiencies relevant to construct representation, content relevance, domain coverage as well as criterion relatedness which question the credibility of the score interpretations and the purposes for which they have been used.

Reference List

Alderson, J.C. (1990). Testing reading comprehension skills. Part two. Getting Students to talk about taking a reading test (a pilot study). *Reading in a Foreign Language*, 7(1), 465–503. From http://nflrc.hawaii.edu/rfl/PastIssues/rfl62alderson.pdf

———. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London, UK: Macmillan.

———. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (eds.), *A new decade of language testing* (pp. 46-57). Arlington, VA: TESOL.

———. (1998). Developments in language testing and assessment, with specific reference to information technology. *Forum for Modern Language Studies* 34 (2) 195-206. From http://fmls.oxfordjournals.org/content/XXXIV/2/195.full. pdf.html

———. (2001). Testing is too important to be left to testers. In C. Coombe (Ed.), *Alternative assessment*. (pp. 1-14) Dubai: TESOL Arabia

———. (2000a). *Assessing Reading*. Cambridge: Cambridge University Press.

———. (2000b).Technology in testing: The present and the future. *System* 28, 593-603.Fromhttp://faculty.ksu.edu.sa/yousif/Research%20Papers/Technology%20in%20testing.pdf

———. (Ed.). (2002). *Common European Framework of Languages: Learning, teaching, assessment: Case studies*. France, Strasbourg: Council of Europe.

———. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, UK: Continuum.

———. (2007). The challenge of diagnostic testing: Do we know what we are measuring? In J. Fox et al. (Eds.), *Language testing reconsidered* (pp. 21–39). Ottawa: University of Ottawa Press.

———. (Ed). (2009). *The politics of language education: Individuals and institutions*. Bristol, UK. Buffalo, Canada, Toronto, NY: Multilingual Matters

———. (2011). The Politics of Aviation English Testing. *Language Assessment Quarterly*, 8(4), 386-403. From http://dx.doi.org/10.1080/15434303.2011.622017

Alderson, J, C., & Bachman, L. F. (Eds.). (2000–2006). *The Cambridge language assessment series*. Cambridge: Cambridge University Press.

Alderson, C, J., & Banerjee, J. (2001). Impact and washback research in language testing. In C, Elder, et al (Eds.). *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 150- 161). Cambridge, UK: Cambridge University Press; UCLES.

Alderson, J. C., & Buck, G. (1993). Standards in testing: a study of the practice of UK examination boards in EFL/ESL testing. *Language Testing* 10 (1) 1-26. From http://ltj.sagepub.com/cgi/content/abstract/10/1/1

Alderson, J. C. & Hughes, A. (1981). (Eds.) *Issues in language testing.* London: The British Council.

Alderson, J. C. & Urquhart, A. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing, 2,* 192-204. From http://ltj.sagepub.com/cgi/content/abstract/2/2/192

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics,* 14 (2), 115-129. Retrieved from http://applij.oxfordjournals.org/content/14/2/115.full.pdf+html

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

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, *1*(5). From http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

——— . (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

——— . (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association..

——— . (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.

——— . (2006). *Publication manual of the American psychological association* (6th ed.). Washington, DC: Author.

Anastasi, A. (1954). Psychological testing. New York: Macmillan.

——— . (1985). Some emerging trends in psychological measurement: A fifty-year perspective. *Applied psychological Measurement*,9 (2) pp.121-138. From http://conservancy.umn.edu/bitstream/handle/11299/102069/1/v09n2p121.pdf

——— . (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37* (1) 15.http://www.annualreviews.org/doi/pdf/10.1146/annurev.ps.37.020186.000245

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.

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

——— . What does language testing have to offer? *TESOL Quarterly*, 25 (4), 671-704. From http://www.jstor.org.www.sndl1.arn.dz/stable/pdfplus/3587082.pdf.

——— . Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing*, 17 (1) 1–42. From ltj.sagepub.com/content/17/1/1.abstract.

——— . (2001a). Some construct validity issues in interpreting scores from performance assessments of language ability. In R. Cooper., E. Shohamy., & J. Walters (Eds), *New perspectives and issues in educational language policy: a festschrift for* Bernard Dov Spolsky (pp. 63- 90). Amsterdam, The Netherlands: John Benjamins Publishing Co.

——— . (2001b). Designing and developing useful language tests. In C, A. Elder, et al (eds.). *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 109-116). Cambridge, UK: Cambridge University Press; UCLES.

——— . (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, *21*(3), 5–18. From http://onlinelibrary. wiley.com/doi/10.1111/j.1745-3992.2002.tb00095.x/pdf.

——— . (2004a). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

——— . (2004b). Linking Observations to Interpretations and Uses in TESOL Research. *TESOL Quarterly*, 38 (4), 723-728. From http://www.jstor.org. www.sndl1.arn.dz/stable /pdf plus/3587082.pdf.

——— . (2005) Building and Supporting a Case for Test Use, Language Assessment Quarterly, 2:1, 1-34. From http://www.tandfonline.com. www.sndl1.arn.dz/doi/pdf/10. 1207/s15434311 laq0201_2.

——— . (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville, C. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 165–207). Dordrecht, the Netherlands: John Benjamins.

——— . (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, & D. Bayliss (Eds.),*What are we measuring?Language testing reconsidered* (pp.41-71) Ottawa, Canada: University of Ottawa Press.

——— . (2013). How is educational measurement supposed to deal with test use?, *Measurement: Interdisciplinary Research and Perspectives.* 11 (1-2) 19-23. From http://www.tandfonline.com.www.sndl1.arn.dz/doi/pdf/10.1080/15366367.2013.784150.

Bachman, L. F., & Cohen, A. (Eds.). (1998). *Interfaces between second language acquisition and language testing research*. Cambridge, UK: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

——— . (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Purpura, J.E. (2008). Language Assessments: Gate-Keepers or Door-Openers? In B, Spolsky and M, H Francis (eds). *The Handbook of Educational Linguistics.*(pp.456-468). Oxford, UK: Blackwell Publishing Ltd

Bachman, L. F., & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal*, 70(4), 380-390 http://www.jstor.org. www.sndl1.arn.dz/stable/pdfplus/10.2307/327688.pdf

Bailey, M. K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing,* 13 (3) 258-.279. from http://ltj.sagepub.com/cgi/content/abstract/13/3/257

——— . (1999). *Washback in Language Testing*. TOEFL Monograph Series 15. Princeton, NJ: Educational Testing Service

Bailey, K. M., & Brown. J. D. (1995). Language testing courses: What are they? In A. Cumming & R. Berwick (Eds.), *Validation in language testing*. (pp. 236–256). Clevedon, UK: Multilingual Matters.

Basturkmen, H. (2003). Specificity and ESP course design. *Regional English Language Council Journal,* *34*(1), 48–63. from http://www.researchgate.net/publication/249768794_Specificity_and_Esp_Course_Design

——— . (2006). *Ideas and Options in English for Specific Purposes.* Mahwah, NJ: L. Erlbaum Associates.

——— . (2010). *Developing courses in English for specific purposes* . Basingstoke, UK : Palgrave/ Macmillan.

Basturkmen, H., & Elder, C. (2004). The practice of LSP. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 672–694). Oxford: Blackwell.

Bejar, I, I. (2011): A validity-based approach to quality control and assurance of automated scoring. *Assessment in Education: Principles, Policy & Practice*, 18 (3) 319-341. Retrieved http://dx.doi.org/10.1080/0969594X.2011.555329

Bejar, I, I., Williamson, M, D., & Mislevy, J. (2006). Human scoring. In M, D,Williamson, I, Bejar & R ,J Mislevy (Eds). *Automated scoring of complex tasks in computer-based testing* (pp.49-82). Mahwah, New Jersey/London:Lawrence Erlbaum Associates, Publishers.

Besnard, B., & Hunter, A. (2008). *Elements of Argumentation*. Massachusetts: The MIT Press

Braun, H., Bejar, I, I., & Williamson, M, D. (2006). Rule-based methods for automated scoring: Application in a licensing context. In M, D,Williamson, I, Bejar & R ,J Mislevy (Eds). *Automated scoring of complex tasks in computer-based testing* (pp.83-122). Mahwah, New Jersey/London:Lawrence Erlbaum Associates, Publishers.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice, 16*(4), 14–20. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1997.tb00604.x/ pdf

———. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

———. (2010). Generalizability Theory and Classical Test Theory, *Applied Measurement in Education,* 24 (1) 1-21. From http://www.tandfonline. com.www.sndl1.arn.dz/doi/pdf/10.1080/08957347. 2011.532417

———. (2013). Commentary on "Validating the interpretations and uses of test scores". *Journal of Educational Measurement*, *50*, (1), *74–83*. From http://onlinelibrary. wiley.com/doi/10.1111/jedm.12001/pdf

Brindley, G. (1998). Describing language development?: Rating scales and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112–140). Cambridge: Cambridge University Press.

Brown, A. (1995). The effect of rater variables in the development of an occupation specific language performance test. *Language Testing,* 12 (1) 16-33. From http://ltj.sagepub.com/cgi/content/abstract/12/1/1

———. (2012). Interlocutor and rater training. In G. Fulcher., & F. Davidson (Eds), *The Routledge Handbook of Language Testing* (pp. 413- 425). Abingdon, Oxon: Routledge.

Brown, J. D. (2003). *Language assessment principles and classroom practice.* Longman.

———. (2012). Classical test theory. In G. Fulcher., & F. Davidson (Eds), *The Routledge Handbook of Language Testing* (pp. 323- 335). Abingdon, Oxon: Routledge.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press; UCLES.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). London, UK: Longman.

——— . (1983). On some dimensions of language proficiency. In J. W. Oller, Jr. (ed.), *Issues in language testing research.* Rowley, MA: Newbury House.

——— . (1984). A communicative approach to language proficiency assessment in a minority setting. In C. Rivera (Ed), *Communicative competence approaches to language proficiency assessment: research and application* (pp.107-122). Clevedon, Avon, England: Multilingual Matters LTD

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47. From http://applij.oxfordjournals.org/content/I/1/1.full.pdf+html

Carroll, B. J. (1961). Fundamental considerations in testing for English Language proficiency of foreign students. In *CAL Testing the English proficiency of foreign students* (pp. 30–40). Washington, DC: Center for Applied Linguistics.

——— . (1968). The psychology of language testing. In A. Davies (ed.), *Language testing symposium: A psycholinguistic approach* (pp. 46-69). London: Oxford University Press.

——— . (1980). *Testing communicative performance*. London: Pergamon Institute of English

Cartier, F, A. (1968). Criterion-Referenced Testing of Language Skills. *TESOL Quarterly*, 2(1), 27-32. http://www.jstor.org. www.sndl1.arn.dz/ stable/pdfplus/10.2307/3585439. pdf

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing 14* (1) 3–22. http://ltj.sagepub.com/cgi/content/abstract/14/1/3 DOI: 10.1177/026553229701400102

——— . (2001). Language testing and technology: Past and future. *Language Learning and Technology, 5*, 95–98. http://llt.msu.edu/vol5num2/deville/default.html

——— . (2002). Technology in standardized language assessments. In R. Kaplan (Ed.), *Oxford handbook of applied linguistics*(471–484). Oxford: Oxford University Press.

——— . (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383. From http://ltj.sagepub.com/cgi/content/refs/20/4/369

Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics* 19, 273-299.

——— . (2005). A look back at and forward to what language testers measure. In H. Hinkel (ed.), *Handbook of research in second language teaching and learning.* (pp. 815-831) Mahwah, NJ: london, UK: Lawrence Erlbaum Associates Publishers.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Second language acquisition and language testing interfaces* (pp. 32–70). Cambridge: CUP.

——— . (1999). Validation in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272. From http://dx.doi.org/10.1017/S0267190599190147

——— . (2005).Computer Assisted Language Learning . In H. Hinkel (ed.), *Handbook of research in second language teaching and learning.* (pp743-755) . Mahwah, NJ; london, UK: Lawrence Erlbaum Associates Publishers.

——— . (2008). Utilizing technology in language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., vol. 7, pp. 301–317). New York, NY : Springer Science & Business Media.

——— . (2010). *Technology in language testing* [video]. Retrieved from http://languagetesting.info/video/main.html

——— . (2012a). Conceptions of validity. In G. Fulcher., & F. Davidson (Eds), *The Routledge Handbook of Language Testing* (pp.21-3). Abingdon, Oxon: Routledge.

——— . (2012b). Validity argument for language assessment: The framework is simple…. *Language Testing*. 29(1) 19–27. From http://ltj.sagepub.com.www.sndl1.arn.dz/content/29/1/19.full. Pdf+html

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language*. New York, NY: Routledge.

——— . (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2009.00165.x/pdf

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing,* 20 (4) 409–439. From http://ltj.sagepub.com/cgi/content/abstract/20/4/409

Cheng , L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., vol. 7, pp. 349–364.). New York, NY : Springer Science & Business Media.

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In R. Jones & B. Spolsky (eds.), *Testing language proficiency* (pp. 10-24). Arlington, VA: Center for Applied Linguistics.

Cohen, L., Manion, L., & Morrison, K. (2007) *Research Methods in Education* (7th ed). London: Routledge.

Commission Nationale des Programmes. (2005). *Programme D'anglais: Deuxième Langue Etrangère (Première Année Secondaire).* Alger: Direction de L'enseignement Secondaire.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

——— . (1980). Validity on parole: how can we go straight? In *New directions for testing and measurement* (Vol. 5, pp. 99–108). San Francisco: Jossey-Bass.

——— . (1984). *Essentials of psychological testing*. (4th ed). New York: Harper and Row.

——— . (1988). Five perspectives on validity argument. In H.Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

——— . (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302. From http://psycnet.apa.org/psycinfo/1956-03730-001

Crooks, T., Kane, M., & Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education*, *3*, 265–285. From http://www.tandfonline.com/doi/pdf/10.1080/0969594960030302

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing* 7 (1) 31-51. From http://ltj.sagepub.com/cgi/content/abstract/7/1/31

Cziko, G. A. (1982). Improving the psychometric, criterion-referenced, and practical qualities of integrative language tests. TESOL Quarterly, 16 (3), 367-379. From http://www.jstor.org/stable/3586636

Davidson, F. (2004). The Identity of Language Testing, *Language Assessment Quarterly*, 1 (1), 85-88. From http://www.tandfonline. com.www.sndl1. arn.dz/doi/pdf/10. 1207/s15434311laq0101_9

——— . (2012). Test specifications and criterion referenced assessment. In G. Fulcher., & F. Davidson (Eds), *The Routledge Handbook of Language Testing* (pp. 197- 207). Abingdon, Oxon: Routledge.

Davidson, F., & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT and London: Yale University Press

Davies, A. (1990). *Principles of language testing*. Oxford: Oxford University Press.

———— . (2001). The logic of testing Languages for Specific Purposes. *Language Testing,* 18 (2) 133–147. From http://ltj.sagepub.com/cgi/content/abstract/18/2/133

———— . (2003). Three heresies of language testing research. *Language Testing* 20 (4) 355–368. From http://ltj.sagepub.com/cgi/content/abstract/20/4/355

———— . (2008). Textbook trends in teaching language testing. *Language Testing*, 25 (3) 327–347. From http://ltj.sagepub.com/cgi/content/abstract/25/3/327

———— . (2007). Assessing Academic English Language Proficiency: 40+ years of U.K. Language Tests 73. In J. Fox, et al (Eds.), *Language Testing Reconsidered* (pp.73-86), Ottawa, Canada: University of Ottawa Press

———— . (2008). *Assessing academic English*. Cambridge, UK: Cambridge University Press

———— . (2010).Test fairness: A response. *Language Testing*. 27(2) 171–176. From http://ltj.sagepub.com.www.sndl1.arn.dz/content/27/2/171.full.pdf+html

———— . (2012a). Ethical codes and unexpected consequences. In G. Fulcher., & F. Davidson (Eds), *The Routledge Handbook of Language Testing* (pp. 455- 468). Abingdon, Oxon: Routledge.

———— .(2012b). Kane, validity and soundness. *Language Testing*. 29(1) 37– 42. From http://ltj.sagepub.com.www.sndl1.arn.dz/content/29/1/37.full. Pdf+html

Davies, A., & Elder, C. (2005). 'Validity and validation in language testing', in E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 795–813). Mahwah, NJ: Lawrence Erlbaum.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (eds.). (1999). *Dictionary of Language Testing*: *Studies in Language Testing, Vol 7*. Cambridge: Cambridge University Press; UCLES

Department of General Secondary Education. (2004). *Syllabuses for English: 1st, 2nd and 3rd Years Literary, Scientific and Technology Streams.* Alger: ONED

Department of Technical Secondary Education. (1995).*Syllabuses for English: Technical Secondary Schools 2nd and 3rd Years technical Streams.* Alger: ONPS

Direction Des Enseignements.(1983). *Programme Anglais*. Alger: IPN.

Dörnyei, Z. (1995). On the Teachability of Communication Strategies. *TESOL quarterly* 29(1), 55-85. From http://www.jstor.org/stable/3587805

——— . (2005). *The psychology of the language learner: Individual differences in second language acquisition* Mahwah, N J & London: Lawrence Erlbaum Associates, Inc.

Dörnyei, Z., & Scott, L, M. (1997). Communication strategies in a second language: Definitions and taxonomies. *Language Learning*, 47 (1), 173-210. From http://onlinelibrary.wiley.com/doi/10.1111/0023-8333.51997005/pdf

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

——— . (2001a). Three problems in testing language for specific purposes: Authenticity, specificity and inseparability. In C Elder et al (Eds). *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 45-51). Cambridge: Cambridge University Press.

——— . (2001b). Language for specific purposes assessment criteria: where do they come from*? Language Testing,* 18 (2) 171–185. From http://ltj.sagepub.com/cgi/content/abstract/18/2/171

——— . (2005). Testing Language for Specific Purposes. (2005).In H. Hinkel (ed.), *Handbook of research in second language teaching and learning.* (pp.857-868). Mahwah, NJ: London,, UK: Lawrence Erlbaum Associates Publishers.

——— . (2010a). *Understanding Language Testing.* London: Hodder Education.

——— .(2010b). This won't hurt a bit: Assessing English for nursing. *Taiwan International ESP Journal,2(2),1-16.* From http://tiespj.tespa.org.tw/this-wont-hurt-a-bit-assessing-english-for-nursing/

——— . (2013). ESP and assessment. In B, Paltridge ., & Starfield, S. Eds. *The handbook of English for specific purposes*, (pp. 367-383). England, Chichester: John Wiley & Sons, Inc

Downing, S, M. (2006). Selected-response item formats in test development. In S. M, Downing., & T, M, Haladyna (Eds), *Handbook of test development* (pp.287-301). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*(1), 61–82. From http://www.tandfonline.com/doi/pdf/10.1207/s15324818ame1001_4

Dudley - Evans , T. (1997). Five Questions for LSP Teacher Training. In R. Howard & G. Brown (Eds.), *Teacher Education for Languages for Specific Purposes* (pp. 58–67). Philadelphia: Multilingual Matters Ltd

Dudley-Evans, T., & St. John, M. (1998). *Developments in English for Specific Purposes.* Cambridge: Cambridge University Press.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing Original Research Article. *Assessing Writing,*14 (2) 88-115 http://ac.els-cdn.com/S1075293509000130/1-s2.0-S1075293509000130-main.pdf?

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (7th ed). Englewood Cliffs, NJ: Prentice Hall.

Ebel, R. L., & Popham, W,J. (1978). The 1978 Annual Meeting Presidential Debate. *Educational Researcher,*7, 11, 3-10. From http://www.jstor.org/stable/1175378

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*. 25 (2) 155–185. From http://ltj.sagepub.com.www.sndl1.arn.dz/content/25/2/155.full.pdf+html

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, 14 (3) 261–277. Retrieved from http://ltj.sagepub.com/cgi/content/abstract/14/3/261

——— . (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing*, 18 (2) 149–170. From http://ltj.sagepub.com/cgi/content/abstract/18/2/149

Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumley, T., et al. (Eds.). (2001). *Experimenting with uncertainty: Essays in honour of Alan Davies*. Cambridge: Cambridge University Press.

Ellis, Rod. (2005). Planning and task performance in a second language. *Language learning and language teaching,* 11. Amsterdam, Philadelphia, pa : John Benjamins Publishing Company Exeter.

Finnegan, R. (2006). Using documents. In J Sapsford., & V, Jupp (Eds). *Data collection and analysis* (2nd ed, pp.93-123). London: SAGE Publications; The Open university.

Fulcher, G. (1997),. An English language placement test issues in reliability and validity. *Language Testing*, 14 (2) 113–138. From http://ltj.sagepub.com/cgi/content/abstract/14/2/113

——— . (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics* 20(1), 221–36. From http://applij.oxfordjournals.org/content/20/2/221.full.pdf+html

——— . (2000). The "communicative" legacy in language testing. *System*, *28*, 483–497. Retrieved from www.languagetesting.info/articles/store/FulcherCLT.pdf

——— . (2003a). *Testing second language speaking*. London: Longman; Pearson Education.

——— . (2003b). Interface design in computer-based. *Language Testing 2003 20 (4) 384–408.*From http://ltj.sagepub.com/cgi/content/abstract/14/2/113

——— . (2010). *Practical language testing*. London, UK: Hodder Education.

——— . (2012).Scoring performance tests. In G. Fulcher., & F. Davidson (Eds), *The Routledge Handbook of Language Testing* (pp. 378- 392). Abingdon, Oxon: Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and Assessment. An advanced resource book*. London: Routledge.

——— . (2009). Test architecture, test retrofit. *Language Testing,* 26 (1) 123–144. http://ltj.sagepub.com/cgi/content/abstract/26/1/123

——— . (Eds). (2012). *The Routledge Handbook of Language Testing* (pp. 455- 468). Abingdon, Oxon: Routledge.

Given, L. M. (2008).*The Sage encyclopedia of qualitative research methods* (Vols. 1–2). London, UK: Sage Publications, Inc.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519-521. Retrieved from http://psycnet.apa.org/psycinfo/1964-047004-001

Glaser, R., & Cox, R. C. (1968). Criterion-referenced testing for the measurement of educational outcomes. In R. Weisgerber (Ed.), *Instructional process and media innovation* (pp. 545-550). Chicago: Rand-McNally.

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), *Psychological principles in system development* (pp. 419-474). New York: Holt, Rinehart & Winston.

Glaser, R., & Nitko, J.(1970). *Measurement in learning and instruction: Document resumé.* Washington, D.C: Office of Naval Research, Psychological Sciences Div.

Glendinning, E, H., & Glendinning, N. (2008). *Oxford English for electrical and mechanical engineering.* Oxford: Oxford University Press.

Green, D, R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice*, *17*(2), 16–19. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1998.tb00828.x/pdf

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. B., & Reckase, M. D. (1984). Technical Guidelines for Assessing Computerized Adaptive Tests. *Journal of Educational Measurement,*21 (4) 347-360. From http://www.jstor.org/stable/1434586 .

Gronlund, N. E. (1977). *Constructing achievement tests* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Haertel, E. H. (1999).Validity arguments for high-stakes testing: in search of the evidence. *Educational Measurement: Issues and Practice, 18*(4) 5–9. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1999.tb00276.x/pdf

———— . (2006). Reliability. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger Publishers.

———— . (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives*, 11 (1-2) 1-18. From http://dx.doi.org/10.1080/15366367.2013.783752

Haig, B, D. (2012). From construct validity to theory validation. *Measurement: Interdisciplinary Research and Perspectives*, 10 (1-2) 59-62. From http://www.tandfonline.com.www.sndl1.arn.dz/ doi/pdf/10.1080/15366367. 2012. 681975

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd Ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance: A threat in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2004.tb00149.x/pdf

Halliday, M. A. K. (1973). *Explorations in the Functions of Language.* London: Edward Arnold.

———— . (1975). *Learning How to Mean.* London: Edward Arnold.

———— . (1994). *An introduction to functional grammar* (2nd ed). London: Edward Arnold.

Halliday, M.A.K. (2002). *On grammar*. London: Continuum

———— . (2003). On Language and Linguistics: London: Continuum

———— . (2004) *An introduction to functional grammar* (3rd ed). Oxford: Oxford University Press

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London: Longman

———— . (1985). *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective.* Geelong, Vic.: Deakin University Press.

Halliday, M.A.K., & Martin, J.R. (1993). *Writing science: Literacy and discursive power*. London: The Falmer Press,

Halliday, M.A.K., & Webster, J, J. (Eds). (2009). *Continuum Companion to Systemic Functional Linguistics*. London: Continuum

Halliday, M. A. K., McIntosh, A., & Strevens. P. (1964). *The Linguistic Sciences and Language Teaching*. (Longmans' Linguistic Library.) London: Longmans.

Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: Practice, theory and research*. Cresskill, NJ: Hampton Press.

Hamp-Lyons, L., & Lumley, T. (2001). Assessing language for specific purposes. *Language Testing* 18 (2) 127–132. From http://ltj.sagepub.com/cgi/content/refs/18/2/127

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, 3(1), 49–68. From www.sciencedirect.com/science/article/pii/1060374394900051

Harding, L. (2014). Communicative language testing: Current issues and future research. Language *Assessment Quarterly*, 11(2)186-197. From http://dx.doi.org/10.1080/15434303.2014.895829p.2

Harris, D. (1969). *Testing English as a second language*. New York: McGraw Hill.

Heaton, B. (1975). *Writing English language tests*. London: Longman.

Henning, G. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.

Hinkel, E. (Ed.). (2005). *Handbook of research in second language teaching and learning*. Mahwah, NJ: Lawrence Erlbaum.

Hitchcock, D. (2006). Good reasoning on the Toulmin model. In D. Hitchcock & B. Verheij (Eds). *Arguing on the Toulmin model: New essays in argument analysis and evaluation,* (pp. 203-218). Dordrecht, the Netherlands: Springer

Hudson, T. (1991). Relationships among IRT item discrimination and item fit indices in criterion-referenced language testing. *Language Testing,* 8 (2) 160-181. From http://ltj.sagepub.com/cgi/content/abstract/8/2/160

Hudson, T. (1993). Surrogate indices for item information functions in criterion-referenced language testing. *Language Testing,* 10 (2) 171-191. From http://ltj.sagepub.com/cgi/content/abstract/8/2/160

Hudson, T., Lynch, B. (1984). A criterion-referenced measurement approach to ESL achievement testing. *Language Testing.*1(1)171-201. from http://ltj.sagepub.com/cgi/content/abstract/1/2/171

Hughes, A. (1987). *Testing for language teachers*. Cambridge: Cambridge University Press.

———. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Hutchinson, T., & Waters, A. (1987). *English for Specific Purposes: A learning-centered approach.* Cambridge: Cambridge University Press.

Hyland, K. (2006) *English for academic purposes: An advanced resource book* (London: Routledge).

Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Harmandsworth, UK: Penguin.

Ingram, E. (1977). Basic concepts in testing. In J. P. B. Allen & A. Davies (Eds.), *Testing and experimental methods: Edinburgh course in applied linguistics* (vol. 4., pp. 11–37). London: Oxford University Press.

Johnson, R., Penny, J., & Gordon, B. (2000). The relationship between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education,* 13 (2), 121–138. From http://www.tandfonline.com/doi/pdf/10.1207/S15324818AME1302_1

——— . (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication*, 18 (2), 229–249. From http://wcx.sagepub.com/content/18/2/229.full.pdf+html

——— . (2009). *Assessing performance: Designing, scoring, and validating performance.* New York, NY: The Guilford Press.

Johnson, R., Penny, J., Fisher, S., & Kuhs, T. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education*, 16 (4), 299–322. http://www.tandfonline.com/doi/pdf/10.1207/S15324818AME1604_3

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S, P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly,* 2(2), 117–146. From http://www.tandfonline.com/doi/pdf/10.1207/s15434311laq0202_2

Jupp, V. (2006a). Documents and Critical Research. In J Sapsford., & V, Jupp (Eds). *Data collection and analysis* (2nd ed, pp.272-290). London: SAGE Publications; The Open university.

——— . (Ed.). (2006b). *The Sage dictionary of social research methods*. London, Thousand Oaks New Delhi: Sage Publications.

Kane, M. (1986).The role of reliability in criterion-referenced tests. *Journal of Educational Measurement,* 23 (3) 221-224. From http://www.jstor.org/stable/1434609

——— . (1992). An argument-based approach to validation. *Psychological Bulletin*, *112*, 527–535.

——— . (1996). The precision of measurements. *Applied Measurement in Education*, *9*, 355– 379.

——— . (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2001.tb01130.x/pdf

——— . (2002). Validating high stakes testing programs. *Educational measurement: Issues and practice*, 21(1), 31-41. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2002.tb00083.x/pdf

——— . (2004a). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 135–170. From http://www.tandfonline.com/doi/pdf/10.1207/s15366359mea0203_1

——— . (2004b). The analysis of interpretive arguments: some observations inspired by the comments. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 192–200. http://www.tandfonline.com.www.sndl1.arn.dz/doi/pdf/10.1207/s15366359mea0203_3

——— . (2006). Content-related validity evidence in test development. In S. M, Downing & T, M, Haladyna (Eds), *Handbook of test development* (pp.131-153). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

——— . (2007). Validating measures of Mathematical Knowledge for Teaching. Measurement: *Interdisciplinary Research and Perspectives.*5 (2-3)180-187. From http://www.tandfonline. com.www.sndl1.arn.dz/doi/pdf/10.1080/15366360701492807

——— . (2010). Validity and fairness. *Language Testing,* 27(2) 177–182. http://ltj.sagepub.com.www.sndl1.arn.dz/content/27/2/177.full.pdf+html

——— . (2012a).Articulating a validity argument. In G. Fulcher., & F. Davidson (Eds), *The Routledge Handbook of Language Testing* (pp.34-47). Abingdon, Oxon: Routledge.

——— . (2012b). Validating score interpretations and uses. *Language Testing*. 29(1) 3-17. From http://ltj.sagepub.com. www.sndl1.arn.dz/content/29/1/3.full. Pdf+html

——— . (2013).Validating the interpretations and uses of test scores. *Journal of Educational Measurement,* 50 (1) 1–73. From http://www.tandfonline.com.www.sndl1. arn.dz/doi/pdf/10.1080/15366367. 2013.857208

Kane, M., & Tannenbaum, R. J. (2013). The role of construct maps in standard setting. *Measurement: Interdisciplinary Research and Perspectives*, 11 (4) 177-180.Fromhttp://www.tandfonline.com.www.sndl1.arn.dz/doi/pdf/10.1080/15366367.2013.857208

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1999.tb00010.x/pdf

Kempf-Leonard, K. (Ed). (2005). *Encyclopedia of Social Measurement* (Vol.1). Amsterdam, the Netherlands: Elsevier Science Publishing Co Inc

Kerlinger, F. N. (1973). *Foundations of behavioral research (*2nd ed.). New York: Holt, Reinhart and Winston.

——— . (1986). *Foundations of behavioral research (*3rd ed.). New York: Holt, Reinhart and Winston.

Kerlinger, F. N., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research.* New York: Holt, Rinehart and Winston.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing,* 26 (2) 275–304. Retrieved from http://ltj.sagepub.com/cgi/content/refs/26/2/275

——— . (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? Original Research Article. *Assessing Writing*, 16 () 81–96.

Kothari, C. R. (2008). *Research methodology: Methods and techniques* (2nd ed.). New Delhi: New Age International Publishers.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, 70(4), 366-372. http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4781.1986.tb05291.x/pdf

Kunnan A, J. (2005). Language assessment from a wider context . In H. Hinkel (ed.), *Handbook of research in second language teaching and learning* (pp. 779- 794) .Mahwah, NJ; London, UK: Lawrence Erlbaum Associates Publishers.

——— . (1999). Language Testing: Fundamentals. In B. Spolsky (Ed.), *Concise Encyclopedia Of Educational Linguistics* (pp. 707- 714). Oxford, UK: Elsevier Science Ltd

——— . Regarding language assessment. *Language Assessment Quarterly,* 1(1), 1.http://www.tandfonline.com.www.sndl1.arn.dz/doi/pdf/10.1207/s15434311laq0101_2

——— . (2010). Test fairness and Toulmin's argument structure. *Language Testing,* 27(2) 183–189. http://ltj.sagepub.com.www.sndl1.arn.dz/ content/27/2/183.full.pdf+ html

——— . (2008). Large Scale Language Assessments. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., vol. 7, pp. 135–155.). New York, NY : Springer Science & Business Media.

——— . (Ed.). (1998). *Validation in language assessment*. Mahwah, NJ: Erlbaum.

——— . (Ed.). (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language testing research colloquium, Orlando, Florida*. Cambridge, UK: Cambridge University Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book*. New York: McGraw Hill.

Li, H. (2003). The Resolution of Some Paradoxes Related to Reliability and Validity. *Journal of Educational and Behavioral Statistics.* 28 (2). 89–95. http://www.jstor.org.www.sndl1.arn.dz/stable/pdf/10.2307/3701256.pdf

Lindvall, C. M., & Nitko, A. J. (1969). *Criterion-referenced testing and the individualization of education: Document Resume.* Washington, D.C: Bureau of Research. From http://files.eric.ed.gov/fulltext/ED036167.pdf

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, *16*(2), 14–16. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1997.tb00587.x/pdf

——— . (1998a). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice*, *17*(2), 28–30.

——— . (1998b). Validating inferences from national assessment of educational progress achievement-level reporting. *Applied Measurement in Education*, 11 (1) 23-47 http://dx.doi.org/10.1207/s15324818ame1101_2

———. (2006). The standards for educational and psychological testing. In S. M, Downing., & T, M, Haladyna (Eds), *Handbook of test development* (pp.27-38). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

——— . (2009). The concept of validity in the context of NCLB. In R. Lissitz (Ed.), *The concept of validity* (pp. 195–212). Charlotte, NC: Information Age Publishers.

Linn, R. L., & Gronlund, N. E. (1995). *Measuring and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Livingston, S. A. (2006). Item Analysis. In S. M, Downing., & T, M, Haladyna (Eds), *Handbook of test development,* (pp. 424-441). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Long, M, H. (Ed). (2005). *Second language needs analysis*. Cambridge: Cambridge University Press.

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. From http://ltj.sagepub.com/cgi/content/abstract/19/3/246

———. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Lumely, T., & Brown, A. (2005). Research Methods in Language Testing. In H. Hinkel (ed.), *Handbook of research in second language teaching and learning* (pp. 833-855) Mahwah, NJ: London,, UK: Lawrence Erlbaum Associates Publishers.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing,* 12 (1) 54-71. From http://ltj.sagepub.com/cgi/content/abstract/12/1/54

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Lynch, B. K. (1997). In search of the ethical test. *Language Testing*, 14 (3) 315–327. From http://ltj.sagepub.com/cgi/content/abstract/14/3/315

——— . (2001). The ethical potential of alternative language assessment. In C, Elder, et al (eds.). *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 228-239). Cambridge, UK: Cambridge University Press; UCLES

Lynch, B, K., & Davidson, F. (1994). The Construct Validation of Some Components of Communicative Proficiency. *TESOL Quarterly*, 28 (4), 727-743. http://www.jstor.org. www.sndl1.arn.dz/ stable/pdfplus/3587557.pdf

Lynch, B. K., & Hamp-Lyons, L. (1999). Perspectives on research paradigms and validity: Tales from the Language Testing Research Colloquium. *Melbourne Papers in Language Testing,* 8(1), 57–93.

Madsen, H. (1983). *Techniques in testing*. New York and Oxford: Oxford University Press.

Malone, M. (2008). Training in language assessment In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., vol. 7, pp. 225–239.). New York, NY : Springer Science & Business Media.

Mason, E, J.(2007). Measurement issues in high stakes testing. *Journal of Applied School Psychology*, 23 (2) 27-46. http://www.tandfonline. com.www.sndl1.arn.dz/doi/pdf/10. 1300/J370 v23n02_03

McKay, P. (2006): *Assessing Young Language Learners*. Cambridge , UK; Cambridge University Press

McNamara, T. F. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing,* 7 (2) 52-76. From http://ltj.sagepub.com/cgi/content/abstract/7/2/52

——— . (1996). *Measuring second language performance*. Harlow, Essex: Pearson Education.

——— . (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18(4), 446–466.* From http://applij.oxfordjournals.org/content/18/4/446.full.pdf+html

——— . (1999). Language Testing: Users and Uses. In B. Spolsky (ed.), *Concise Encyclopedia Of Educational Linguistics* (pp. 724-728). Oxford, UK: Elsevier Science Ltd

——— . (2000). *Language testing*. Oxford: Oxford University Press.

——— . (2001). Language assessment as social practice: challenges for research. *Language Testing,* 18 (4) 333-349. http://ltj.sagepub.com/cgi/content/abstract/18/4/333

——— . (2004). Language Testing. In A. Davies & C. Elder (Eds.), *The Handbook of Applied Linguistics*, (pp.763- 783), Malden, MA, Oxford UK, Carlton, Victoria: Blackwell Publishing Ltd

——— . (2005). Second language testing and assessment: Introduction. In E. Hinkel (Ed.). *Handbook of research in second language teaching and learning*. (pp 775–778). Mahwah: Lawrence Erlbaum Associates.

——— . (2006a). Validity and values: Inferences and generalizability in language testing. In M, Chalhoub-Deville,. C, A Chapelle,.& P, Duff (eds). *Inference and Generalizability in Applied Linguistics: Multiple perspectives.* (27-45). Amsterdam, The Netherlands, Philadelphia pa: Benjamins Publishing Co.

——— . (2006b). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31–51. From http://dx.doi.org/10.1207/s15434311laq0301_3

——— . (2007). Language Testing: A Question of Context. In J. Fox, M. Wesche,, D. Bayliss, L. Cheng, E Carolyn. & C D, Turner (eds.), *Language Testing Reconsidered* (pp.131- 131). Ottawa, Canada: University of Ottawa Press.

——— . (2008). The socio-political and power dimensions of tests In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., vol. 7, pp. 415–427). New York, NY : Springer Science & Business Media.

——— . (2009). Language tests and social policy: A commentary . In G, Hogan-Brun., C. Mar-Molinero., & P. Stevenson (eds). *Discourses on Language and Integration: Critical perspectives on language testing regimes in Europe*. (pp. 153-163). Amsterdam ,The Netherlands/ Philadelphia, pa : John Benjamins Publishing Co.

——— . (2011). Applied linguistics and measurement: A dialogue. *Language Testing* 28(4) 435–440. From http://ltj.sagepub.com. www.sndl1.arn.dz/content/28/4/435. full.pdf+html

——— . (2014). 30 years on—evolution or revolution? *Language Assessment Quarterly,*11 (2) 226-232.http://dx.doi.org/10.1080/15434303.2014.895830

McNamara, T., & Roever, C. (2006). *Language Testing: The social dimension*. Oxford: Blackwell Publishing.

McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics*, *18*, 89–95. http://onlinelibrary.wiley.com/doi/10.1111/j.1473-4192.2008.00191.x/pdf

McNamara, T ., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics,* 22, 221–242. From http://dx.doi.org/10.1017/S0267190502000120

McNamara, T., Knoch, U., & Davies, A. (2012). The Rasch wars: The emergence of Rasch measurement in language testing *Language Testing*. 29(1) 37– 42. From http://ltj.sagepub.com .www.sndl1.arn.dz/content/29/1/37.full. Pdf+html

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027.

——— . (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, *10*(9), 9–20. From http://www.jstor.org.www.sndl1.arn.dz/stable/pdf/1174731.pdf

——— . (1982). The values of ability testing: Implications of multiple perspectives about criteria and standards. *Educational Measurement: Issues and Practice*, *1*(3), 9–12. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1982.tb00660.x/pdf

——— . (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215-237. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1984.tb01030.x/pdf

——— . (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

——— . (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington DC: The American Council on Education and the National Council on Measurement in Education.

——— . (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*, 13–23. From http://www.jstor.org.www.sndl1.arn.dz/stable/pdf/1176219.pdf

——— . (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50 (9)741-749

——— . (1996). Validity and washback in language testing. *Language Testing* 13 (3) 241-256. From http://ltj.sagepub.com/cgi/content/abstract/13/3/241

Ministry of National Education. (1987-88).*Think it over: An Algerian course book: Student's Book 3 A.S.* Alger: IPN

——— . (1988-89*). New Skills: English for Science and Technology 2 AS Pupil's Book*. Alger: IPN

——— . (1988-89).*New Lines: Learn English With Us. An Algerian Course book. Book III.1 AS* .Alger: IPN

——— . (1991). *Modern World: English for Science and Technology 3 A.S.Pupil's Book.* Alger: ENAG

——— . (1992). *Syllabuses for English.* Alger: Department of Secondary Education.

——— . (1995). *Syllabuses for English: Technical secondary schools 2nd and 3rd years technical streams.* Alger: Department of Technical Secondary Education

——— . (1997). *Comet: A communicative English Teaching Course book for all streams.* Alger: IPN

——— . (1998). *The New Midlines: English for 2 AS Pupils.* Alger: ONPS

——— . (1998-1999). *My New Book of English: 1 AS.* Alger: IPN

——— . (1988-1999).*Midlines. An Algerian Course book. Book IV 2 AS. Pupil's Book.* Alger: IPN

Mislevy, R. J. (2003). Argument substance and argument structure in educational assessment. *Law, Probability and Risk, 2*(4), 237–258.

——— . (2004). Can There Be Reliability without "Reliability?" *Journal of Educational and Behavioral Statistics.* 29 (2). 241–244. From http://www.jstor.org.www.sndl1.arn.dz/stable/pdf/3701267.pdf

——— . (2007).Validity by Design. *Educational Researcher*, 36 (8) 463-469. From http://www.jstor.org.www.sndl1.arn.dz/stable/pdf/4621101.pdf.

——— . (2012). The Case for informal argument. *Measurement: interdisciplinary research and perspectives,* 10 (1-2) 93-96. From http://www.tandfonline.com.www.sndl1.arn.dz/doi/pdf/10.1080/15366367.2012.682525

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). Mahwah, NJ: Erlbaum

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496. From http://ltj.sagepub.com/cgi/content/refs/19/4/477

——— . (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1 (1) 3-62. From http://dx.doi.org/10.1207/S15366359MEA0101_02

Moller, A. (1981). Reaction to the Morrow paper. In J. C. Alderson & A. Hughes (Eds.), *Issues in Language testing: ELT documents 111* (pp. 39-45). London: The British Council

Morrow, K. (1981). Communicative language testing: Revolution or evolution? In J. C. Alderson & A. Hughes (Eds.), *Issues in language testing: ELT documents 111*(pp. 9–25). London. London, UK: The British Council.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23(2), 5–12. From http://www.jstor.org.www.sndl1.arn.dz/stable/pdf/1176218.pdf.

——— . (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, *4*(2), 5–13. http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1995.tb00854.x/pdf

——— . (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, *17*(2), 6–12. http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1998.tb00826.x/pdf

——— . (2007). Reconstructing validity. *Educational Researcher*, *36*, 470–476. From http://www.jstor.org.www.sndl1.arn.dz/stable/pdf/4621102.pdf.

——— . (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement*, *50*, (1), 91–98. From http://dx.doi.org/10.1207/S15366359MEA0101_02

Mumby, J. (1978). *Communicative syllabus design*. Cambridge, UK: Cambridge University Press.

Naoua, M.(2006). *An analysis of some factors leading to underachievement in English as a foreign language for secondary school pupils: The case of Technology streams in Eloued* (Unpublished magister dissertation). Mohamed Kheider University, Biskra, Algeria.

Nevo, N. (1989).Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing*, 6, 199-215. From http://ltj.sagepub.com/cgi/content/refs/6/2/199

Newton, P, E.(2012). Clarifying the consensus definition of validity, *Measurement: Interdisciplinary Research and Perspectives,* 10 (1-2) 1-29. From http://www.tandfonline.com.www.sndl1.arn.dz/doi/pdf/10.1080/15366367.2012.669666

——— . (2013). Two Kinds of Argument? *Journal of Educational Measurement*, *50*, (1), 105–109. http://onlinelibrary.wiley.com/doi/10.1111/jedm.12004/pdf

Noijion, J. (1994). Testing computer-assisted language testing: Towards a checklist for CALT. *CALICO Journal*, *12*(1), 37–58. From www.equinoxpub.com/journals/index.php/CALICO/article/.../19429

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.

——— . (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

——— . (1999). *Second language* teaching & *learning*. Boston, Massachusetts: Heinle & Heinle Publishers

———. (2004). *Task-based language teaching*. Cambridge: Cambridge University Press.

Oller, J.W., Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman. Palgrave Macmillan

———. (Ed). (1983). *Issues in language testing research* . Rowley, MA: Newbury House.

———. (2012). Grounding the argument-based framework for validating score interpretations and uses. *Language Testing*. 29(1) 29-36. From http://ltj.sagepub.com. www.sndl1.arn.dz/content/29/1/29.full. Pdf+html

Osterlind, S. J. (1990a). Establishing criteria for meritorious test items. *Educational Research Quarterly,* 14(3), 26. From http://www.researchgate.net/publication/232501238_Establishing_criteria_for_meritorious_test_items

———. (1990b). Toward a uniform definition of a test item. *Educational Research Quarterly,* 14(4), 2-5. From http://www.researchgate.net/publication/232501238_Establishing_criteria_for_meritorious_test_items

———. (2002). Constructing test items: Multiple-choice, constructed-response, performance, and other formats (2nd ed). New York, NY: Kluwer Academic Publishers.

Osterlind, S. J., & Everson, H. T. (2009). Differential item functioning (2nd ed.). Thousand Oaks, CA: Sage.

Paltridge, B., & Starfield, S. (2013). *The handbook of English for specific purposes*. (Eds). England, Chichester: John Wiley & Sons, Inc

Parkinson, J. (2013). English for Science and Technology. In B, Paltridge ., & Starfield, S. Eds. *The handbook of English for specific purposes*,(pp. 156-173). England, Chichester: John Wiley & Sons, Inc.

Pellegrino, J, W., Chudowsky, N,. & Glaser, R.(2001). *Knowing what students know the science and design of educational assessment.* Washington, DC: National Academy Press

Penny, J., Johnson, R, L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric Original Research Article. *Assessing Writing,*7 (2) 143-164. http://ac.els-cdn.com/S107529350000012X/1-s2.0-S107529350000012X-main.pdf

———. (2000b). Using rating augmentation to expand the scale of an analytic rubric. *Journal of Experimental Education*, 68, 269–287. Retrieved from http://www.tandfonline.com/doi/pdf/10.1080/00220970009600096

Popham, W. J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice Hall.

——— . (1988). Judgment-based teacher evaluation. In S. J. Stanley & W. J. Popham (Eds.), *Teacher evaluation: Six prescriptions for success* (pp. 56–77). Alexandria, VA: ASCD.

——— . (1997). Consequential validity: Right concern – Wrong concept. *Educational Measurement: Issues and Practice*, *16*(2), 9–13. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1997.tb00586.x/pdf.

——— . (2000). *Modern educational measurement: Practical guidelines for educational leaders*. Needham, MA: Allyn and Bacon.

——— . (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: ASCD.

——— . (2003).*Test better, teach better :The instructional role of assessment.* Alexandria, VA: ASCD.

——— . (2004). *America's **"Failing"** schools**:** How parents and teachers can cope with no child left behind**.*** New York, NY, London: Routledge Falmer.

——— . (2008). *Transformative assessment.* Alexandria, VA: ASCD.

——— . (2009). *Instruction that measures up: Successful teaching in the age of accountability.* Alexandria, Virginia USA: ASCD publications

——— . (2011a). *Classroom assessment: What teachers need to know* (6th ed.). Boston: Pearson

——— . (2011b). *Transformative assessment in action : An inside look at applying the process.*

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 6 (1) 1-9. From http://www.jstor.org/stable/1433917 .

Purpura, J. A. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

——— . (2008). Assessing communicative language ability: Models and their components. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd ed., vol. 7, pp. 53–68). New York, NY : Springer Science & Business Media.

Rassool, N. (1999). *Literacy for sustainable development in the age of information.* Clevedon Philadelphia, Toronto Sydney Johannesburg: Multilingual Matters Ltd.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Rea-Dickins, P. (2000). Assessment in early years language learning contexts. *Language Testing* 2000 17 (2) 115–122. http://ltj.sagepub.com.

Reed, D,. & Cohen, C. (2001). Revisiting raters and ratings in oral language assessment. In C, Elder. , A, et al (eds.). *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 82- 96). Cambridge, UK: Cambridge University Press; UCLES

Richards, J. C. & Schmidt, R. W. (2002*). Longman dictionary of language teaching and applied linguistics* (3rd ed.). London: Longman Group.

Ryan, K. (2002). Assessment validation in the context of high-stakes testing assessment. *Educational Measurement: Issues and Practice*, *21*(1), 7–15. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.2002.tb00080.x/pdf

Sapsford, J., & Jupp, V. (2006). *Data collection and analysis*. (Eds). 2nd Ed. London: SAGE Publications/The Open university.

Savignon, S. J. (1972). *Communicative competence: An experiment in foreign language teaching*. Philadelphia, PA: Center for Curriculum Development.

Savignon, S.J. 1983: Communicative competence: theory and classroom practice. Reading, MA: Addison-Wesley.

——— . (1985). Evaluation of communicative competence: The ACTFL provisional proficiency guidelines. The Modern Language Journal, 69(2), 129-134 http://www.jstor.org.www.sndl1.arn.dz/stable/pdfplus/10.2307/326817.pdf

——— . (1991). Communicative Language Teaching: State of the Art. *TESOL Quarterly*, 25(2), 261-277. http://www.jstor.org.www.sndl1.arn.dz/stable/pdfplus/10.2307/3587463.pdf

——— . (2002). *Interpreting communicative language teaching: Contexts and concerns in teacher education.* Yale: Yale University Press

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405–450). Washington, DC: American Educational Research Association.

——— . (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24. From http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1997.tb00585.x/pdf

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing,* 25 (4) 465–493. http://ltj.sagepub.com.www.sndl1.arn.dz/content/26/2/275.full.pdf+html

Shohamy, E.(1990). Discourse analysis in language testing. *Annual Review of Applied Linguistics* 11, 115–28. Retrieved from http://dx.doi.org/10.1017/S026719050000

——— . (1992). Beyond performance testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal*, *76*(4), 513–521. From http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4781.1992.tb05402.x/pdf

——— . (1993). *The power of tests: The impact of language tests on teaching and learning.* Washington, DC: The National Foreign Language Center at Johns Hopkins University.

——— . (1997). Testing methods, testing consequences: are they ethical? Are they fair? *Language Testing* 14 (3) 340–349. From http://ltj.sagepub.com/cgi/content/abstract/14/3/340

——— . (1999). Language testing impact . In B. Spolsky (ed.), *Concise Encyclopedia Of Educational Linguistics* (pp. 711-715). Oxford, UK: Elsevier Science Ltd

——— . (2001a). *The power of tests: A critical perspective on the uses of language tests.* Harlow: Longman/Pearson

——— . (2001b). The social responsibility of the language testers. In R, L.Cooper,. E, Shohamy., & J .Walters (eds*). New Perspectives and Issues in Educational Language Policy: A festschrift for Bernard Dov Spolsky.* (pp. 113-130), Amsterdam ,The Netherlands/ Philadelphia, pa: John Benjamins Publishing Co.

——— . (2006). *Language policy: Hidden agendas and new approaches*. London: Routledge.

——— . (2007). Language tests as language policy tools. *Assessment in Education: Principles, Policy & Practice,* 14 (1) 117-130 http://dx.doi.org/10.1080/09695940701272948

——— . (2008). Introduction to volume 7: Language testing and assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Volume 7. Language testing and assessment* (2nd ed. pp. xiii-xxii). New York: Springer Science & Business Media.

——— . (2009). Language tests for immigrants Why language? Why tests? Why citizenship?In G, Hogan-Brun., C. Mar-Molinero., & P. Stevenson (eds). *Discourses on Language and Integration: Critical perspectives on language testing regimes in Europe*. (pp. 45-59). Amsterdam ,The Netherlands, Philadelphia, pa : John Benjamins Publishing Co.

——— . (2013).The discourse of language testing as a tool for shaping national, global, and transnational identities. *Language and Intercultural Communication,*13 (2) 225-236 from http://dx.doi.org/10.1080/14708477.2013.770868

Shohamy, E., & McNamara, T. F. (2009): Language Tests for citizenship, immigration, and asylum. *Language Assessment Quarterly*, 6 (1) 1-5. From www.tandfonline.com/doi/pdf/10.1080/15434300802606440

Simon, G, B. (1969). Comments on "implications of criterion-referenced measurement". *TESOL Quarterly*, 6, 4, 259-260. From http://www.jstor.org/stable/1434027 .

Sireci, G. S. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, *50*, (1), 99–104. http://onlinelibrary.wiley.com/doi/10.1111/jedm.12005/pdf

Spolsky, B. (1968). Language testing: the problem of validation. *TESOL Quarterly, 2*, 88–94. From http://www.jstor.org/stable/3586083 .

——— . (1979). Introduction: Linguists and language testers. In B. Spolsky (ed.), *Approaches to language testing* (pp. v-x). Arlington, VA: Center for Applied Linguistics.

——— . (1981). Some ethical questions about language testing. In C. Klein-Braley & D. Stevenson (eds.), *Practices and problems in language testing* (pp. 5-30). Frankfurt: Peter Lang.

——— . (1990). Oral examinations: an historical note. *Language Testing* 7 (2) 158-173. From http://ltj.sagepub.com/cgi/content/abstract/7/2/158

——— . (1995a). *Measured words*. Oxford: Oxford University Press.

——— . (1995b). Prognostication and language aptitude testing – 1925–62. *Language Testing,* 12(3), 321–340. From http://ltj.sagepub.com/cgi/content/abstract/12/3/321

——— . (1997). The ethics of gatekeeping tests: What have we learned in a hundred years? *Language Testing,* 14(3)242–247. From http://ltj.sagepub.com/cgi/content/abstract/14/3/242

——— . (1999). Language Testing. In B. Spolsky (ed.), *Concise Encyclopedia Of Educational Linguistic,* (pp. 695- 703), Oxford, UK: Elsevier Science Ltd

——— . (2001). Cheating language tests can be dangerous. In C, Elder, et al (eds.). *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 212-221). Cambridge, UK: Cambridge University Press; UCLES

——— . (2001-2005). The state of the art in language assessment. *Russian Language Journal*, 55(180-182), 169-187. From http://russnet.org/previousIssues/RLJ%20 Volume%2055.pdf

——— . (2004*). Language Policy*. Cambridge: Cambridge University Press.

——— . (2008a). Introduction: Language testing at 25: Maturity and responsibility? *Language Testing*, *25*(3), 297–305. http://ltj.sagepub.com/cgi/content/refs/25/3/297

——— . (2008b). Language assessment in historical and future perspective. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education*: *Language testing and assessment* (2nd ed., vol. 7, pp. 445–454). New York: Springer Science.

Stansfield, C. W. (2008). Lecture: Where we have been and where we should go. *Language Testing,* 2008 25 (3) 311–326. From http://ltj.sagepub.com/cgi/content/refs/25/3/311

Strevens, P. (1977). *New Orientations in the Teaching of English*. Oxford: Oxford University Press

Swales , J. M. (1990). *Genre Analysis: English in Academic and Research Settings* . Cambridge : Cambridge University Press.

———. (2000). Languages for specific purposes. *Annual Review of Applied Linguistics* 20, 59–76. From http://dx.doi.org/10.1017/S026719050020

Tarone, E. (1981). Some Thoughts on the Notion of Communication Strategy. *TESOL Quarterly*, 15 (3), 285-295. From http://www.jstor.org/stable/3586754 .

———. (2001). Assessing language for Skills specific purposes: Describing and analyzing the 'behavior domain'. In C.A. Elder, et al (eds.). *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 53- 60). Cambridge, UK: Cambridge University Press; UCLES.

Tavakoli, H. (2012). *A Dictionary of Research Methodology and Statistics in Applied Linguistics*. Tehran, Iran: Rahnama Press.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21–36. Retrieved from http://dx.doi.org/10.1017/S0267190509090035

———. (2014). General language proficiency (GLP): Reflections on the “Issues Revisited” from the perspective of a UK examination board. *Language Assessment Quarterly,*11 (2) 136-151. From http://dx.doi.org/10.1080/15434303.2014.896366

Toulmin, S.E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Toulmin, S. E. (2001). *Return to reason*. Cambridge, MA: Harvard University Press.

———. (2003). *The uses of argument* (2nd ed.). Cambridge: Cambridge University Press.

———. (2006). Good reasoning on the Toulmin model. In D. Hitchcock & B. Verheij (Eds). *Arguing on the Toulmin model: New essays in argument analysis and evaluation,* (pp. 203-218). Dordrecht, the Netherlands: Springer

Toulmin, S., Rieke, R., & Janik, A. (1984). *An introduction to reasoning* (2nd ed). New York, NY: Macmillan Publishing Co., Inc.

Urbina, S. (2004). *Essentials of Psychological Testing.* Hoboken, N J: John Wiley & Sons, Inc.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? *System 28, 499–509.* From www.elsevier.com/locate/system

———. (2012). Washback. In G. Fulcher., & F. Davidson (Eds), *The Routledge Handbook of Language Testing* (pp.79-92). Abingdon, Oxon: Routledge.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing,* 10(3), 41-69. http://ltj.sagepub.com/cgi/content/abstract/10/1/41

Walters, F. S. (2012).Fairness. In G. Fulcher., & F. Davidson (Eds), *The Routledge Handbook of Language Testing* (pp. 455- 468). Abingdon, Oxon: Routledge.

Weigle , S.C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15 (2) 263–287. From http://ltj.sagepub.com/cgi/content/abstract/15/2/263

———. (1994). Effects of training on raters of ESL compositions. *Language Testing,* 11 (2) 197-223. Retrieved from http://ltj.sagepub.com/cgi/content/abstract/11/2/197

———. (2002). *Assessing writing*. Cambridge: Cambridge University Press

Weir, C.J. (1990). *Communicative Language Testing*. New York: Prentice Hall.

——— . (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, UK: Palgrave Macmillan.

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford: Oxford University Press.

———. (1979). *Explorations in Applied Linguistics*1. Oxford: Oxford University Press.

———. (1983*). Learning purpose and language use*. London: Oxford University Press

———. (1984). *Explorations in Applied Linguistics* 2. Oxford: Oxford University Press

———. (2001). Communicative language testing: The art of the possible. In C.A. Elder., et al (eds.). *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 12-21). Cambridge, UK: Cambridge University Press; UCLES

———. (2003). *Defining Issues in English Language Teaching*. Oxford: Oxford University Press.

——— . (2004). *Text, context, pretext Critical issues in discourse analysis*. Oxford: Blackwell Publishing Ltd

Williamson, M, D., Bejar,I.I., & Mislevy, R ,J. (2006a). Automated scoring of complex tasks in computer-based testing: An introduction. In M, D,Williamson, I, Bejar & R ,J Mislevy (Eds) . *Automated scoring of complex tasks in computer-based testing* (pp. 1-14). Mahwah, New Jersey/London:Lawrence Erlbaum Associates, Publishers.

——— . (Eds) (2006b). *Automated scoring of complex tasks in computer-based testing*. Mahwah, New Jersey/London:Lawrence Erlbaum Associates, Publishers.

Wilson, M,. & Sapsford, R. (2006). Asking Questions. In J Sapsford., & V, Jupp (Eds). *Data collection and analysis* (2nd ed, pp.93-123). London: SAGE Publications; The Open university.

Reference List in Arabic

الشروق اليومي.(2008). صناع أسئلة بكالوريا 2008 في "معسكر مغلق.
www.echoroukonline.com/ara/?news=3576.

ـــــــــــ . (2009) التعليمات الأخيرة الموجهة لمعدّي الأسئلة: هذه هي مسودة "صناع" البكالوريا في طريقة إعداد الأسئلة.http://www.echoroukonline.com/ara/?news=3576

الديوان الوطني للامتحانات والمسابقات. (2013). الوجه الآخر للبكالوريا : كيفية إعداد أسئلة الامتحانات..www.eddirasa.com

مديرية التربة لولاية الوادي. (2001) .تقويم نتائج امتحان شهادة البكالوريا دورة 2001. الوادي: مركز التوجيه المهني والمدرسي.

ـــــــــــ .(2002).تقويم نتائج امتحان شهادة البكالوريا دورة 2002. الوادي: مركز التوجيه المهني والمدرسي.

ـــــــــــ . (2003).تقويم نتائج امتحان شهادة البكالوريا دورة 2003. الوادي: مركز التوجيه المهني والمدرسي.

ـــــــــــ . (2004) .تقويم نتائج امتحان شهادة البكالوريا دورة 2004. الوادي: مركز التوجيه المهني والمدرسي.

ـــــــــــ . (2005). :تقويم نتائج امتحان شهادة البكالوريا دورة 2005. الوادي: مركز التوجيه المهني والمدرسي.

ـــــــــــ . (2006). تقويم نتائج امتحان شهادة البكالوريا دورة 2006. الوادي: مركز التوجيه المهني والمدرسي.

مديرية التعليم الثانوي العام. (1994).منشور رقم 94/32 الصادر في 94/11/16 يتعلق بتنقيح وإعادة كتابة البرامج التعليمية وطبعها. مجموعة المناشير الصادرة خلال الموسم الدراسي 1991-1992. الجزائر: المديرية الفرعية للوثائق.

وزارة التربية الوطنية . (1997). منشور رقم 95/78 الصادر في 95/10/23 يتعلق بالبرامج المعتمدة في شعب التعليم الثانوي التقني والتكنولوجي. مجموعة المناشير الصادرة خلال الموسم الدراسي 1994-1995. الجزائر: المديرية الفرعية للوثائق.

ـــــــــــ . (2000).دليل بناء اختبار مادة الإنجليزية في امتحان البكالوريا. الجزائر: الديوان الوطني للامتحانات والمسابقات.

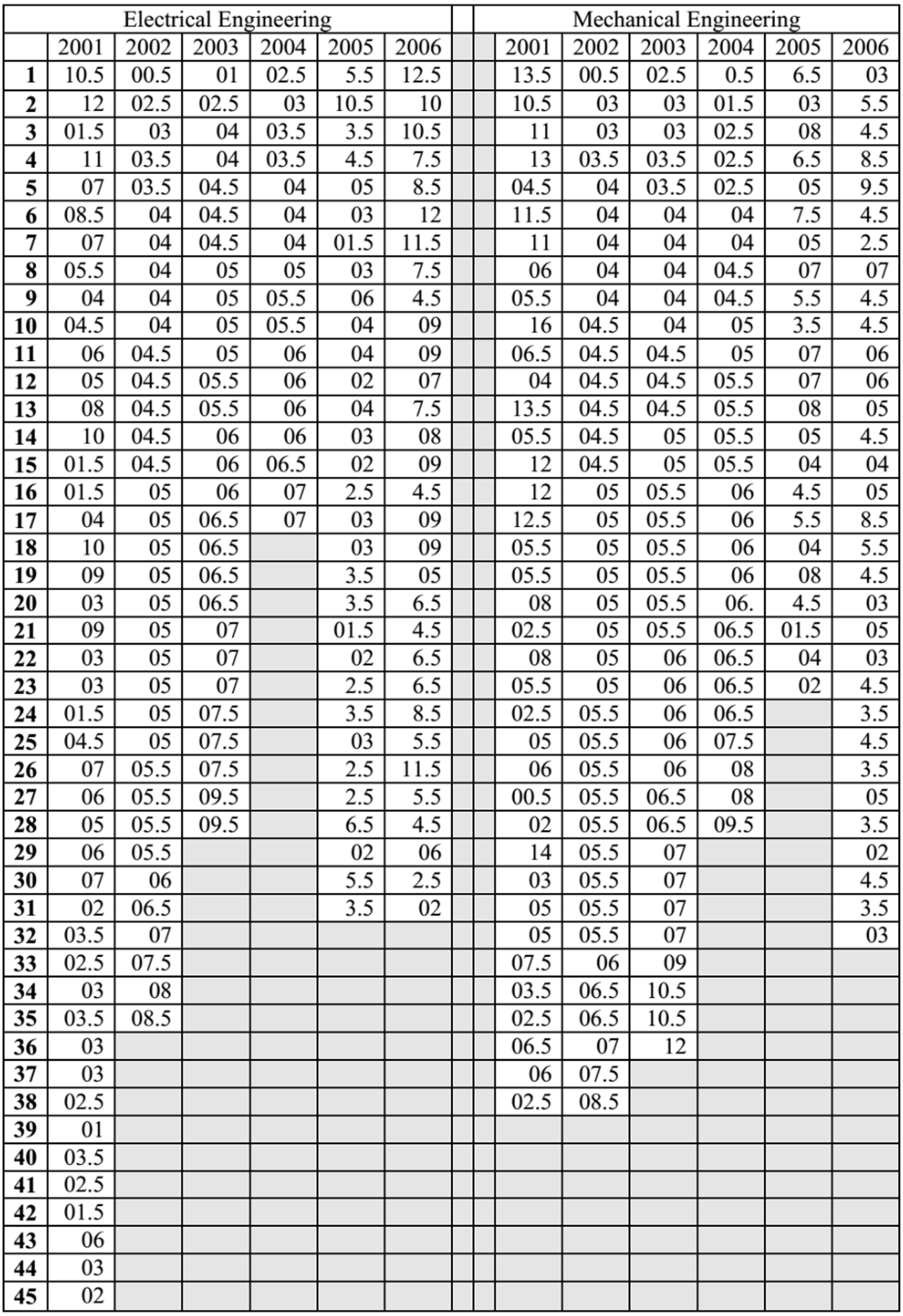
ــــــــــــــ . (2002) البرامج والمواقيت في التعليم الأسـاسي والثانوي: النشرة الرسمية للتربية الوطنية. الجزائر: المديرية الفرعية للتوثيق.

ــــــــــــــ . (2005). مـشروع إعادة تنظيـم التّعليـم والتّكويـن مـا بعد الإلزامي: 1. التّعليـم الثّـانوي العـام والتّكنـولـوجـي. الجـزائـر: مديرية التعليم الثانوي العـام والتّكنـولـوجـي.

Appendices

Appendix A: Technology Pupils BAC English Scores from 2001-2006

Table A 1: The Scores obtained by Technology streams in Guémar Technical School organized according the school results register



	Electrical Engineering							Mechanical Engineering					
	2001	2002	2003	2004	2005	2006		2001	2002	2003	2004	2005	2006
1	10.5	00.5	01	02.5	5.5	12.5		13.5	00.5	02.5	0.5	6.5	03
2	12	02.5	02.5	03	10.5	10		10.5	03	03	01.5	03	5.5
3	01.5	03	04	03.5	3.5	10.5		11	03	03	02.5	08	4.5
4	11	03.5	04	03.5	4.5	7.5		13	03.5	03.5	02.5	6.5	8.5
5	07	03.5	04.5	04	05	8.5		04.5	04	03.5	02.5	05	9.5
6	08.5	04	04.5	04	03	12		11.5	04	04	04	7.5	4.5
7	07	04	04.5	04	01.5	11.5		11	04	04	04	05	2.5
8	05.5	04	05	05	03	7.5		06	04	04	04.5	07	07
9	04	04	05	05.5	06	4.5		05.5	04	04	04.5	5.5	4.5
10	04.5	04	05	05.5	04	09		16	04.5	04	05	3.5	4.5
11	06	04.5	05	06	04	09		06.5	04.5	04.5	05	07	06
12	05	04.5	05.5	06	02	07		04	04.5	04.5	05.5	07	06
13	08	04.5	05.5	06	04	7.5		13.5	04.5	04.5	05.5	08	05
14	10	04.5	06	06	03	08		05.5	04.5	05	05.5	05	4.5
15	01.5	04.5	06	06.5	02	09		12	04.5	05	05.5	04	04
16	01.5	05	06	07	2.5	4.5		12	05	05.5	06	4.5	05
17	04	05	06.5	07	03	09		12.5	05	05.5	06	5.5	8.5
18	10	05	06.5		03	09		05.5	05	05.5	06	04	5.5
19	09	05	06.5		3.5	05		05.5	05	05.5	06	08	4.5
20	03	05	06.5		3.5	6.5		08	05	05.5	06.	4.5	03
21	09	05	07		01.5	4.5		02.5	05	05.5	06.5	01.5	05
22	03	05	07		02	6.5		08	05	06	06.5	04	03
23	03	05	07		2.5	6.5		05.5	05	06	06.5	02	4.5
24	01.5	05	07.5		3.5	8.5		02.5	05.5	06	06.5		3.5
25	04.5	05	07.5		03	5.5		05	05.5	06	07.5		4.5
26	07	05.5	07.5		2.5	11.5		06	05.5	06	08		3.5
27	06	05.5	09.5		2.5	5.5		00.5	05.5	06.5	08		05
28	05	05.5	09.5		6.5	4.5		02	05.5	06.5	09.5		3.5
29	06	05.5			02	06		14	05.5	07			02
30	07	06			5.5	2.5		03	05.5	07			4.5
31	02	06.5			3.5	02		05	05.5	07			3.5
32	03.5	07						05	05.5	07			03
33	02.5	07.5						07.5	06	09			
34	03	08						03.5	06.5	10.5			
35	03.5	08.5						02.5	06.5	10.5			
36	03							06.5	07	12			
37	03							06	07.5				
38	02.5							02.5	08.5				
39	01												
40	03.5												
41	02.5												
42	01.5												
43	06												
44	03												
45	02												

Table A 2: The ordered listing of the scores obtained by Technology streams in Guémar Technical School from 2001-2006

	Electrical Engineering							Mechanical Engineering					
	2001	2002	2003	2004	2005	2006		2001	2002	2003	2004	2005	2006
1	01	00.5	01	02.5	01.5	02		00.5	00.5	02.5	0.5	01.5	02
2	01.5	02.5	02.5	03	01.5	2.5		02	03	03	01.5	02	2.5
3	01.5	03	04	03.5	02	4.5		02.5	03	03	02.5	03	03
4	01.5	03.5	04	03.5	02	4.5		02.5	03.5	03.5	02.5	3.5	03
5	01.5	03.5	04.5	04	02	4.5		02.5	04	03.5	02.5	04	03
6	01.5	04	04.5	04	02	4.5		02.5	04	04	04	04	03
7	02	04	04.5	04	2.5	05		03	04	04	04	04	3.5
8	02	04	05	05	2.5	5.5		03.5	04	04	04.5	4.5	3.5
9	02.5	04	05	05.5	2.5	5.5		04	04	04	04.5	4.5	3.5
10	02.5	04	05	05.5	2.5	06		04.5	04.5	04	05	05	3.5
11	02.5	04.5	05	06	03	6.5		05	04.5	04.5	05	05	04
12	03	04.5	05.5	06	03	6.5		05	04.5	04.5	05.5	05	4.5
13	03	04.5	05.5	06	03	6.5		05	04.5	04.5	05.5	5.5	4.5
14	03	04.5	06	06	03	07		05.5	04.5	05	05.5	5.5	4.5
15	03	04.5	06	06.5	03	7.5		05.5	04.5	05	05.5	6.5	4.5
16	03	05	06	07	03	7.5		05.5	05	05.5	06	6.5	4.5
17	03	05	06.5	07	3.5	7.5		05.5	05	05.5	06	07	4.5
18	03	05	06.5		3.5	08		05.5	05	05.5	06	07	4.5
19	03.5	05	06.5		3.5	8.5		06	05	05.5	06	07	4.5
20	03.5	05	06.5		3.5	8.5		06	05	05.5	06.	7.5	4.5
21	03.5	05	07		3.5	09		06	05	05.5	06.5	08	05
22	04	05	07		04	09		06.5	05	06	06.5	08	05
23	04	05	07		04	09		06.5	05	06	06.5	08	05
24	04.5	05	07.5		04	09		07.5	05.5	06	06.5		05
25	04.5	05	07.5		4.5	09		08	05.5	06	07.5		5.5
26	05	05.5	07.5		05	10		08	05.5	06	08		5.5
27	05	05.5	09.5		5.5	10.5		10.5	05.5	06.5	08		06
28	05.5	05.5	09.5		5.5	11.5		11	05.5	06.5	09.5		06
29	06	05.5			06	11.5		11	05.5	07			07
30	06	06			6.5	12		11.5	05.5	07			8.5
31	06	06.5			10.5	12.5		12	05.5	07			8.5
32	06	07						12	05.5	07			9.5
33	07	07.5						12.5	06	09			
34	07	08						13	06.5	10.5			
35	07	08.5						13.5	06.5	10.5			
36	07							13.5	07	12			
37	08							14	07.5				
38	08.5							16	08.5				
39	09												
40	09												
41	10												
42	10												
43	10.5												
44	11												
45	12												

Table A 3: The Scores Obtained by Technology Pupils in Eloued in 2001 Sessions

School	Specialty	Number of Pupils	SCORES above Average	Percentage
Guémar	Mechanical Engineering	38	12	31.57%
	Electrical Engineering	45	05	11.11%
Eloued	Mechanical Engineering	38	09	23.68%
	Electrical Engineering	32	10	31.25%
Robbah	Mechanical Engineering	37	14	37.83%
	Electrical Engineering	28	11	39.28%
Debila	Mechanical Engineering	35	13	37.14
	Electrical Engineering	34	10	29.41
Lemghair	Mechanical Engineering	33	12	36.36%
	Electrical Engineering	40	15	37.5
Djemaa	Civil Engineering	35	18	51.41%

Table A 4: The Scores Obtained by Technology Pupils in Eloued in 2002Session

School	Specialty	Number of Pupils	SCORES above Average	Percentage
Guémar	Mechanical Engineering		00	00%
	Electrical Engineering		00	00%
Eloued	Mechanical Engineering		00	00%
	Electrical Engineering		00	00%
Robbah	Mechanical Engineering		00	00%
	Electrical Engineering		00	00%
Debila	Mechanical Engineering		00	00%
	Electrical Engineering		00	00%
Lemghair	Mechanical Engineering		00	00%
	Electrical Engineering		00	00%
Djemaa	Civil Engineering		00	00%

Table A 5: The Scores Obtained by Technology Pupils in Eloued in 2003 Session

School	Specialty	Number of Pupils	Scores above Average	Percentage
Guémar	Mechanical Engineering	28	00	00%
	Electrical Engineering	36	03	08.33%
Eloued	Mechanical Engineering	21	00	00%
	Electrical Engineering	24	00	00%
Robbah	Mechanical Engineering	39	00	00%
	Electrical Engineering	28	11	
Debila	Mechanical Engineering	22	00	00%
	Electrical Engineering	27	01	03.7%
Lemghair	Mechanical Engineering	29	00	00%
	Electrical Engineering	29	01	03.45%
Djemaa	Civil Engineering	33	04	12.12%

Table A6: The Scores Obtained by Technology Pupils in Eloued in 2004 Session

School	Specialty	Number of Pupils	Scores above Average	Percentage
Guémar	Mechanical Engineering	20	00	00 %
	Electrical Engineering	26	00	00 %
Eloued	Mechanical Engineering	16	00	00 %
	Electrical Engineering	21	00	00 %
Robbah	Mechanical Engineering	39	00	00 %
	Electrical Engineering	33	00	00 %
Debila	Mechanical Engineering	24	00	00 %
	Electrical Engineering	33	00	00 %
Lemghair	Mechanical Engineering	20	00	00 %
	Electrical Engineering	29	00	00 %
Djemaa	Civil Engineering	25	00	00 %

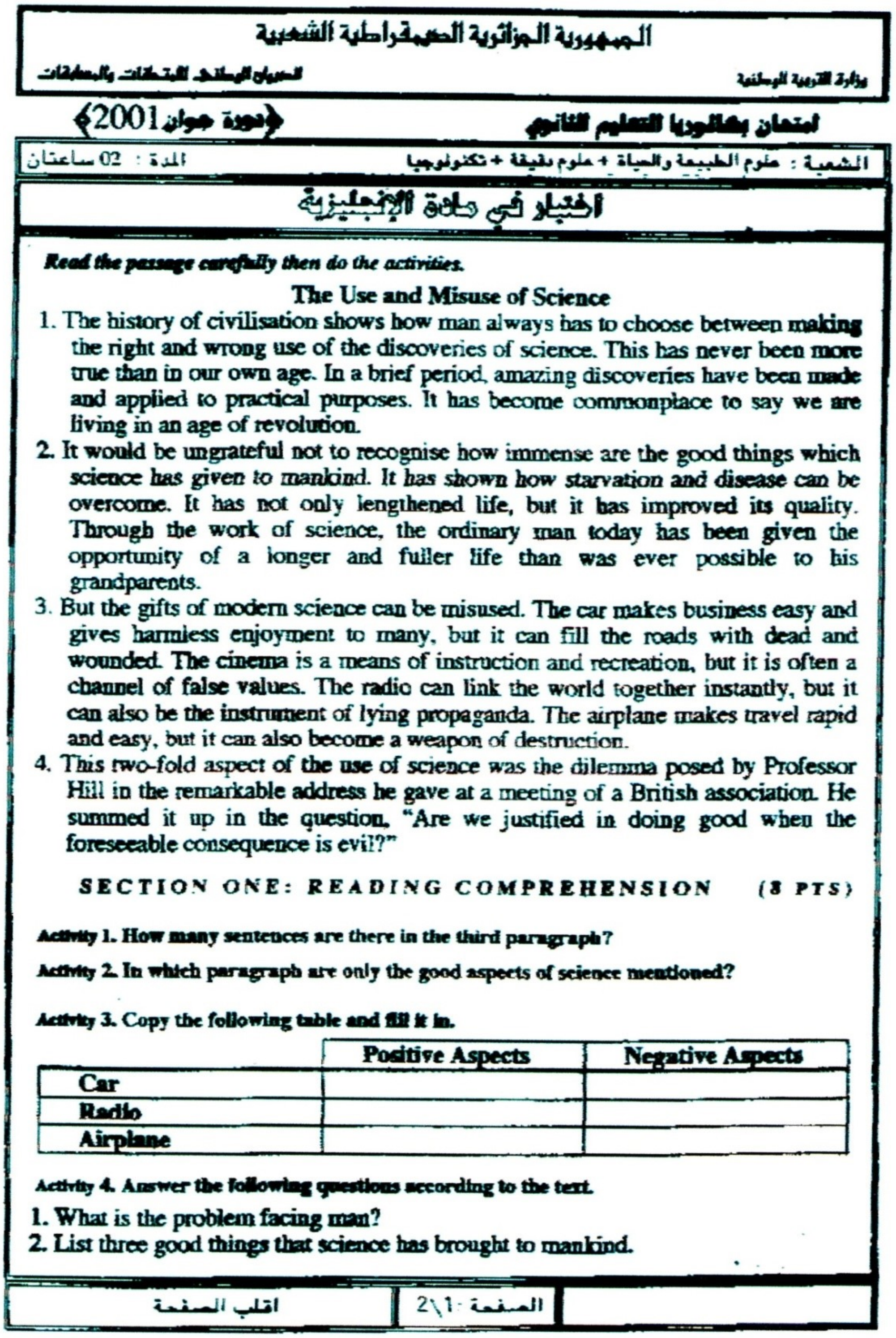
Table A7: The Scores Obtained by Technology Pupils in Eloued in 2005 Session

School	Specialty	Number of Pupils	Scores above Average	Percentage
Guémar	Mechanical Engineering			00 %
	Electrical Engineering			00 %
Eloued	Mechanical Engineering			00 %
	Electrical Engineering			00 %
Robbah	Mechanical Engineering			00 %
	Electrical Engineering			00 %
Debila	Mechanical Engineering			00 %
	Electrical Engineering			00 %
Lemghair	Mechanical Engineering			00 %
	Electrical Engineering			00 %
Djemaa	Civil Engineering			00 %

Table A 8: The Scores Obtained by Technology Pupils in Eloued in 2006 Session

School	Specialty	Number of Pupils	Scores above Average	Percentage
Guémar	Mechanical Engineering	31	00	00 %
	Electrical Engineering	32	06	18.75 %
Eloued	Mechanical Engineering	29	00	00 %
	Electrical Engineering	32	00	00 %
Robbah	Mechanical Engineering	28	03	10.71 %
	Electrical Engineering	29	02	06.89 %
Debila	Mechanical Engineering	33	00	00 %
	Electrical Engineering	32	00	00 %
Lemghair	Mechanical Engineering	25	02	08 %
	Electrical Engineering	27	03	11.11%
Djemaa	Civil Engineering	31	05	16.12 %

Appendix B: Technology Streams' BAC English tests from 2001 to 2006



الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التربية الوطنية

الديوان الوطني للامتحانات والمسابقات

امتحان بكالوريا التعليم الثانوي

﴿دورة جوان 2001﴾

الشعبة : علوم الطبيعة والحياة + علوم دقيقة + تكنولوجيا

المدة : 02 ساعتان

اختبار في مادة الإنجليزية

Read the passage carefully then do the activities.

The Use and Misuse of Science

1. The history of civilisation shows how man always has to choose between making the right and wrong use of the discoveries of science. This has never been more true than in our own age. In a brief period, amazing discoveries have been made and applied to practical purposes. It has become commonplace to say we are living in an age of revolution.
2. It would be ungrateful not to recognise how immense are the good things which science has given to mankind. It has shown how starvation and disease can be overcome. It has not only lengthened life, but it has improved its quality. Through the work of science, the ordinary man today has been given the opportunity of a longer and fuller life than was ever possible to his grandparents.
3. But the gifts of modern science can be misused. The car makes business easy and gives harmless enjoyment to many, but it can fill the roads with dead and wounded. The cinema is a means of instruction and recreation, but it is often a channel of false values. The radio can link the world together instantly, but it can also be the instrument of lying propaganda. The airplane makes travel rapid and easy, but it can also become a weapon of destruction.
4. This two-fold aspect of the use of science was the dilemma posed by Professor Hill in the remarkable address he gave at a meeting of a British association. He summed it up in the question, "Are we justified in doing good when the foreseeable consequence is evil?"

SECTION ONE: READING COMPREHENSION (8 PTS)

Activity 1. How many sentences are there in the third paragraph?

Activity 2. In which paragraph are only the good aspects of science mentioned?

Activity 3. Copy the following table and fill it in.

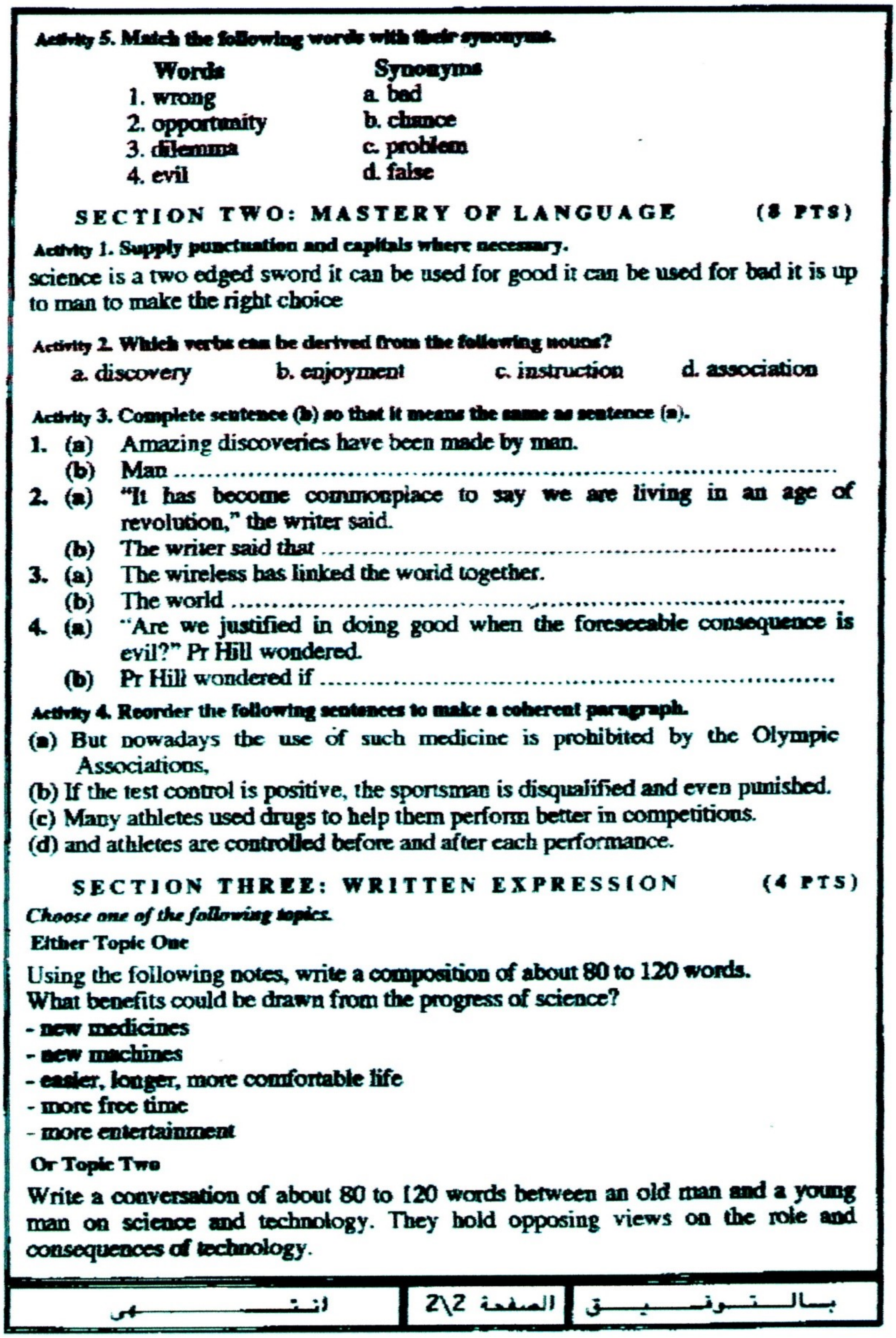
	Positive Aspects	Negative Aspects
Car		
Radio		
Airplane		

Activity 4. Answer the following questions according to the text.

1. What is the problem facing man?
2. List three good things that science has brought to mankind.

اقلب الصفحة

الصفحة : 2\1



Activity 5. Match the following words with their synonyms.

Words	Synonyms
1. wrong	a. bad
2. opportunity	b. chance
3. dilemma	c. problem
4. evil	d. false

SECTION TWO: MASTERY OF LANGUAGE (8 PTS)

Activity 1. Supply punctuation and capitals where necessary.

science is a two edged sword it can be used for good it can be used for bad it is up to man to make the right choice

Activity 2. Which verbs can be derived from the following nouns?

a. discovery b. enjoyment c. instruction d. association

Activity 3. Complete sentence (b) so that it means the same as sentence (a).

1. (a) Amazing discoveries have been made by man.
   (b) Man ....
2. (a) "It has become commonplace to say we are living in an age of revolution," the writer said.
   (b) The writer said that ....
3. (a) The wireless has linked the world together.
   (b) The world ....
4. (a) "Are we justified in doing good when the foreseeable consequence is evil?" Pr Hill wondered.
   (b) Pr Hill wondered if ....

Activity 4. Reorder the following sentences to make a coherent paragraph.

(a) But nowadays the use of such medicine is prohibited by the Olympic Associations,

(b) If the test control is positive, the sportsman is disqualified and even punished.

(c) Many athletes used drugs to help them perform better in competitions.

(d) and athletes are controlled before and after each performance.

SECTION THREE: WRITTEN EXPRESSION (4 PTS)

Choose one of the following topics.

Either Topic One

Using the following notes, write a composition of about 80 to 120 words.
What benefits could be drawn from the progress of science?

- new medicines
- new machines
- easier, longer, more comfortable life
- more free time
- more entertainment

Or Topic Two

Write a conversation of about 80 to 120 words between an old man and a young man on science and technology. They hold opposing views on the role and consequences of technology.

انتــــــــــــهى	الصفحة 2\2	بــالـــتـــوفـــيـــق

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التربية الوطنية

الديوان الوطني للامتحانات والمسابقات

امتحان بكالوريا التعليم الثانوي

دورة سبتمبر 2001

الشعبة : علوم الطبيعة والحياة + علوم دقيقة + تكنولوجيا

المدة : ساعتان

اختبار في مادة الإنجليزية

Read the passage carefully then do the activities.

1. Scientists know there are two basic approaches to prolonging life. One is the elimination of diseases such as cancer, heart and brain attacks that affect older people. The other is the slowing down of the process of growing old, the delaying of the deterioration of the body.
2. Scientists believe that they will soon develop the knowledge and ability to delay the ageing process by ten to fifteen years. The result will be that more people will live longer. Scientists believe that with the right diet, exercise, medical advice and mental attitude, many people can live to be 100 years old.
3. Gerontologists, people who specialise in the study of the process of growing old, are investigating why the body cells die. They are studying the activity of cells and the effect of diet on ageing. If their studies are successful, the result should help to improve the quality of life of the next generation.
4. What will some of the effects of longer life be? For one thing by adding extra more healthful years to a person's life, youth and middle age will be prolonged. A person's productivity and efficiency will be increased.
5. On the other hand, the longer lives would bring a major problem, that of money. Pensions would have to last longer, which means that governments would have to provide enough money to meet the increased cost of pensions. Otherwise, it would be tragic if man were to live longer but not have any financial security.
6. Today, gerontologists think that by the next decade, the results of their research will be apparent and that there will be a significant increase in the number of longer lives among the general population.

SECTION ONE: READING COMPREHENSION (8 PTS)

Activity **1. Are there any interrogative sentences in the passage? If so, how many?**

Activity **2. On your answer sheet, copy the title which is the most appropriate.**

a) Financial Security for Old People
b) Prolonging Life
c) Causes of Early Death

Activity **3. Answer the following questions according to the text.**

1. According to scientists, what should people do to live longer?
2. What does a gerontologist do?
3. What impact will living longer have on governments?
4. When will the results of the gerontologists' research be apparent?

اقلب الصفحة

الصفحة : 1\2

Activity 4. Match the following words from the text with their synonyms.

Words	Synonyms
1. improve	a) effects
2. provide	b) methods
3. approaches	c) make better
4. impacts	d) supply

SECTION TWO: MASTERY OF LANGUAGE (8 PTS)

Activity 1. Give the plural form of the following words.

a) life b) process c) body d) youth

Activity 2. Put the verbs in brackets in the correct form.

1. I (already, read) that book.
2. He (arrive) soon.

Activity 3. Complete sentence (b) so that it means the same as sentence (a).

1. (a) The writer said, "If their studies are successful, the results should help to improve the quality of life of the next generation."
   (b) The writer said that ....
2. (a) Scientists are studying the activity of cells.
   (b) The activity of cells ....

Activity 4. Reorder the following sentences to make a coherent paragraph.

(a) The man with the new heart lived for only eighteen days.
(b) He took a healthy heart from the body of a girl
(c) and put it into a man's body.
(d) In 1967, Dr Christian Barnard transplanted a heart for the first time.

Activity 5. Reorder these words to make a meaningful sentence.

environment	street	the	by	influenced	is	children	of	behaviour	the

Activity 6. Supply punctuation and capitals where necessary.

among the many effects of longer life expectation is the scarcity of food supply in certain regions of the world rapid development in agriculture is therefore necessary to cover a higher demand for food

SECTION THREE: WRITTEN EXPRESSION (4 PTS)

Choose one of the following topics.

Either Topic One

Using the following notes, write a composition of about 80 to 120 words.
What should people do to live longer?

- follow the instructions of their doctors
- go on a strict diet
- practise sport
- spend plenty of time outdoors
- avoid all excesses
- refrain from smoking

Or Topic Two

Write a composition of about 80 to 120 words on the following topic.
What kind of sport do you prefer? State your reasons.

بالتوفيق	الصفحة :2\2	انتهى

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التربية الوطنية　　　　الديوان الوطني للامتحانات والمسابقات

امتحان بكالوريا التعليم الثانوي　　﴿ دورة جوان 2002 ﴾

الشـعـب : علوم الطبيعة والحياة + علوم دقيقة + تكنولوجيا　　　　المـدة : ساعتان

اختبار في مادة الأنجليزية (لغة أجنبية ثانية)

Read the passage carefully then do the activities.

It is easy to think of the world's oceans as indestructible, bodies so deep and wide they can absorb anything. And enormous they are - 300 million cubic miles of water spread over 70 percent of the earth's surface. The only trouble is that we have managed to clog all the seas of the world with something like 20 billion tons of rubbish, including everything from soda cans to radioactive waste and exotic chemicals to heavy metals. And now, perhaps the oceans are finally telling us that enough is enough, and that those waters have suffered the worst effects of pollution.

At bottom, the problem is one of overpopulation in coastal areas and inadequate waste management. In the world-wide web of pollution, almost no one is blameless.

The irony is that the technology and expertise already exist to alleviate some of the worst effects. For instance, there are treatment plants that can take the heavily contaminated water and make it drinkable. Such facilities are terribly expensive, but it may eventually become clear that the costs of not investing in them are even higher.

Section One: Reading Comprehension **(8 pts)**

1. *How many paragraphs are there in the above passage?*

2. *Choose the general idea of the text.*
   a) Pollution of the environment.
   b) The world's polluted oceans.
   c) Measures taken to fight water pollution.

3. *Are these statements True, False or Not Mentioned ?*
   a) Oceans tell people to stop throwing rubbish.
   b) Demographic explosion is a cause of water pollution.
   c) Polluted waters cannot be treated.
   d) Coastal areas play the most important role in the chain of life.

4. *Answer the following questions according to the text.*
   a) What makes people think that oceans can absorb anything?
   b) What can be done to alleviate some of the effects of pollution?

5. *Match words and their definitions.*

WORDS	DEFINITIONS
a. to clog	1. to make less severe
b. contaminated	2. to fill , to block
c. to alleviate	3. not pure

اقلب الصفحة

Section Two: Mastery of Language **(8 pts)**

1. *Add three more words to the list.*

environment	pollution			

2. *Supply punctuation and capitalisation.*
   whales are sea-living mammals they breathe air but cannot survive on land
3. *Reorder the words to make a coherent sentence.*
   produce / radioactive / of / remain / all / nuclear / wastes / which / stations / for / years / thousands / power / dangerous
4. *Complete the following chart as shown in the example.*

Verb	Adjective	Noun
to think	*thoughtful*	*a thought*
to exist		
	blameless	
		pollution

5. *Classify these words according to the pronunciation of their final 's'.*
   wastes - bodies - chemicals - sons – facilities – thinks

/ s /	/ z /

6. *Rewrite sentence (b) so that it means the same as sentence (a).*
   a1. Polluted water can be treated.
   b1. We ....
   a2. "How many casualties were recorded during the Chernobyl accident? " he asked.
   b2. He asked ....
   a3. Radioactive waste and chemicals are spoiling our environment.
   b3. Our environment ....

Section Three: Written Expression **(4 pts)**

Choose ONE of the following topics.

Either topic one.

Write a composition of 80 – 120 words on the following topic.

According to you what are the measures that should be taken to protect our environment?

Or topic two.

This is a conversation between a journalist and a whale hunter. Complete what the journalist says.

Hunter: Can I help you?
Journalist: ...
Hunter: Of course. I know them well. There are two main groups of whales: toothed like the dolphin and toothless like the blue whales.
Journalist: ...
Hunter: Well! For their oil, their meat and a curious product called 'ambergris'.
Journalist: ...
Hunter: A substance produced by the whale, and it is used in the production of perfumes.
Journalist: ...
Hunter: I know we are destroying the whale stocks... But what can we do instead?
Journalist: ...

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التربية الوطنية　　　　الديوان الوطني للامتحانات والمسابقات

امتحان بكالوريا التعليم الثانوي　　(دورة جوان 2003)

الشعب : علوم الطبيعة والحياة + علوم دقيقة + تكنولوجيا　　المدة : ساعتان

اختبار في مادة الأنجليزية (لغة أجنبية ثانية)

Read the passage carefully then do the activities.

Research has shown that the physically fit person is able to withstand fatigue for longer periods than the unfit person; that the physically fit person is better equipped to tolerate physical stress; that the physically fit person has a stronger and more efficient heart; and that there is a relationship between good mental alertness, absence of nervous tension and physical fitness.

One way of being fit is through weight control. The major purpose of weight control is to reduce the amount of fat and to increase the amount of muscle. It is in reality a programme of fat control rather than weight control. This control can be exerted only by coupling a sensible dietary programme with a regular balanced programme of exercise.

When we eat, the food is used, stored or discarded. The body stores fuel or calories as fat. The more fuel we consume, and the less of it we use, then the more of it is stored in the body in the form of fat. The human body is not like the petrol tank of a car that will overflow when it is full. Our bodies accept all the calories that we put into them, and store those that we do not use.

When you exercise, you burn calories. As muscle is slightly heavier than fat, you may very well notice an increase in your weight rather than a reduction. However, it must be stressed that this muscle weight is useful weight and will improve the way you look and feel.

Research has shown clearly that the most effective way of taking off weight and keeping it off is through a programme which combines diet and exercise.

Section One: Reading Comprehension **(8 pts)**

1. *Are there any negative sentences in the third paragraph? If so, how many?*
2. *Are the following sentences true or false?*
   a) As compared to the physically unfit person, the fit person has a stronger and healthier life.
   b) A dietary programme is necessary for fat control.
   c) The human body rejects some calories.
   d) According to research, practising sport and special diet are very effective ways of taking off weight.
3. *Here are the answers to some questions about the text. Ask the questions.*
   a) The food is used, stored or discarded.
   b) Fuel or calories as fat.
   c) When you exercise.
4. *Find in the text words or phrases opposite in meaning to the following.*

a) weaker (§ 1)　　b) reject (§ 3)　　c) useless (§ 4)

Section Two: Mastery of Language **(8 pts)**

1. *Supply capitals and punctuation.*

the next olympic games will be held in athens athletes from different parts of the world will take part in the event the algerian athletes will certainly represent their country in an honourable way

2. *Divide the following words into roots and affixes.*
unfit - reality – ineffective

Prefix	Root	Suffix

3. *Complete the following chart as shown in the example.*

Verb	Noun	Adjective
produce	*product*	*productive*
	thought	
		known / knowledgeable
endanger		

4. *Complete sentence (b) so that it means the same as sentence (a).*

a1. "The muscle weight will improve the way we look", the writer says.
b1. The writer says that ....
a2. Solar energy is changed into chemical energy by plant cells.
b2. Plant cells ....
a3. The candidates had revised English before they slept last night.
b3. After ....

5. *Reorder these sentences to make a coherent paragraph. One irrelevant sentence must be left out.*

1. you will gain an extra pound.
2. and use only 2 600 of them in your activity,
3. When you accumulate about 4 000 of these calories,
4. you will lose 400 calories.
5. the remaining 400 calories will be stored in the body.
6. If you eat food that has a value of 3 000 calories

6. *Classify the following words according to the pronunciation of their final 'ed'.*

equipped - used - discarded - stored - accepted - reduced

/ t /	/ d /	/ id /

Section Three: Written Expression **(4 pts)**

Choose ONE of the following topics.

Either topic one:

Using the following notes, write a short paragraph of about 80 – 120 words on the following topic.
Activity and diet play a beneficial role in man's health.

- control weight	-decrease stress and anguish	- reduce heart problems
- activate the respiratory system	- make life more enjoyable	- feel and look well

Or topic two:

Write a composition of about 80 – 120 words on the following topic.
Do you like to practise sport? Give your reasons.

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التربية الوطنية — الديوان الوطني للامتحانات والمسابقات

امتحان بكالوريا التعليم الثانوي ﴿ دورة جوان 2004 ﴾

الشعبة : علوم الطبيعة والحياة + علوم دقيقة + تكنولوجيا — المدة : ساعتان

اختبار في مادة الإنجليزية ــ لغة أجنبية ثانية ــ

SECTION ONE: READING COMPREHENSION (8 points)

Read the text carefully then do the activities.

Soccer is probably the most popular sport in the world. Two teams of 11 players attempt to guide an inflated ball into goal cages opposite ends of a playing field. Soccer is unique because of **its** restriction on the use of the hands; only the goal keeper may handle the ball, and then only within a limited area.

The continuous action and fast pace of soccer have made it a major spectator sport throughout the world, and for **the same reasons** it has attracted millions of players. Since the late 1960s and early 1970s its growth in the United States, especially on the amateur level, has been substantial. The name of the game presents some confusion. In countries other than the United States soccer is called football. The word 'soccer' is short for 'association' football.

It is hard to believe that a game as fast and exciting as soccer had its origin in a religious ceremony several thousand years ago in Egypt. After putting an armor, two teams fought with sticks over a round stone. The custom of teams competing for control of a round object, or ball, spread across North Africa, the Arabic countries and Persia.

The international governing body of soccer is the Fédération Internationale de Football Association (FIFA), with headquarters in Zurich, Switzerland. Every 4 years national teams – made up of the top players from each country (**who** may play professionally for teams in other countries) - compete for the World Cup, soccer's most coveted prize. It is the most popular athletic event, possibly with the exception of the Summer Olympics. The 2002 World Cup Finals were hosted by two Asian countries: South Korea and Japan.

1. Are there any passive sentences in the above passage? If so, how many?

2. Are the following statements true, false or not mentioned?

a) Soccer is the most popular sport in America.
b) Football isn't played in the USA.
c) Soccer could be found in North Africa, long time ago.
d) Millions of viewers watched the last world cup finals.

3.On your answer sheet, copy the title which you think is the most appropriate.

a) The Last World Cup Finals
b) Football and Soccer
c) The History of Soccer

4.What or who do the underlined words or phrases refer to in the text?

a)… its restrictions ….(§1)
b)… the same reasons … (§2)
c)… who may play … (§4)

5. Find in the text words , phrases or expressions closest in meaning to:

a) try (§1) b) all over (§2) c) award (§4)

6. Find in the text words, phrases or expressions opposite in meaning to:

a) least (§1) b) minor (§2) c) slow (§3)

اقلب الصفحة — الصفحة 1 / 2

SECTION TWO: MASTERY OF LANGUAGE (8 points)

1. Supply punctuation, capitals and apostrophes where necessary.

mother the little boy said I want to see the game it s all right you may go she answered

2. On your answer sheet, copy the odd one out.

a) tennis volleyball handball basketball

b) passed watched succeed earned·

3. Divide the following words into their roots and affixes.

a) shortening b) athletic c) international

4. Complete sentence (b) so that it means the same as sentence (a).

1 (a) Eleven players guide an inflated ball into goal cages.
1 (b) An inflated …………………………………………

2 (a) The 2002 World Cup Finals had been hosted by Korea and Japan.
2 (b) Korea and Japan ………………………………………………

3 (a) He didn't watch the basketball match. He didn't study.
3 (b) He neither ……………………………………………

5. Fill in the gaps so that the passage makes sense.

One of the top women athletes Algeria ……1…….. ever known is Hassiba Boulmerka. This athlete has ……2…….. part in different running competitions all over the world. Thanks to her fitness ……3……. determination, she has won medals and become a star long distance ………4……… . She was the youngest world champion ever in the 1500-metre competition.

6. Classify the following words according to the number of their syllables.

a) drowned b) spectator c) game d) against

SECTION THREE: WRITTEN EXPRESSION (4 points)

Choose one of the following topics.

1. Either topic one: Complete the following dialogue.

A: ……………………………………
B: Unfortunately I did. It was a pity.
A: ……………………………………
B: Don't blame the referee. Our team didn't play well.
A: …………………………………………
B: Yes, you're right, we must support them.
A: …………………………………………
B: …………………………………………

2. Or topic two: Write a composition of about 80 to 120 words on the following topic.

Team sports contribute more to international understanding than individual sports. Do you agree? Give examples to justify your point of view.

بالتوفيق	الصفحة 2 / 2	انتهى

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التربية الوطنية — الديوان الوطني للامتحانات والمسابقات

امتحان بكالوريا التعليم الثانوي (دورة جوان 2005)

الشعبة : علوم الطبيعة والحياة + علوم دقيقة + تكنولوجيا. — المدة : ساعتان

اختبار في مادة الإنجليزية لغة أجنبية ثانية

SECTION ONE: READING COMPREHENSION (8 POINTS)

Read the passage carefully then do the activities.

Not all Americans are free to celebrate holidays at all times. Whether they must work or not depends upon the importance of the holiday, the demands of seasonal work, holidays agreed to in union contracts and other factors. Many newspaper reporters, radio broadcasters, hospital workers, police, fire fighters and workers who provide other essential services must work on holidays. All working Americans, however, do get vacation time. Most take their vacations during the summer months as is common in other nations. The amount of vacation time varies greatly, but most people get one or two weeks a year, after working for the same company for a year or more. More vacation is given after longer periods of work.

This brief description of holidays shows that for some of these special times, the customs of all or most Americans are very much the same. For others, however, the customs can vary greatly. Those who feel strongly about the labour unions, for example, see Labour Day as a day on which to demonstrate labour solidarity in a public way. For others, Labour Day means a day off to go for a ride in a car, to go for a final summer swim or to hold a family get-together.

Activity 1. Choose a title to the text.

a) American Celebrations
b) Holidays in America
c) American Workers

Activity 2. Answer these questions according to the text.

a) American holidays depend on several factors. Mention two of them.
b) *Do all Americans behave in the same way during* Labour Day? Justify your answer.

Activity 3. In which paragraph is it mentioned that in some kinds of jobs the Americans must work during holidays?

Activity 4. Here are the answers to questions about the text. Write the questions.

a) During the summer months. (§1)
b) More vacation time. (§1)

Activity 5. Find in the text words that are closest in meaning to:

a) differs (§1) b) usual (§1) c) short (§2) d) traditions (§2)

Section Two: Mastery of Language (8 points)

Activity 1. Supply punctuation and capitals where necessary.

the american student spends six hours a day five days a week 180 days a year in school children in the united states start pre-school at the age of four

Activity 2. Which adjectives can be derived from the following nouns?

a) life b) length c) child

Activity 3. Every sentence contains one mistake only. Write the sentence without the mistake.

a) Money bring money.

b) Nowadays. people do not like read.

Activity 4. Combine the following pairs using the connector provided.

a) Reading the newspaper / the telephone (ring) (while)

b) You get to London / (start) speaking English (as soon as)

Activity 5. Read the passage and delete the unnecessary words.

In during their free time. the students spend much time watching TV. Students they also listen to music.

Activity 6. Classify the following words according to the pronunciation of 'ed'.

worked depended reported showed

/d/	/t/	/id/

Section Three: Written Expression (4 points)

Choose one of the following topics.

Either Topic One

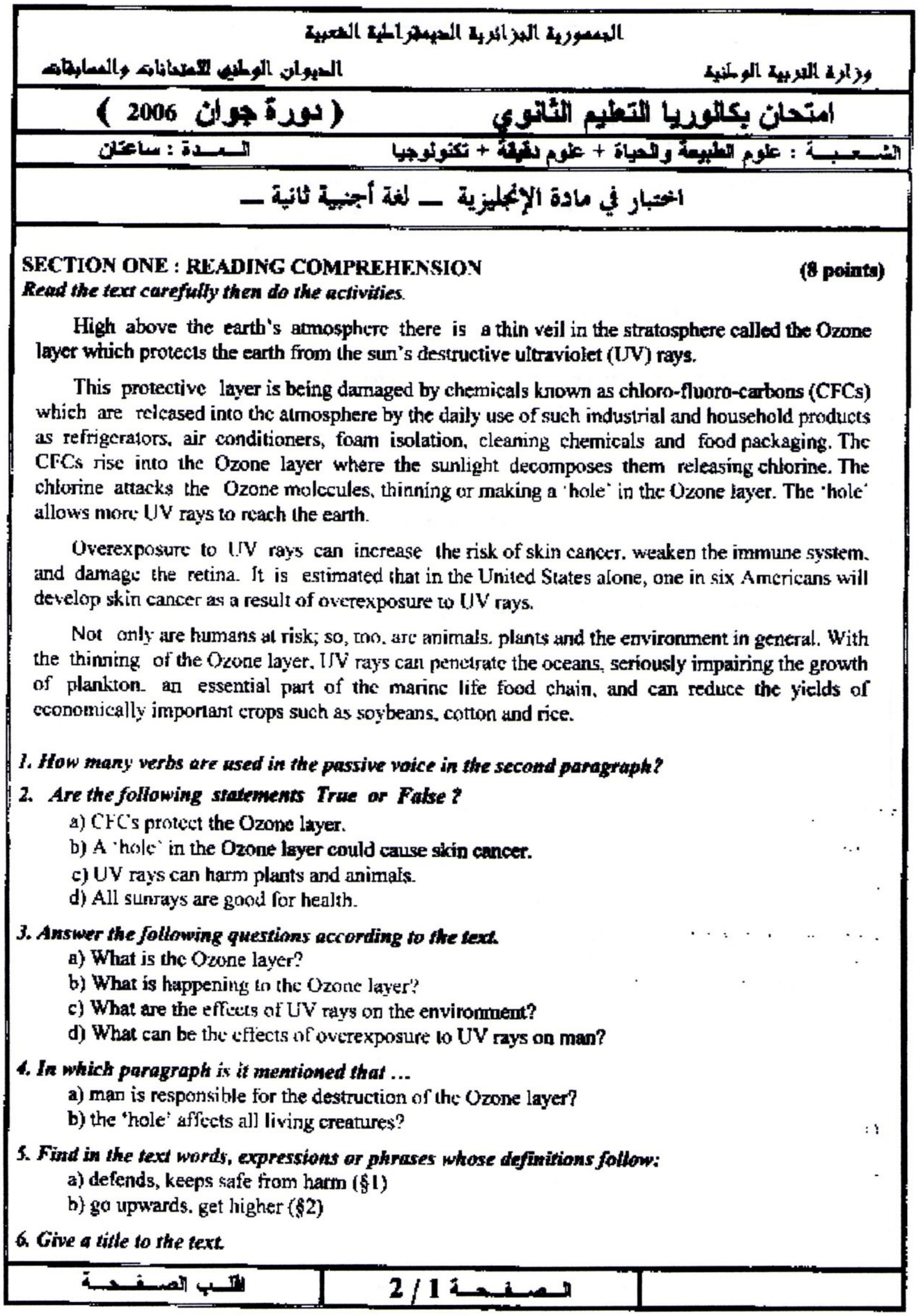
Using the following notes, write a composition of about 80 to 120 words.

- Algerian holidays : mostly in summer
- A few Algerians work - only necessary services
- They spend holidays : seaside – weddings – with family
- Holidays still expensive (hotels food)
- A lot stay at home.

Or Topic Two

Write a composition of about 80 to 120 words on the following topic.

Do you prefer to spend your summer holidays in your country or abroad? Give your reasons.



الجمهورية الجزائرية الديمقراطية الشعبية

الديوان الوطني للامتحانات والمسابقات وزارة التربية الوطنية

امتحان بكالوريا التعليم الثانوي (دورة جوان 2006)

الشـــعـبـــة : علوم الطبيعة والحياة + علوم دقيقة + تكنولوجيا الـمـــدة : ساعتان

اختبار في مادة الإنجليزية ـــ لغة أجنبية ثانية ـــ

SECTION ONE : READING COMPREHENSION (8 points)

Read the text carefully then do the activities.

High above the earth's atmosphere there is a thin veil in the stratosphere called the Ozone layer which protects the earth from the sun's destructive ultraviolet (UV) rays.

This protective layer is being damaged by chemicals known as chloro-fluoro-carbons (CFCs) which are released into the atmosphere by the daily use of such industrial and household products as refrigerators, air conditioners, foam isolation, cleaning chemicals and food packaging. The CFCs rise into the Ozone layer where the sunlight decomposes them releasing chlorine. The chlorine attacks the Ozone molecules, thinning or making a 'hole' in the Ozone layer. The 'hole' allows more UV rays to reach the earth.

Overexposure to UV rays can increase the risk of skin cancer, weaken the immune system, and damage the retina. It is estimated that in the United States alone, one in six Americans will develop skin cancer as a result of overexposure to UV rays.

Not only are humans at risk; so, too, are animals, plants and the environment in general. With the thinning of the Ozone layer, UV rays can penetrate the oceans, seriously impairing the growth of plankton, an essential part of the marine life food chain, and can reduce the yields of economically important crops such as soybeans, cotton and rice.

1. How many verbs are used in the passive voice in the second paragraph?

2. Are the following statements True or False ?

a) CFC's protect the Ozone layer.
b) A 'hole' in the Ozone layer could cause skin cancer.
c) UV rays can harm plants and animals.
d) All sunrays are good for health.

3. Answer the following questions according to the text.

a) What is the Ozone layer?
b) What is happening to the Ozone layer?
c) What are the effects of UV rays on the environment?
d) What can be the effects of overexposure to UV rays on man?

4. In which paragraph is it mentioned that ...

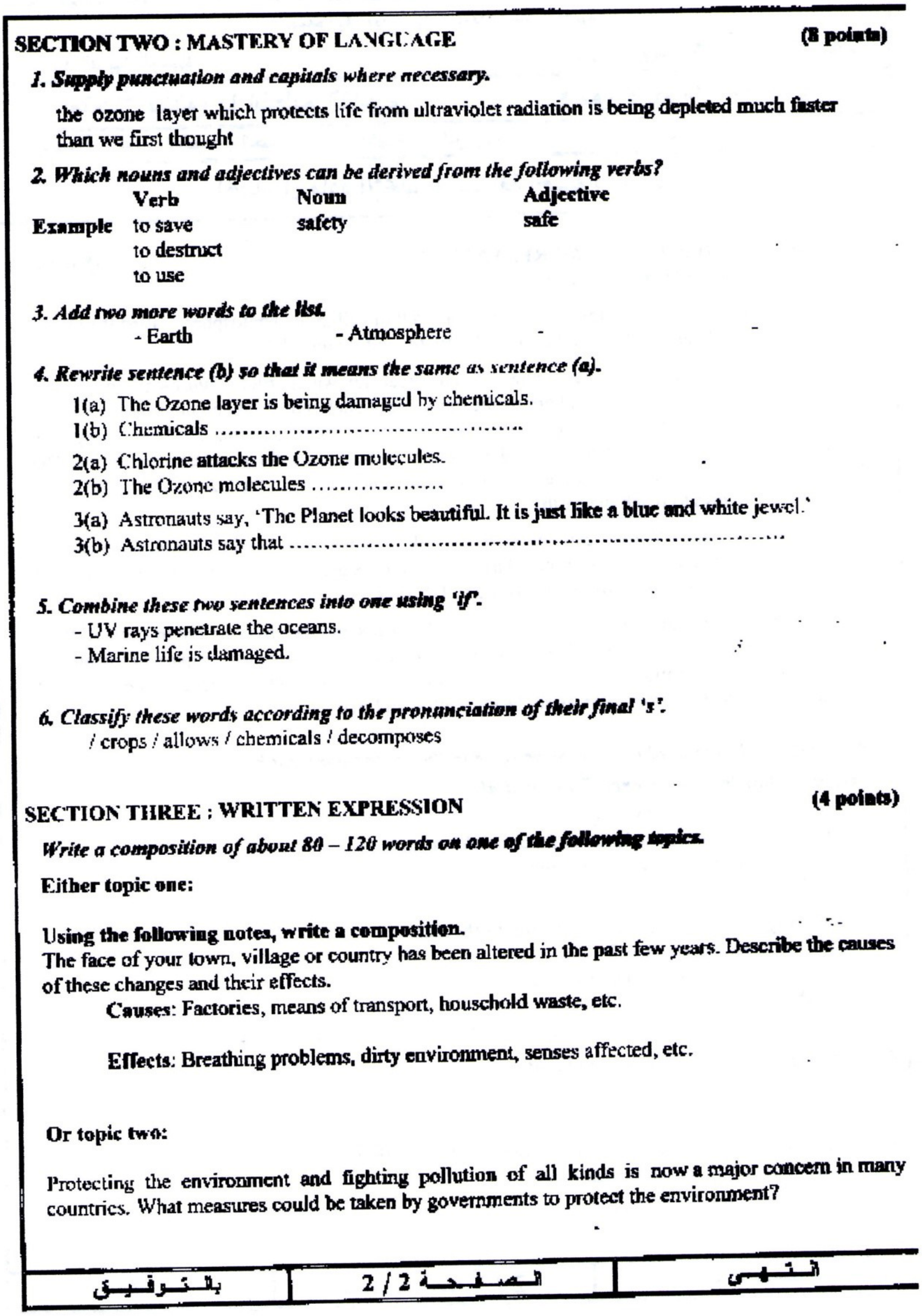
a) man is responsible for the destruction of the Ozone layer?
b) the 'hole' affects all living creatures?

5. Find in the text words, expressions or phrases whose definitions follow:

a) defends, keeps safe from harm (§1)
b) go upwards, get higher (§2)

6. Give a title to the text.

اقلب الصفحة الصفحة 2 / 1



SECTION TWO : MASTERY OF LANGUAGE **(8 points)**

1. Supply punctuation and capitals where necessary.

the ozone layer which protects life from ultraviolet radiation is being depleted much faster than we first thought

2. Which nouns and adjectives can be derived from the following verbs?

	Verb	Noun	Adjective
Example	to save	safety	safe
	to destruct		
	to use		

3. Add two more words to the list.

- Earth - Atmosphere - -

4. Rewrite sentence (b) so that it means the same as sentence (a).

1(a) The Ozone layer is being damaged by chemicals.
1(b) Chemicals ……………………………………

2(a) Chlorine attacks the Ozone molecules.
2(b) The Ozone molecules ………………

3(a) Astronauts say, 'The Planet looks beautiful. It is just like a blue and white jewel.'
3(b) Astronauts say that ……………………………………………………………

5. Combine these two sentences into one using 'if'.

- UV rays penetrate the oceans.
- Marine life is damaged.

6. Classify these words according to the pronunciation of their final 's'.

/ crops / allows / chemicals / decomposes

SECTION THREE : WRITTEN EXPRESSION **(4 points)**

Write a composition of about 80 – 120 words on one of the following topics.

Either topic one:

Using the following notes, write a composition.
The face of your town, village or country has been altered in the past few years. Describe the causes of these changes and their effects.

Causes: Factories, means of transport, houschold waste, etc.

Effects: Breathing problems, dirty environment, senses affected, etc.

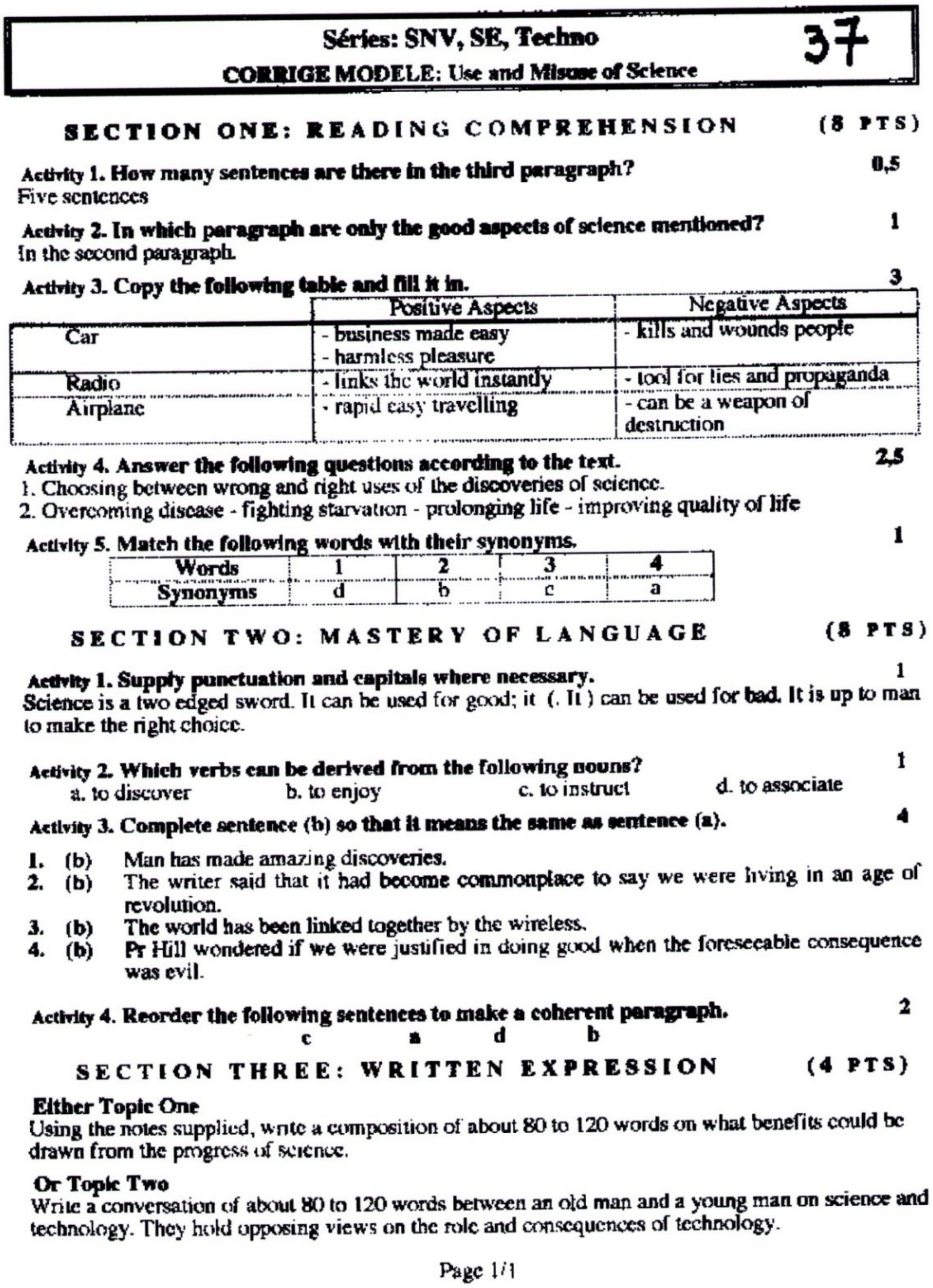
Or topic two:

Protecting the environment and fighting pollution of all kinds is now a major concern in many countries. What measures could be taken by governments to protect the environment?

بالتوفيق	الصفحة 2 / 2	انتهى

Appendix C: Model Corrections of Technology Streams' BAC English Tests

- June 2001



Séries: SNV, SE, Techno

37

CORRIGE MODELE: Use and Misuse of Science

SECTION ONE: READING COMPREHENSION (8 PTS)

Activity 1. How many sentences are there in the third paragraph? 0,5
Five sentences

Activity 2. In which paragraph are only the good aspects of science mentioned? 1
In the second paragraph.

Activity 3. Copy the following table and fill it in. 3

	Positive Aspects	Negative Aspects
Car	- business made easy - harmless pleasure	- kills and wounds people
Radio	- links the world instantly	- tool for lies and propaganda
Airplane	- rapid easy travelling	- can be a weapon of destruction

Activity 4. Answer the following questions according to the text. 2,5
1. Choosing between wrong and right uses of the discoveries of science.
2. Overcoming disease - fighting starvation - prolonging life - improving quality of life

Activity 5. Match the following words with their synonyms. 1

Words	1	2	3	4
Synonyms	d	b	c	a

SECTION TWO: MASTERY OF LANGUAGE (8 PTS)

Activity 1. Supply punctuation and capitals where necessary. 1
Science is a two edged sword. It can be used for good; it (. It) can be used for bad. It is up to man to make the right choice.

Activity 2. Which verbs can be derived from the following nouns? 1
a. to discover b. to enjoy c. to instruct d. to associate

Activity 3. Complete sentence (b) so that it means the same as sentence (a). 4

1. (b) Man has made amazing discoveries.
2. (b) The writer said that it had become commonplace to say we were living in an age of revolution.
3. (b) The world has been linked together by the wireless.
4. (b) Pr Hill wondered if we were justified in doing good when the foreseeable consequence was evil.

Activity 4. Reorder the following sentences to make a coherent paragraph. 2
c a d b

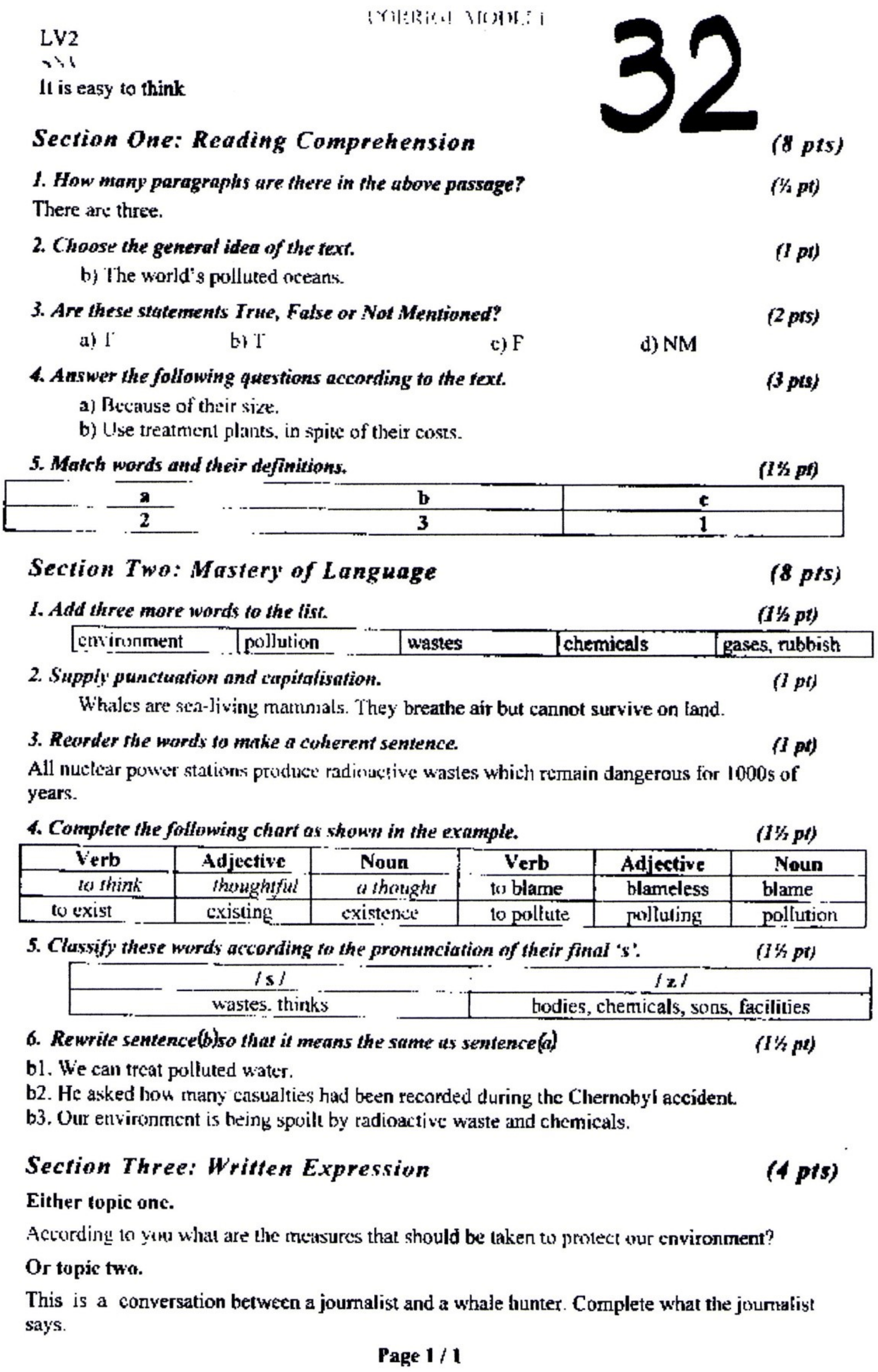
SECTION THREE: WRITTEN EXPRESSION (4 PTS)

Either Topic One
Using the notes supplied, write a composition of about 80 to 120 words on what benefits could be drawn from the progress of science.

Or Topic Two
Write a conversation of about 80 to 120 words between an old man and a young man on science and technology. They hold opposing views on the role and consequences of technology.

Page 1/1

2002



CORRIGE MODELE

32

LV2

[illegible]

It is easy to think

Section One: Reading Comprehension (8 pts)

1. How many paragraphs are there in the above passage? **(½ pt)**

There are three.

2. Choose the general idea of the text. **(1 pt)**

b) The world's polluted oceans.

3. Are these statements True, False or Not Mentioned? **(2 pts)**

a) T b) T c) F d) NM

4. Answer the following questions according to the text. **(3 pts)**

a) Because of their size.

b) Use treatment plants, in spite of their costs.

5. Match words and their definitions. **(1½ pt)**

a	b	c
2	3	1

Section Two: Mastery of Language (8 pts)

1. Add three more words to the list. **(1½ pt)**

environment	pollution	wastes	chemicals	gases, rubbish

2. Supply punctuation and capitalisation. **(1 pt)**

Whales are sea-living mammals. They breathe air but cannot survive on land.

3. Reorder the words to make a coherent sentence. **(1 pt)**

All nuclear power stations produce radioactive wastes which remain dangerous for 1000s of years.

4. Complete the following chart as shown in the example. **(1½ pt)**

Verb	Adjective	Noun	Verb	Adjective	Noun
to think	*thoughtful*	*a thought*	to blame	blameless	blame
to exist	existing	existence	to pollute	polluting	pollution

5. Classify these words according to the pronunciation of their final 's'. **(1½ pt)**

/s/	/z/
wastes, thinks	bodies, chemicals, sons, facilities

6. Rewrite sentence(b)so that it means the same as sentence(a) **(1½ pt)**

b1. We can treat polluted water.

b2. He asked how many casualties had been recorded during the Chernobyl accident.

b3. Our environment is being spoilt by radioactive waste and chemicals.

Section Three: Written Expression (4 pts)

Either topic one.

According to you what are the measures that should be taken to protect our environment?

Or topic two.

This is a conversation between a journalist and a whale hunter. Complete what the journalist says.

Page 1 / 1

CORRIGE MODELE

LV2 Anglais . SNV. BAC 2003
+ SE+TECNO

30

Section One: Reading Comprehension (8 pts)

1. Are there any negative sentences in the third paragraph? If so, how many? ***(½ pt)***

Yes, there are two.

2. Are the following sentences true or false? ***(3 pts)***

a) T b) T c) F d) T

3. Here are the answers to some questions about the text. Ask the questions. ***(3 pts)***

a) What happens to the food we eat?

b) What does the body store?

c) When do we / you burn calories?

4. Find in the text words, or phrases opposite in meaning to the following. ***(1½ pt)***

a) stronger (§ 1) b) accept (§ 3) c) useful (§ 4)

Section Two: Mastery of Language (8 pts)

1. Supply capitals and punctuation. ***(1 pt)***

The next Olympic Games will be held in Athens. Athletes from different parts of the world will take part in the event. The Algerian athletes will certainly represent their country in an honourable way.

2. Divide the following words into roots and affixes. ***(1 pt)***

Prefix	Root	Suffix	Prefix	Root	Suffix	Prefix	Root	Suffix
un	fit			real	ity	in	effect	ive

3. Complete the following chart as shown in the example. ***(1½ pt)***

Verb	Noun	Adjective	Verb	Noun	Adjective
produce	*product*	*productive*	know	knowledge	knowledgeable
think	thought	thoughtful	endanger	danger	dangerous

4. Complete sentence b so that it means the same as sentence a. ***(1½ pt)***

b1. The writer says that the muscle weight will improve the way we look.

b2. Plant cells change solar energy into chemical energy.

b3. After revising (they had revised) English, the candidates slept last night.

5. Reorder these sentences to make a coherent paragraph. ***(1½ pt)***

6. If you eat food that has a value of 3 000 calories

2. and use only 2 600 of them in your activity,

5. the remaining 400 calories will be stored in the body.

3. When you accumulate about 4 000 of these calories,

1. you will gain an extra pound.

One irrelevant sentence must be left out.

4. you will lose 400 calories.

6. Classify the following words according to the pronunciation of their final 'ed'. ***(1½ pt)***

/ t /	/ d /	/ id /
(used) , equipped, reduced	(used) , stored	accepted, discarded

Section Three: Written Expression (4 pts)

Choose ONE of the following topics.

Either topic one: Activity and diet play a beneficial role in man's health.

Or topic two: Do you like to practise sport? Give your reasons.

Page 1 / 1

تكتب الإجابة النموذجية على هذه الورقة ولا تقبل سواها

الإجابة النموذجية لموضوع مقترح لبكالوريا دورة : ....2007...

اختبار مادة : ...الإنجليزية...... شعبة : ع ط ج ......2سا...... المدة : ......2سا......

الإجابة النموذجية 30
وسلم التنقيط

عدد الصفحات : 01

<table>
<tr><th colspan="2">العلامة</th><th rowspan="2">عناصر الإجابة</th><th rowspan="2">محاور الموضوع</th></tr>
<tr><th>المجموع</th><th>مجزأة</th></tr>
<tr><td>1</td><td>1</td><td>1) yes, there are 3. (is called. made up – ~~has been~~ were hosted)</td><td>Section1 (8pts)</td></tr>
<tr><td>2</td><td>4 x 0,5</td><td>2) a) NM – b) F – c) T – d) NM</td><td></td></tr>
<tr><td>0,5</td><td>0,5</td><td>3) The History of Soccer</td><td></td></tr>
<tr><td>1,5</td><td>0,5
0,5
0,5</td><td>4) a) its : soccer
b) the same reasons: Continuous action and fast pace
c) who : top players</td><td></td></tr>
<tr><td>1,5</td><td>3 x 0,5</td><td>5) a) attempt – b) throughout – c) prize</td><td></td></tr>
<tr><td>1,5</td><td>3x0,5</td><td>6) a) most – b) major – c) fast</td><td></td></tr>
<tr><td>1,5</td><td>1,5</td><td>1) 'Mother,' said the little boy, 'I want to see the game.' 'It's all right', 'you may go,' she answered.</td><td>Section2 (8pts)</td></tr>
<tr><td>0,5</td><td>2 x 0,25</td><td>2) a) tennis – b) succeed</td><td></td></tr>
<tr><td>1,5</td><td>3 x 0,5</td><td>3)
<table>
<tr><th>Prefix</th><th>Root</th><th>Suffix</th></tr>
<tr><td></td><td>short</td><td>ening</td></tr>
<tr><td></td><td>athlet(e)</td><td>ic</td></tr>
<tr><td>inter</td><td>nation</td><td>al</td></tr>
</table></td><td></td></tr>
<tr><td>1,5</td><td>3 x 0,5</td><td>4) a) an inflated ball is guided
b) They had hosted ...
c) He neither watched the basketball game nor studied.</td><td></td></tr>
<tr><td>2</td><td>4 x 0,5</td><td>5) a) has – b) taken – c) and – d) runner</td><td></td></tr>
<tr><td>1</td><td>4 x 0,25</td><td>6) 1 syl : drowned + game – 2 syl : against – 3 syl : spectator</td><td></td></tr>
<tr><td>4 pts</td><td></td><td>Written Expression</td><td></td></tr>
</table>

28 KEYS 2005 SN' LV2 Not all Americans

SECTION ONE: READING COMPREHENSION (8 POINTS)

Activity 1. Choose a title to the text. (1)

b) Holidays in America

Activity 2. Answer these questions according to the text. (2)

a) The importance of the holiday / the demand / of seasonal work

b) No. Activities differ.

Activity 3. In paragraph one. (1)

Activity 4. Here are the answers to questions about the text. Write the questions. (2)

a) When do most Americans take their vacations?

b) What is given after longer periods of work?

Activity 5. Find in the text words that are closest in meaning. (2)

a) varies (§1) b) common (§1) c) brief (§2) d) customs (§2)

SECTION TWO: MASTERY OF LANGUAGE (8 POINTS)

Activity 1. Supply punctuation and capitals where necessary. (1.5)

The American student spends six hours a day, five days a week, 180 days a year in school. Children in the United States start pre-school at the age of four.

Activity 2. Which adjectives can be derived from the following nouns? (1.5)

a) alive / living b) long c) childish

Activity 3. Every sentence contains one mistake. Write the sentence without the mistake. (1)

a) Money brings money.

b) Nowadays, people do not like (to read) or (reading).

Activity 4. Combine the following pairs using the connector provided. (2)

a) WHILE I was reading the newspaper the telephone rang.

b) AS SOON AS you get to London, (you must / should / will…) start speaking English.

Activity 5. Read the passage and delete the unnecessary words. (1)

In their free time, (the) students spend much time watching TV. They also listen to music.

Activity 6. Classify the following words according to the pronunciation of 'ed'. (1)

/t/	/d/	/id/
worked	showed	reported / depended

SECTION THREE: WRITTEN EXPRESSION (4 POINTS)

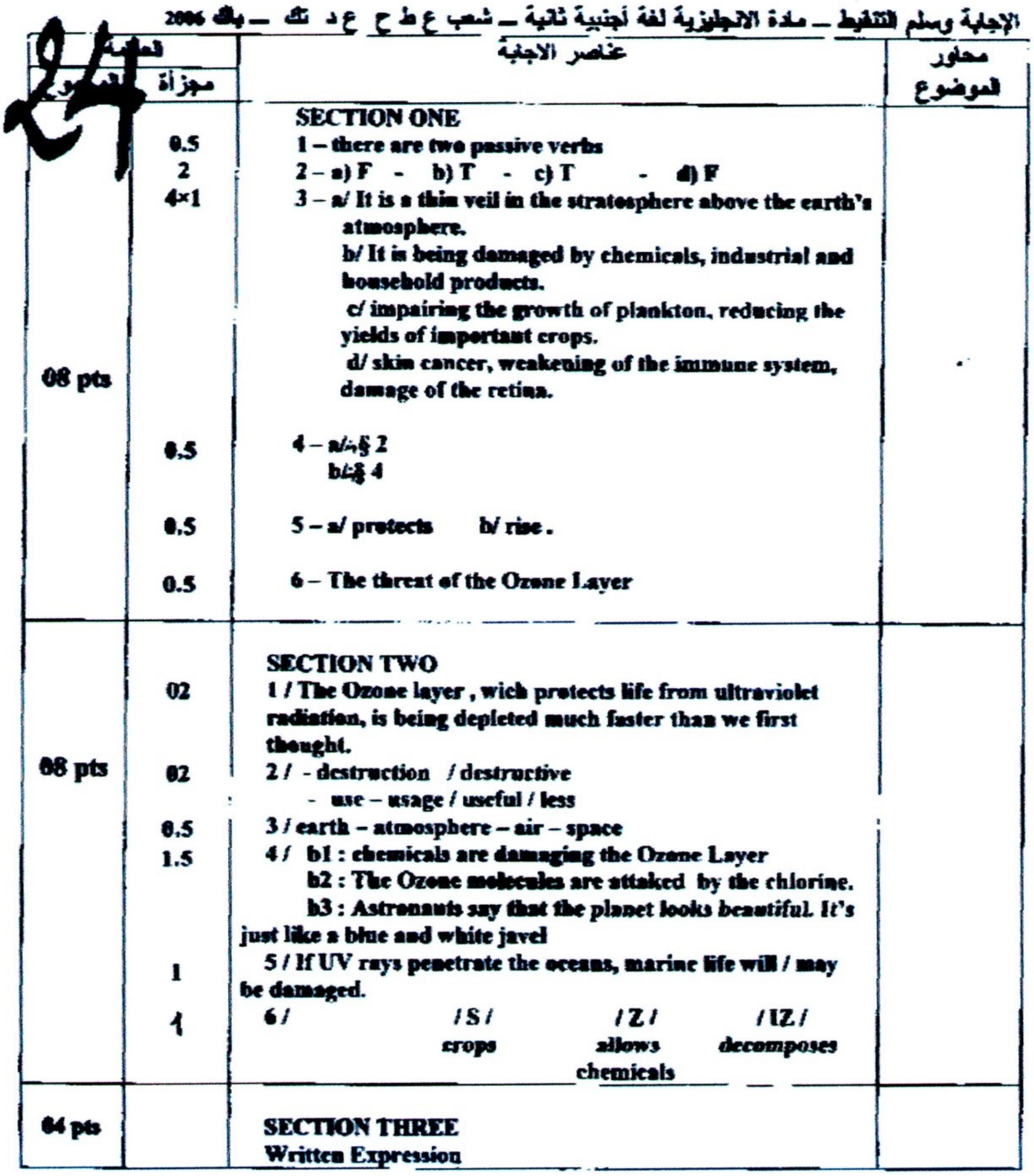
Choose one of the following topics.

Topic one

Using the following notes, write a composition of about 80 to 120 words on **Algerians and holidays.**

Topic two

Do you prefer to spend your summer holidays in your country or abroad? Give your reasons.

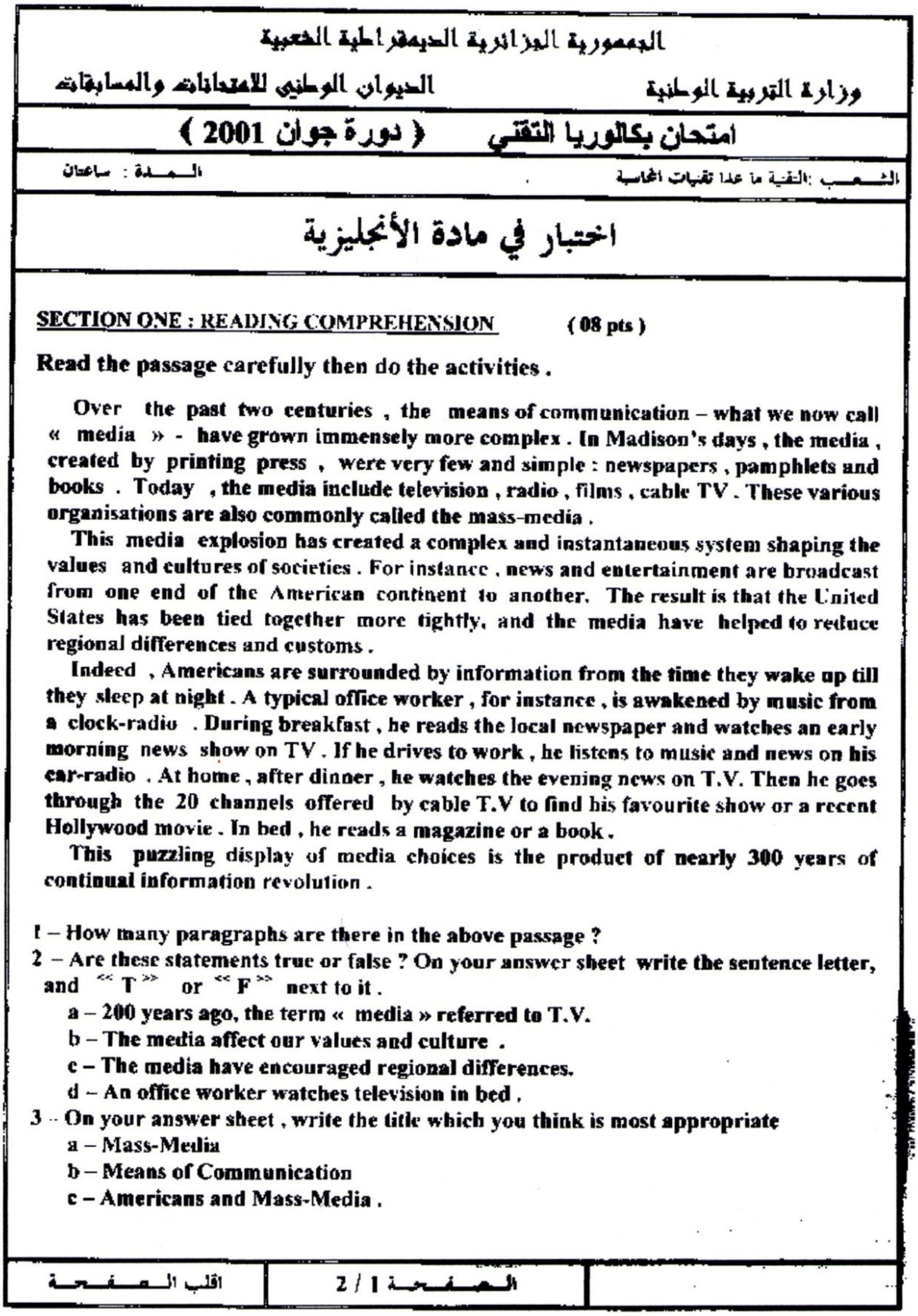


الإجابة وسلم التنقيط – مادة الانجليزية لغة أجنبية ثانية – شعب ع ط ح ع د تك – باك 2006

24

العلامة		عناصر الاجابة	محاور الموضوع
المجموع	مجزأة		
		SECTION ONE	
	0.5	1 – there are two passive verbs	
	2	2 – a) F - b) T - c) T - d) F	
	4×1	3 – a/ It is a thin veil in the stratosphere above the earth's atmosphere. b/ It is being damaged by chemicals, industrial and household products. c/ impairing the growth of plankton, reducing the yields of important crops. d/ skin cancer, weakening of the immune system, damage of the retina.	
08 pts			
	0.5	4 – a/→§ 2 b/→§ 4	
	0.5	5 – a/ protects b/ rise.	
	0.5	6 – The threat of the Ozone Layer	
		SECTION TWO	
	02	1 / The Ozone layer , wich protects life from ultraviolet radiation, is being depleted much faster than we first thought.	
08 pts	02	2 / - destruction / destructive - use – usage / useful / less	
	0.5	3 / earth – atmosphere – air – space	
	1.5	4 / b1 : chemicals are damaging the Ozone Layer b2 : The Ozone molecules are attaked by the chlorine. b3 : Astronauts say that the planet looks beautiful. It's just like a blue and white javel	
	1	5 / If UV rays penetrate the oceans, marine life will / may be damaged.	
	1	6 / /S/ *crops* – /Z/ *allows* chemicals – /IZ/ *decomposes*	
04 pts		**SECTION THREE** Written Expression	

Appendix D: Technical Streams' BAC English Tests from 2001 to 2006



الجمهورية الجزائرية الديمقراطية الشعبية

الديوان الوطني للامتحانات والمسابقات وزارة التربية الوطنية

امتحان بكالوريا التقني (دورة جوان 2001)

الشعب :التقنية ما عدا تقنيات المحاسبة المدة : ساعتان

اختبار في مادة الأنجليزية

SECTION ONE : READING COMPREHENSION (08 pts)

Read the passage carefully then do the activities .

Over the past two centuries , the means of communication – what we now call « media » - have grown immensely more complex . In Madison's days , the media , created by printing press , were very few and simple : newspapers , pamphlets and books . Today , the media include television , radio , films , cable TV . These various organisations are also commonly called the mass-media .

This media explosion has created a complex and instantaneous system shaping the values and cultures of societies . For instance , news and entertainment are broadcast from one end of the American continent to another. The result is that the United States has been tied together more tightly, and the media have helped to reduce regional differences and customs .

Indeed , Americans are surrounded by information from the time they wake up till they sleep at night . A typical office worker , for instance , is awakened by music from a clock-radio . During breakfast , he reads the local newspaper and watches an early morning news show on TV . If he drives to work , he listens to music and news on his car-radio . At home , after dinner , he watches the evening news on T.V. Then he goes through the 20 channels offered by cable T.V to find his favourite show or a recent Hollywood movie . In bed , he reads a magazine or a book .

This puzzling display of media choices is the product of nearly 300 years of continual information revolution .

1 – How many paragraphs are there in the above passage ?

2 – Are these statements true or false ? On your answer sheet write the sentence letter, and « T » or « F » next to it .

a – 200 years ago, the term « media » referred to T.V.

b – The media affect our values and culture .

c – The media have encouraged regional differences.

d – An office worker watches television in bed .

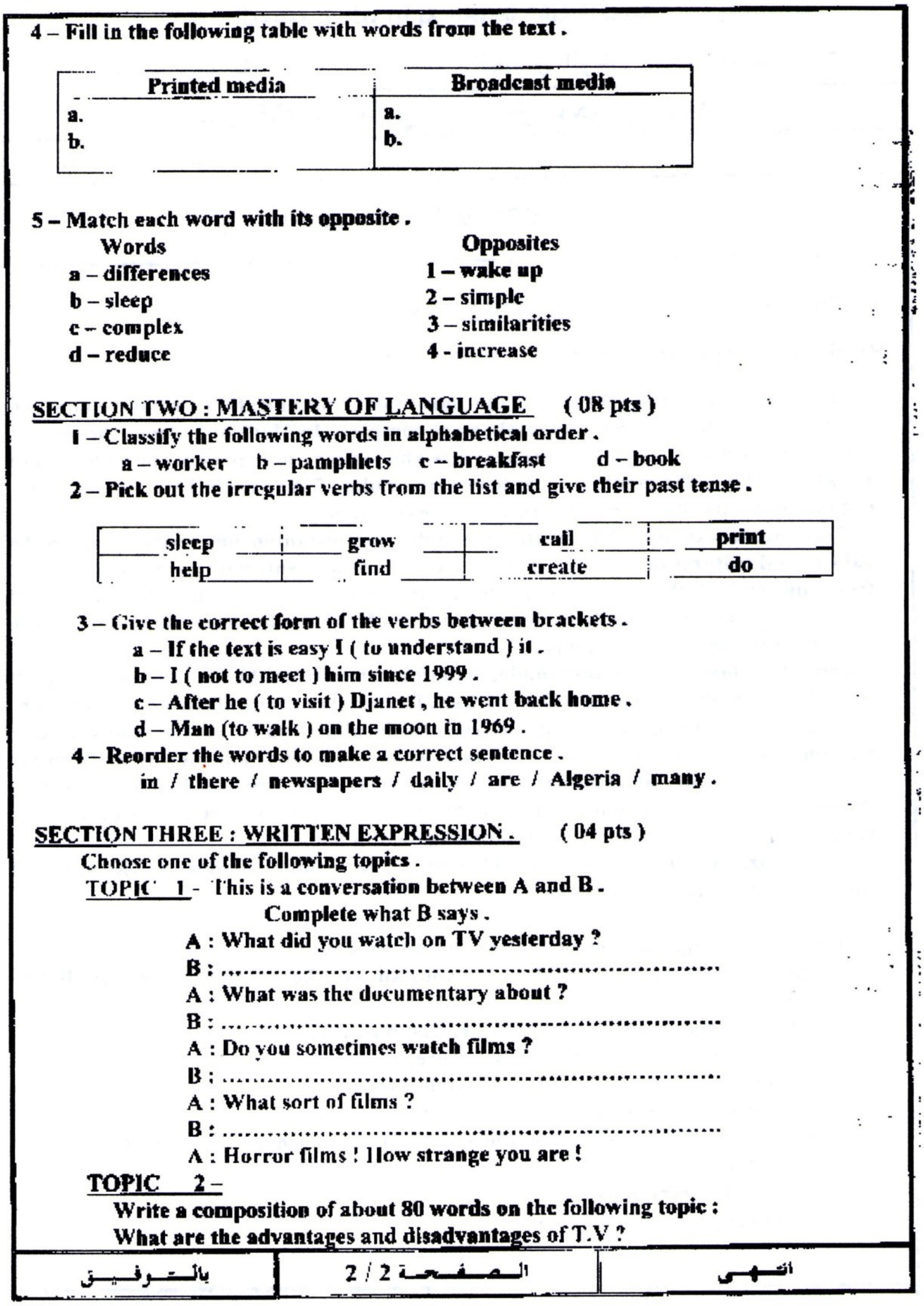
3 – On your answer sheet , write the title which you think is most appropriate

a – Mass-Media

b – Means of Communication

c – Americans and Mass-Media .

اقلب الصفحة	الصفحة 1 / 2	



4 – Fill in the following table with words from the text .

Printed media	Broadcast media
a.	a.
b.	b.

5 – Match each word with its opposite .

Words	Opposites
a – differences	1 – wake up
b – sleep	2 – simple
c – complex	3 – similarities
d – reduce	4 - increase

SECTION TWO : MASTERY OF LANGUAGE (08 pts)

1 – Classify the following words in alphabetical order .

a – worker b – pamphlets c – breakfast d – book

2 – Pick out the irregular verbs from the list and give their past tense .

sleep	grow	call	print
help	find	create	do

3 – Give the correct form of the verbs between brackets .

a – If the text is easy I (to understand) it .

b – I (not to meet) him since 1999 .

c – After he (to visit) Djanet , he went back home .

d – Man (to walk) on the moon in 1969 .

4 – Reorder the words to make a correct sentence .

in / there / newspapers / daily / are / Algeria / many .

SECTION THREE : WRITTEN EXPRESSION . (04 pts)

Choose one of the following topics .

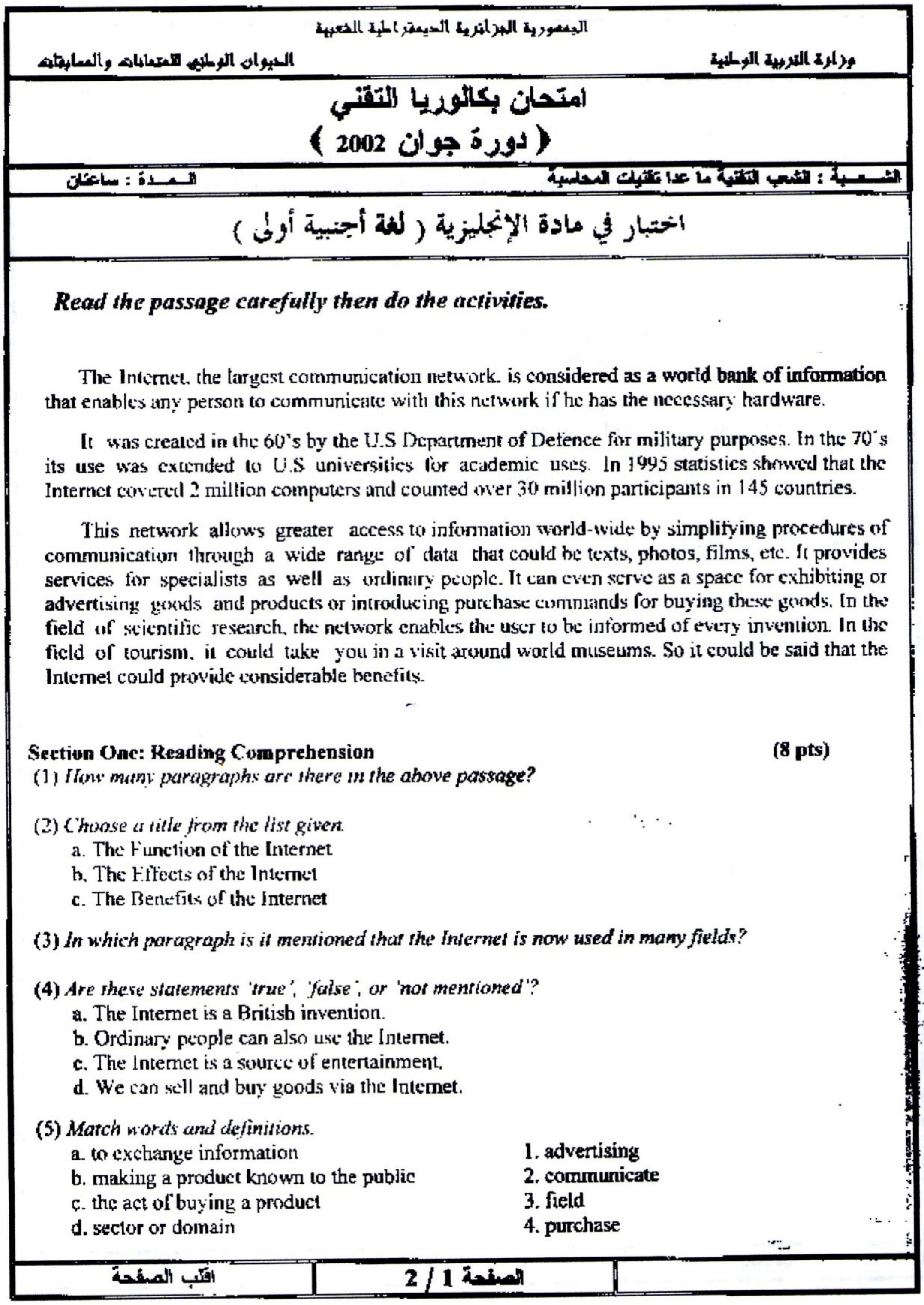
TOPIC 1 - This is a conversation between A and B .
Complete what B says .

A : What did you watch on TV yesterday ?
B : ....
A : What was the documentary about ?
B : ....
A : Do you sometimes watch films ?
B : ....
A : What sort of films ?
B : ....
A : Horror films ! How strange you are !

TOPIC 2 –

Write a composition of about 80 words on the following topic :
What are the advantages and disadvantages of T.V ?

بالتوفيق	الصفحة 2 / 2	انتهى



الجمهورية الجزائرية الديمقراطية الشعبية

الديوان الوطني للامتحانات والمسابقات — وزارة التربية الوطنية

امتحان بكالوريا التقني

(دورة جوان 2002)

الشعبة : الشعب التقنية ما عدا تقنيات المحاسبة — المدة : ساعتان

اختبار في مادة الإنجليزية (لغة أجنبية أولى)

Read the passage carefully then do the activities.

The Internet, the largest communication network, is considered as a world bank of information that enables any person to communicate with this network if he has the necessary hardware.

It was created in the 60's by the U.S Department of Defence for military purposes. In the 70's its use was extended to U.S universities for academic uses. In 1995 statistics showed that the Internet covered 2 million computers and counted over 30 million participants in 145 countries.

This network allows greater access to information world-wide by simplifying procedures of communication through a wide range of data that could be texts, photos, films, etc. It provides services for specialists as well as ordinary people. It can even serve as a space for exhibiting or advertising goods and products or introducing purchase commands for buying these goods. In the field of scientific research, the network enables the user to be informed of every invention. In the field of tourism, it could take you in a visit around world museums. So it could be said that the Internet could provide considerable benefits.

Section One: Reading Comprehension **(8 pts)**

(1) *How many paragraphs are there in the above passage?*

(2) *Choose a title from the list given.*
a. The Function of the Internet
b. The Effects of the Internet
c. The Benefits of the Internet

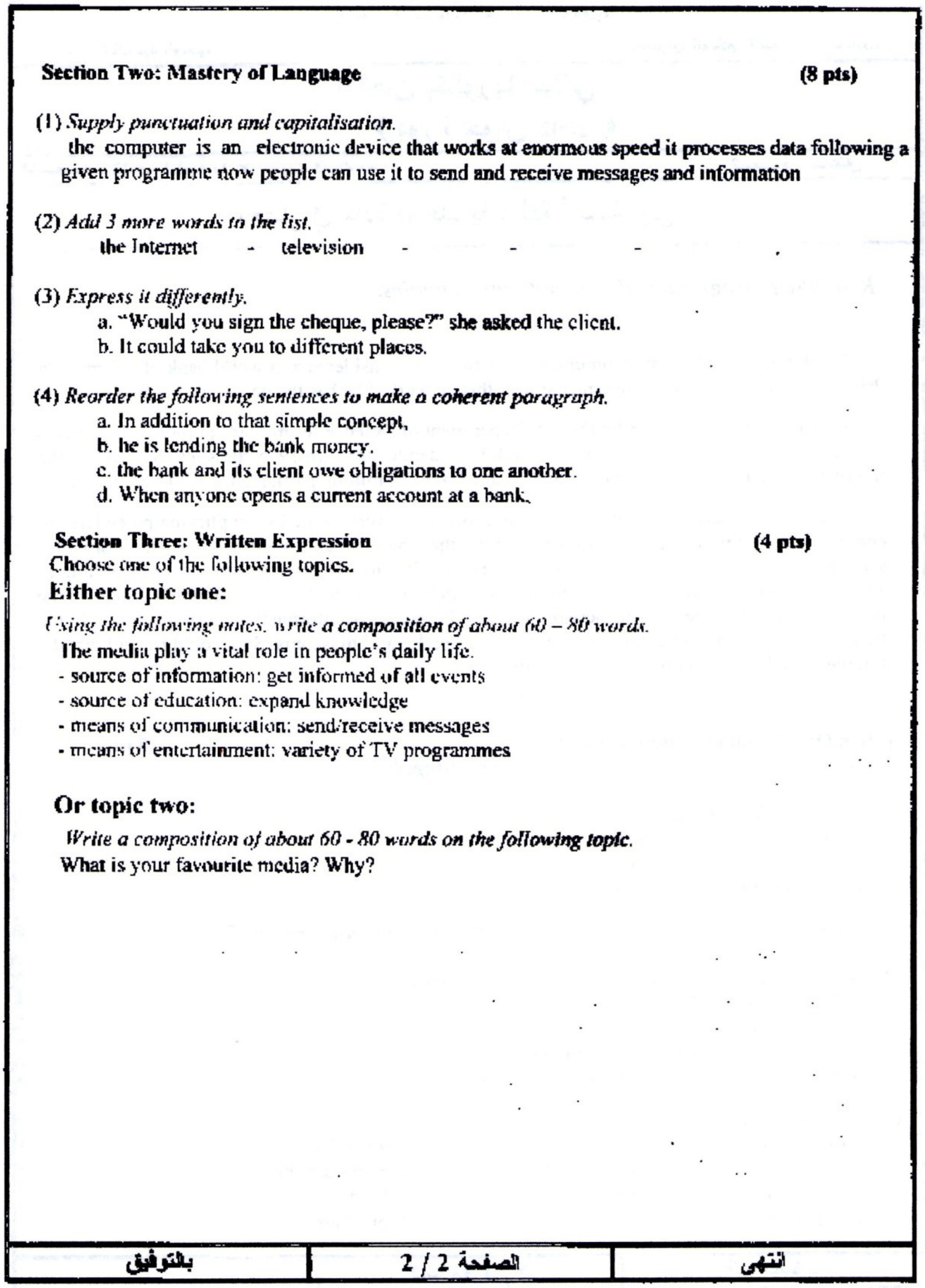
(3) ***In which paragraph is it mentioned that the Internet is now used in many fields?***

(4) *Are these statements 'true', 'false', or 'not mentioned'?*
a. The Internet is a British invention.
b. Ordinary people can also use the Internet.
c. The Internet is a source of entertainment.
d. We can sell and buy goods via the Internet.

(5) *Match words and definitions.*

a. to exchange information	1. advertising
b. making a product known to the public	2. communicate
c. the act of buying a product	3. field
d. sector or domain	4. purchase

اقلب الصفحة — الصفحة 1 / 2



Section Two: Mastery of Language **(8 pts)**

(1) *Supply punctuation and capitalisation.*

the computer is an electronic device that works at enormous speed it processes data following a given programme now people can use it to send and receive messages and information

(2) *Add 3 more words to the list.*

the Internet - television - - - .

(3) *Express it differently.*

a. "Would you sign the cheque, please?" she asked the client.

b. It could take you to different places.

(4) *Reorder the following sentences to make a coherent paragraph.*

a. In addition to that simple concept,

b. he is lending the bank money.

c. the bank and its client owe obligations to one another.

d. When anyone opens a current account at a bank,

Section Three: Written Expression **(4 pts)**

Choose one of the following topics.

Either topic one:

Using the following notes, write a composition of about 60 – 80 words.

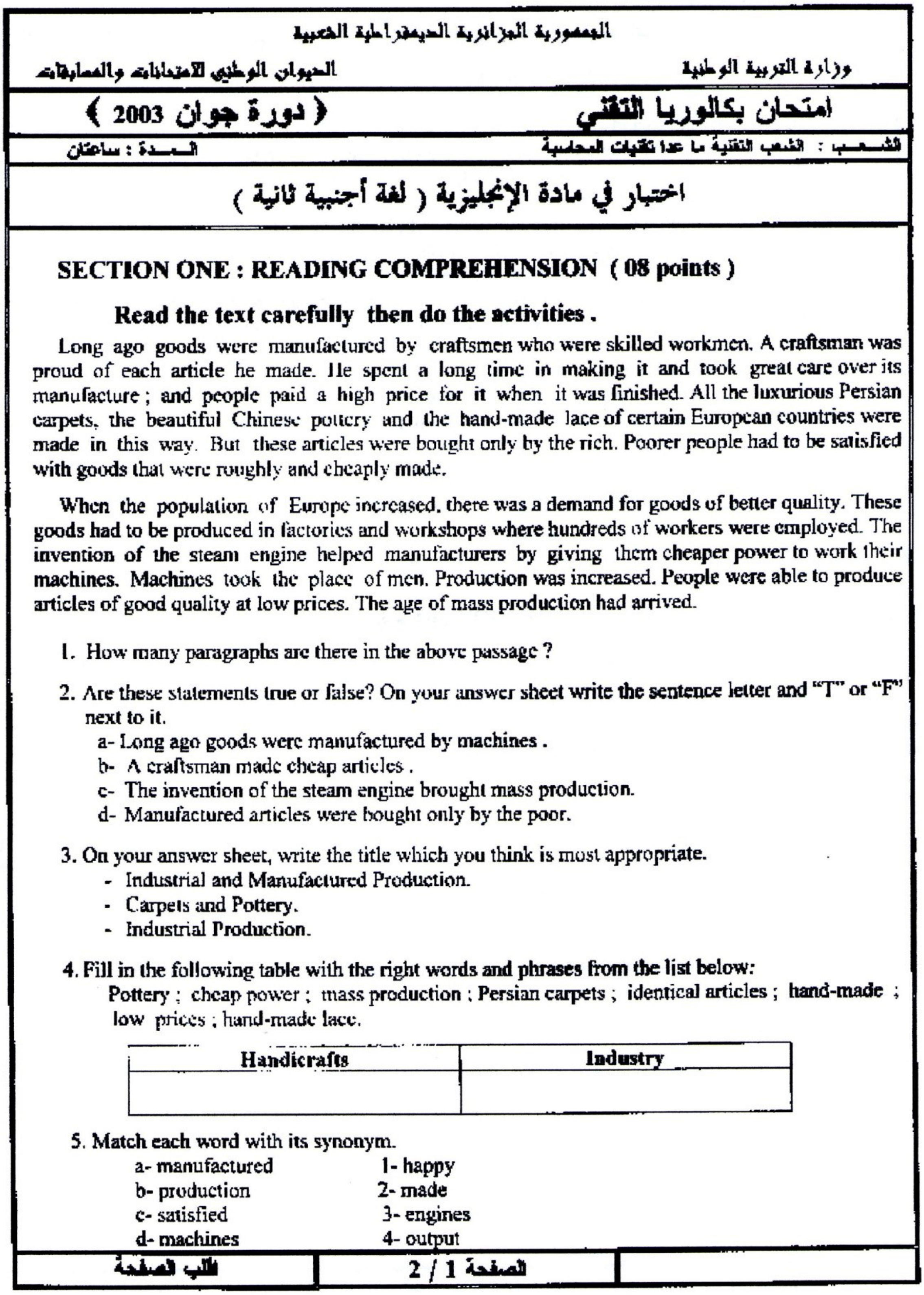
The media play a vital role in people's daily life.
- source of information: get informed of all events
- source of education: expand knowledge
- means of communication: send/receive messages
- means of entertainment: variety of TV programmes

Or topic two:

Write a composition of about 60 - 80 words on the following topic.

What is your favourite media? Why?

بالتوفيق	الصفحة 2 / 2	انتهى



الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التربية الوطنية — الديوان الوطني للامتحانات والمسابقات

امتحان بكالوريا التقني (دورة جوان 2003)

الشعب : الشعب التقنية ما عدا تقنيات المحاسبة — المدة : ساعتان

اختبار في مادة الإنجليزية (لغة أجنبية ثانية)

SECTION ONE : READING COMPREHENSION (08 points)

Read the text carefully then do the activities .

Long ago goods were manufactured by craftsmen who were skilled workmen. A craftsman was proud of each article he made. He spent a long time in making it and took great care over its manufacture ; and people paid a high price for it when it was finished. All the luxurious Persian carpets, the beautiful Chinese pottery and the hand-made lace of certain European countries were made in this way. But these articles were bought only by the rich. Poorer people had to be satisfied with goods that were roughly and cheaply made.

When the population of Europe increased, there was a demand for goods of better quality. These goods had to be produced in factories and workshops where hundreds of workers were employed. The invention of the steam engine helped manufacturers by giving them cheaper power to work their machines. Machines took the place of men. Production was increased. People were able to produce articles of good quality at low prices. The age of mass production had arrived.

1. How many paragraphs are there in the above passage ?

2. Are these statements true or false? On your answer sheet write the sentence letter and "T" or "F" next to it.
   a- Long ago goods were manufactured by machines .
   b- A craftsman made cheap articles .
   c- The invention of the steam engine brought mass production.
   d- Manufactured articles were bought only by the poor.

3. On your answer sheet, write the title which you think is most appropriate.
   - Industrial and Manufactured Production.
   - Carpets and Pottery.
   - Industrial Production.

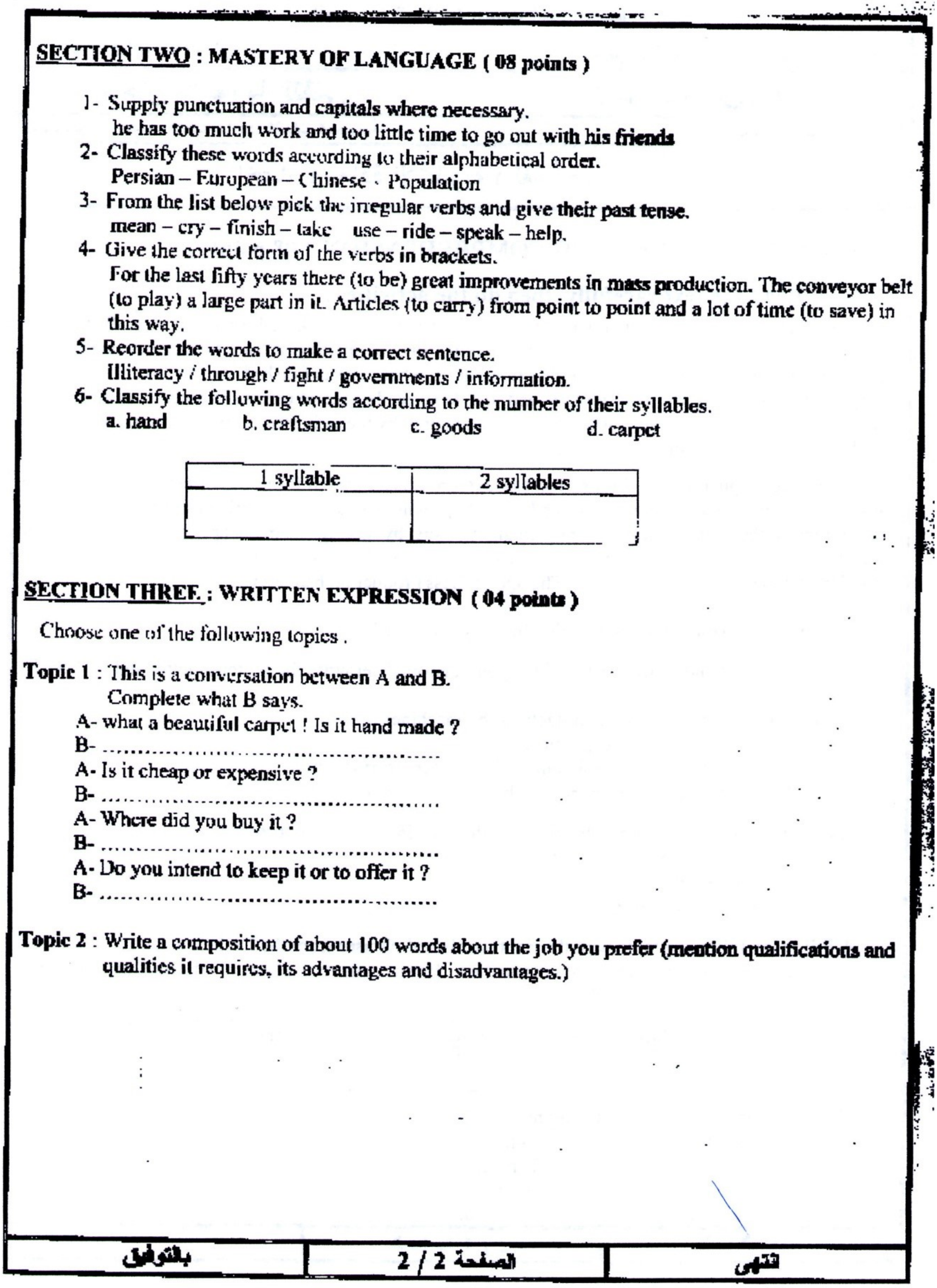
4. Fill in the following table with the right words and phrases from the list below:
   Pottery ; cheap power ; mass production ; Persian carpets ; identical articles ; hand-made ; low prices ; hand-made lace.

Handicrafts	Industry

5. Match each word with its synonym.

a- manufactured	1- happy
b- production	2- made
c- satisfied	3- engines
d- machines	4- output

اقلب الصفحة — الصفحة 1 / 2



SECTION TWO : MASTERY OF LANGUAGE (08 points)

1- Supply punctuation and capitals where necessary.
he has too much work and too little time to go out with his friends

2- Classify these words according to their alphabetical order.
Persian – European – Chinese - Population

3- From the list below pick the irregular verbs and give their past tense.
mean – cry – finish – take use – ride – speak – help.

4- Give the correct form of the verbs in brackets.
For the last fifty years there (to be) great improvements in mass production. The conveyor belt (to play) a large part in it. Articles (to carry) from point to point and a lot of time (to save) in this way.

5- Reorder the words to make a correct sentence.
Illiteracy / through / fight / governments / information.

6- Classify the following words according to the number of their syllables.
a. hand b. craftsman c. goods d. carpet

1 syllable	2 syllables

SECTION THREE : WRITTEN EXPRESSION (04 points)

Choose one of the following topics .

Topic 1 : This is a conversation between A and B.
Complete what B says.

A- what a beautiful carpet ! Is it hand made ?
B- ....
A- Is it cheap or expensive ?
B- ....
A- Where did you buy it ?
B- ....
A- Do you intend to keep it or to offer it ?
B- ....

Topic 2 : Write a composition of about 100 words about the job you prefer (mention qualifications and qualities it requires, its advantages and disadvantages.)

بالتوفيق	الصفحة 2 / 2	انتهى

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التربية الوطنية الديوان الوطني للامتحانات والمسابقات

امتحان بكالوريا التقني (دورة جوان 2004)

الشعـــب : التقنية ما عدا تقنيات الحاسبة. المـــدة : ساعتان

اختبار في مادة الإنجليزية (لغة أجنبية ثانية)

SECTION ONE : READING COMPREHENSION (08 pts)

Read the passage and do the activities.

Automation and Society

By the mid 1980's automation made progress in manufacturing and in the service industries. Automation brought both benefits and challenges to the larger society.

Automation often reduces cost and improves production both in terms of quantity and quality. If properly used, it can free workers from unpleasant and hazardous jobs. In a growing number of factories, robots are programmed to perform dull and repetitive tasks and to load and unload heavy objects. Various cities have provided professional fire fighters with robot device that can be used into burning buildings in danger of collapse.

In spite of its beneficial effects, increased automation can cause serious problems for workers in manufacturing plants. In many plants, production levels have been greatly increased, while the number of workers has been reduced.

Large factories which once required thousands of employees, need only a few hundred today.

Adapted from : « Articles from the Department of Technology and Society. New york. »

1 – How many paragraphs are there in the text ?

2 – Are these statements true or false ?On your answer sheet, write « T » or « F » next to the sentence letter.

a- Automation reduces cost and products.

b- Robots are used to fight fire.

c- Automation releases workers from dull and dangerous jobs.

d- With automation many workers get new jobs.

3 – Answer the following questions according to the text.

a- Which fields has automation developed ?

b- Give two advantages and one drawback of automation.

4 – What or who do the underlined words refer to in the text ?

a- ... it can free workers from... (§ 2)

b- ... robot device that can be used ... (§ 2)

c- In spite of its beneficial ... (§ 3)

5 – Match each word with its corresponding definition.

Words	Definitions
a- challenges	1. use of machines to do tasks done by people.
b- automation	2. makes better

c- improves 3. not exciting / boring

d- dull 4. problems

SECTION TWO : MASTERY OF LANGUAGE (08 pts)

1 – Supply punctuation and capitals where necessary.

unlike many countries america does not have an official apprenticeship programme

2 – Derive adjectives from the following nouns.

use - danger - automation - technology

3 – Rewrite sentence(b) so that it means the same as sentence(a)

a. 1. The writer said, « It takes me a long time to dial the number ».

b. 1. The writer said that…………………

a. 2.The writer said, « What can people use the internet for ? »

b. 2. The writer asked……………………..

a. 3. Automation has increased the need for specialsts in electronics.

b. 3. The need for specialists in electronics…………………………….

4 – Reorder the following sentences to write a coherent paragraph.

a. It provides information on millions of different subjects ;

b. This system is called electronic mail or e-mail.

c. The Internet is the fastest communication system in human history.

d. it also enables people to write to each other electronically.

SECTION THREE : WRITTEN EXPRESSION (04 pts)

Choose one of the following topics.

TOPIC 1 : This is a conversation between A and B . Complete what B says

A : Is automation beneficial for Man ?

B : …………………

A : Why not ?

B : …………………

A : What sort of negative effects ?

B : ……………………..

A : Automation is not such a good thing , then ?

B :……………………..

TOPIC 2 :

Write a composition of about 80 words on the following topic :

Do you think it is possible to live without computers today ?

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التربية الوطنية الديوان الوطني للامتحانات والمسابقات

امتحان بكالوريا التقني (دورة جوان 2005)

الشعب : التقنية ما عدا تقنيات الحاسبة. المدة : ساعتان

اختبار في مادة الإنجليزية (لغة أجنبية ثانية)

SECTION I : READING COMPREHENSION (08 pts)

Read the passage carefully then do the activities.

According to your needs, you can now choose the way to deliver goods through a variety of means. If you have to transport heavy freight over long distances, the rail is a suitable form of carriage. But remember that an additional means of transport will be necessary to deliver the goods of the customers.

Road transport is a more convenient form of carriage to bring the goods - such as fruit or vegetables – from the wholesaler to the retailer. The road system offers a genuine door – to – door service. Moreover, on motorways, lorries with their trailers, carry loads of up to twenty tons.

If speed is not your chief concern, sea carriage remains less expensive and is effected by cargo boats and trampsteamers. Cargo liners can carry quantities of goods.

If you are in a hurry, the ability to deliver goods within a few hours to any part of the world is air transport. Its main advantage is that mail, films or videotapes, newspapers, pharmaceuticals go by air round the clock.

Adapted from : « Guide de l'Anglais et de
l'Americain des Affaires »
By G. BAXTER , A. LAVIGNAC

1 – How many paragraphs are there in the above passage ?

2 – Match each title with its corresponding paragraph.

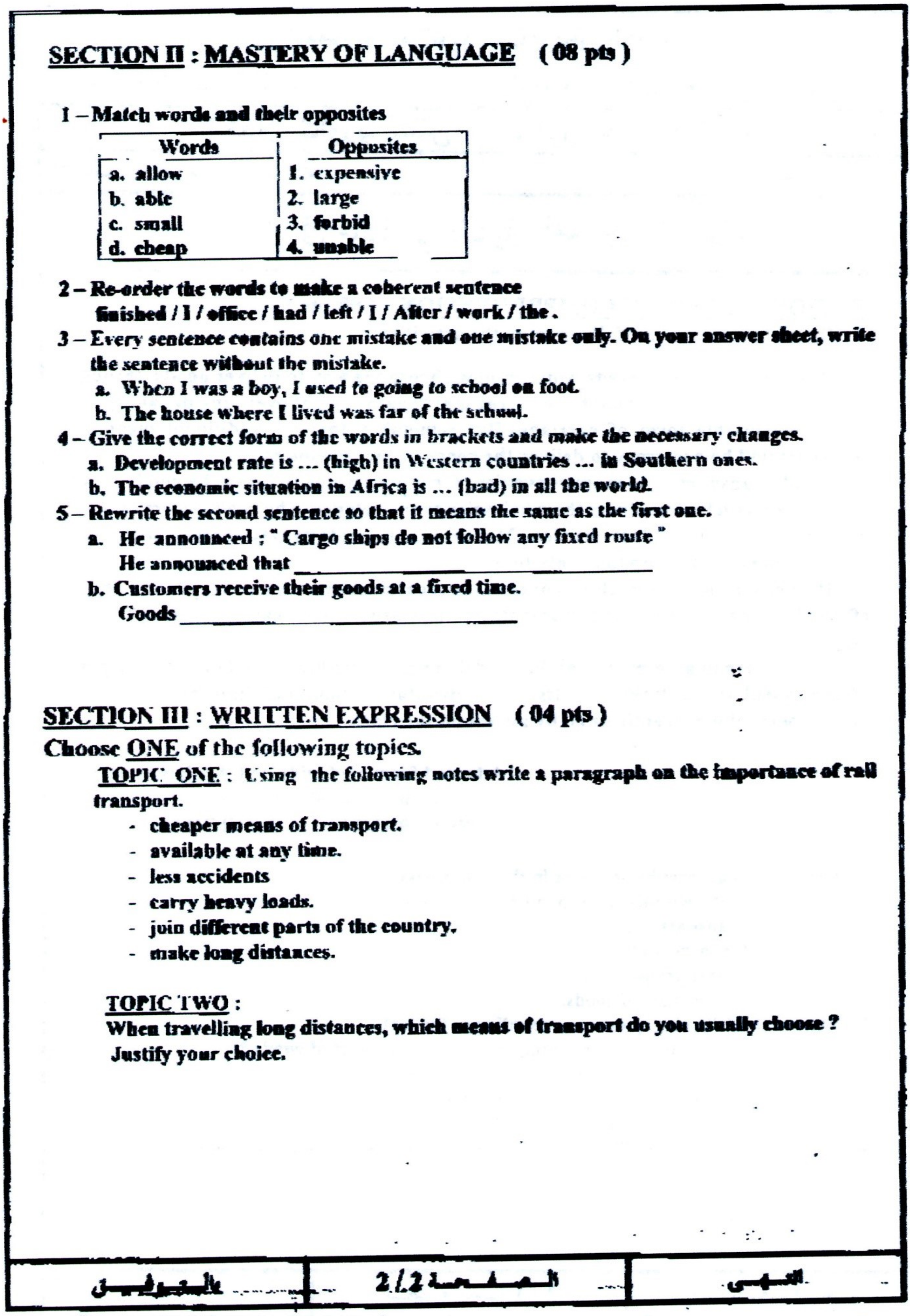
a- Air transport of goods.
b- Rail transport of goods.
c- Sea transport of goods.
d- Road transport of goods.

3 – Answer the following questions according to the text.

a- What is the main disadvantage of the rail transport of goods ?
b- What are the advantages of road transportation ?
c- Do people, in a hurry, use sea carriage ?
d- Why is air transport necessary ?

4 – Find in the text words, phrases or expressions that are closest in meaning to the following :

a- appropriate (§ 1) c- transport (§ 3)
b- Authentic (§ 2) d- main (§ 3)



SECTION II : MASTERY OF LANGUAGE (08 pts)

1 – Match words and their opposites

Words	Opposites
a. allow	1. expensive
b. able	2. large
c. small	3. forbid
d. cheap	4. unable

2 – Re-order the words to make a coherent sentence
finished / I / office / had / left / I / After / work / the .

3 – Every sentence contains one mistake and one mistake only. On your answer sheet, write the sentence without the mistake.

a. When I was a boy, I used to going to school on foot.
b. The house where I lived was far of the school.

4 – Give the correct form of the words in brackets and make the necessary changes.

a. Development rate is ... (high) in Western countries ... in Southern ones.
b. The economic situation in Africa is ... (bad) in all the world.

5 – Rewrite the second sentence so that it means the same as the first one.

a. He announced ; " Cargo ships do not follow any fixed route "
He announced that ____________________________________
b. Customers receive their goods at a fixed time.
Goods ____________________________________

SECTION III : WRITTEN EXPRESSION (04 pts)

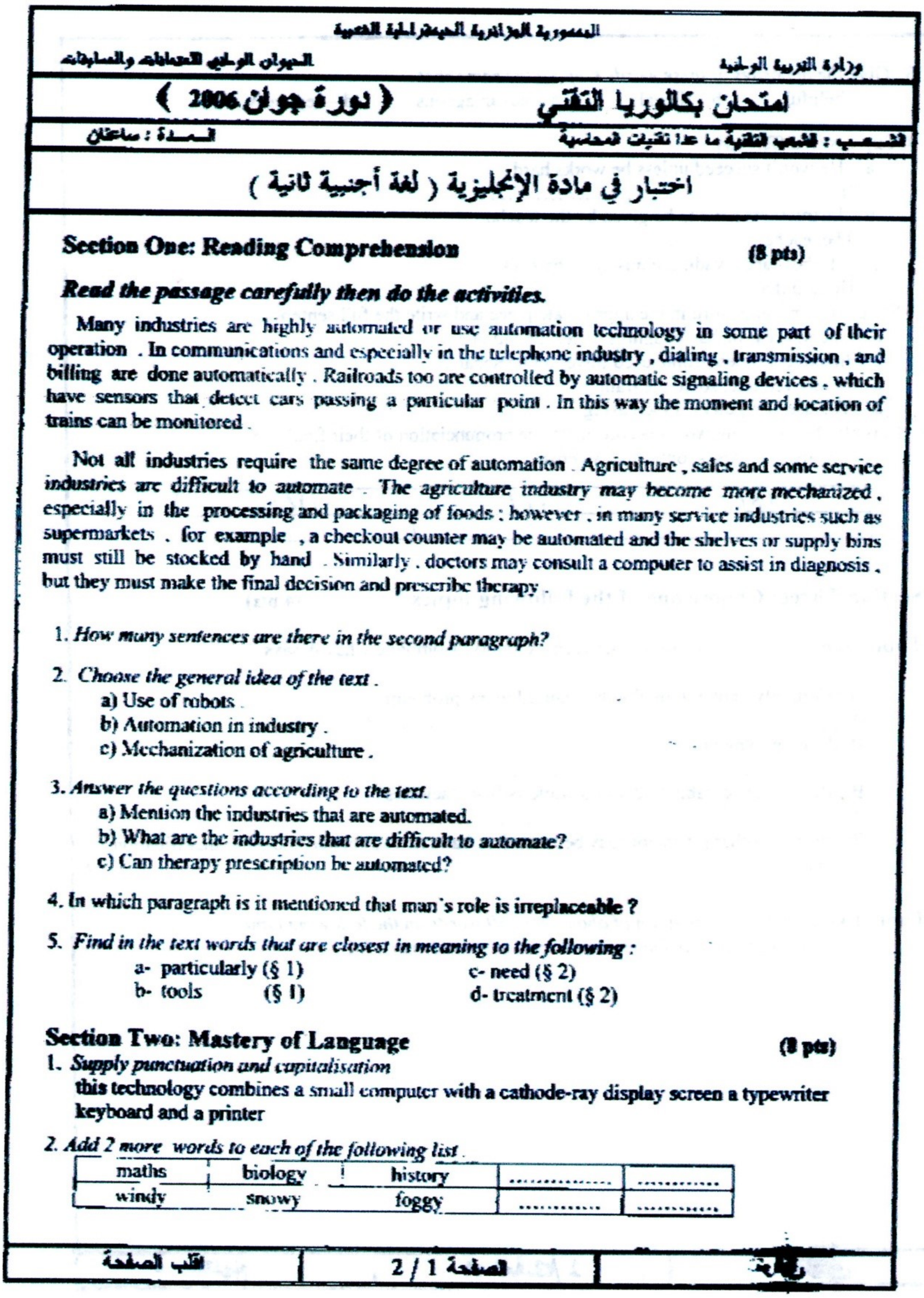
Choose ONE of the following topics.

TOPIC ONE : Using the following notes write a paragraph on the importance of rail transport.

- cheaper means of transport.
- available at any time.
- less accidents
- carry heavy loads.
- join different parts of the country.
- make long distances.

TOPIC TWO :
When travelling long distances, which means of transport do you usually choose ? Justify your choice.

بالتوفيق	الصفحة 2/2	انتهى.



الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التربية الوطنية — الديوان الوطني للامتحانات والمسابقات

امتحان بكالوريا التقني (دورة جوان 2006)

الشعـــب : الشعب التقنية ما عدا تقنيات المحاسبة — المـــدة : ساعتان

اختبار في مادة الإنجليزية (لغة أجنبية ثانية)

Section One: Reading Comprehension (8 pts)

Read the passage carefully then do the activities.

Many industries are highly automated or use automation technology in some part of their operation . In communications and especially in the telephone industry , dialing , transmission , and billing are done automatically . Railroads too are controlled by automatic signaling devices , which have sensors that detect cars passing a particular point . In this way the moment and location of trains can be monitored .

Not all industries require the same degree of automation . Agriculture , sales and some service industries are difficult to automate . The agriculture industry may become more mechanized , especially in the processing and packaging of foods ; however , in many service industries such as supermarkets , for example , a checkout counter may be automated and the shelves or supply bins must still be stocked by hand . Similarly , doctors may consult a computer to assist in diagnosis , but they must make the final decision and prescribe therapy .

1. *How many sentences are there in the second paragraph?*

2. *Choose the general idea of the text .*
   a) Use of robots .
   b) Automation in industry .
   c) Mechanization of agriculture .

3. *Answer the questions according to the text.*
   a) Mention the industries that are automated.
   b) What are the industries that are difficult to automate?
   c) Can therapy prescription be automated?

4. In which paragraph is it mentioned that man's role is irreplaceable ?

5. *Find in the text words that are closest in meaning to the following :*

a- particularly (§ 1)	c- need (§ 2)
b- tools (§ 1)	d- treatment (§ 2)

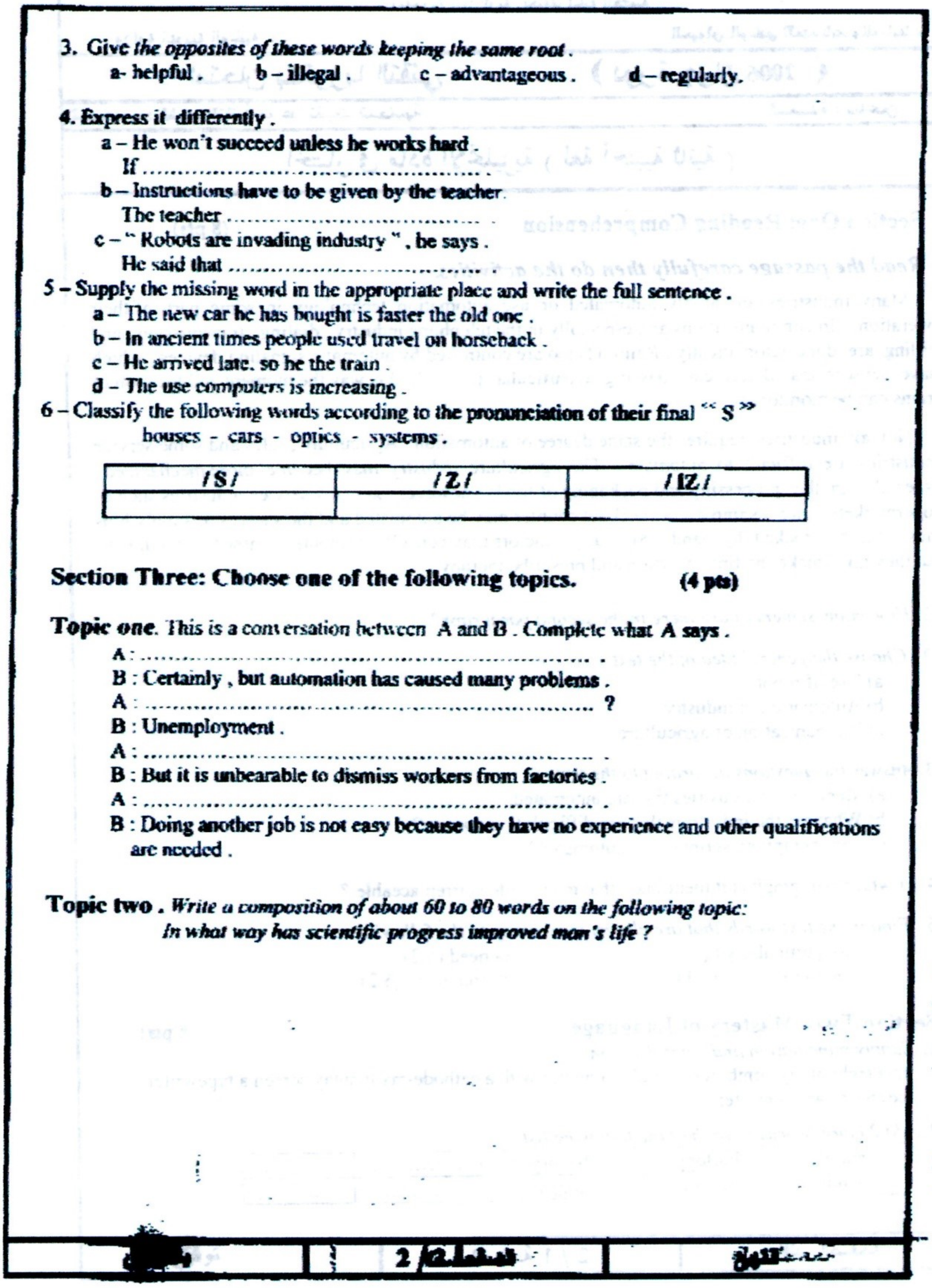
Section Two: Mastery of Language (8 pts)

1. *Supply punctuation and capitalisation*
   this technology combines a small computer with a cathode-ray display screen a typewriter keyboard and a printer

2. *Add 2 more words to each of the following list .*

maths	biology	history	....	....
windy	snowy	foggy	....	....

اقلب الصفحة — الصفحة 1 / 2



3. Give *the opposites of these words keeping the same root* .

a- helpful . b – illegal . c – advantageous . d – regularly.

4. Express it differently .

a – He won't succeed unless he works hard .

If ....

b – Instructions have to be given by the teacher.

The teacher ....

c – " Robots are invading industry " , he says .

He said that ....

5 – Supply the missing word in the appropriate place and write the full sentence .

a – The new car he has bought is faster the old one .

b – In ancient times people used travel on horseback .

c – He arrived late. so he the train .

d – The use computers is increasing .

6 – Classify the following words according to the pronunciation of their final " S "

houses - cars - optics - systems .

/S/	/Z/	/IZ/

Section Three: Choose one of the following topics. (4 pts)

Topic one. This is a conversation between A and B . Complete what A says .

A : ....

B : Certainly , but automation has caused many problems .

A : .... ?

B : Unemployment .

A : ....

B : But it is unbearable to dismiss workers from factories .

A : ....

B : Doing another job is not easy because they have no experience and other qualifications are needed .

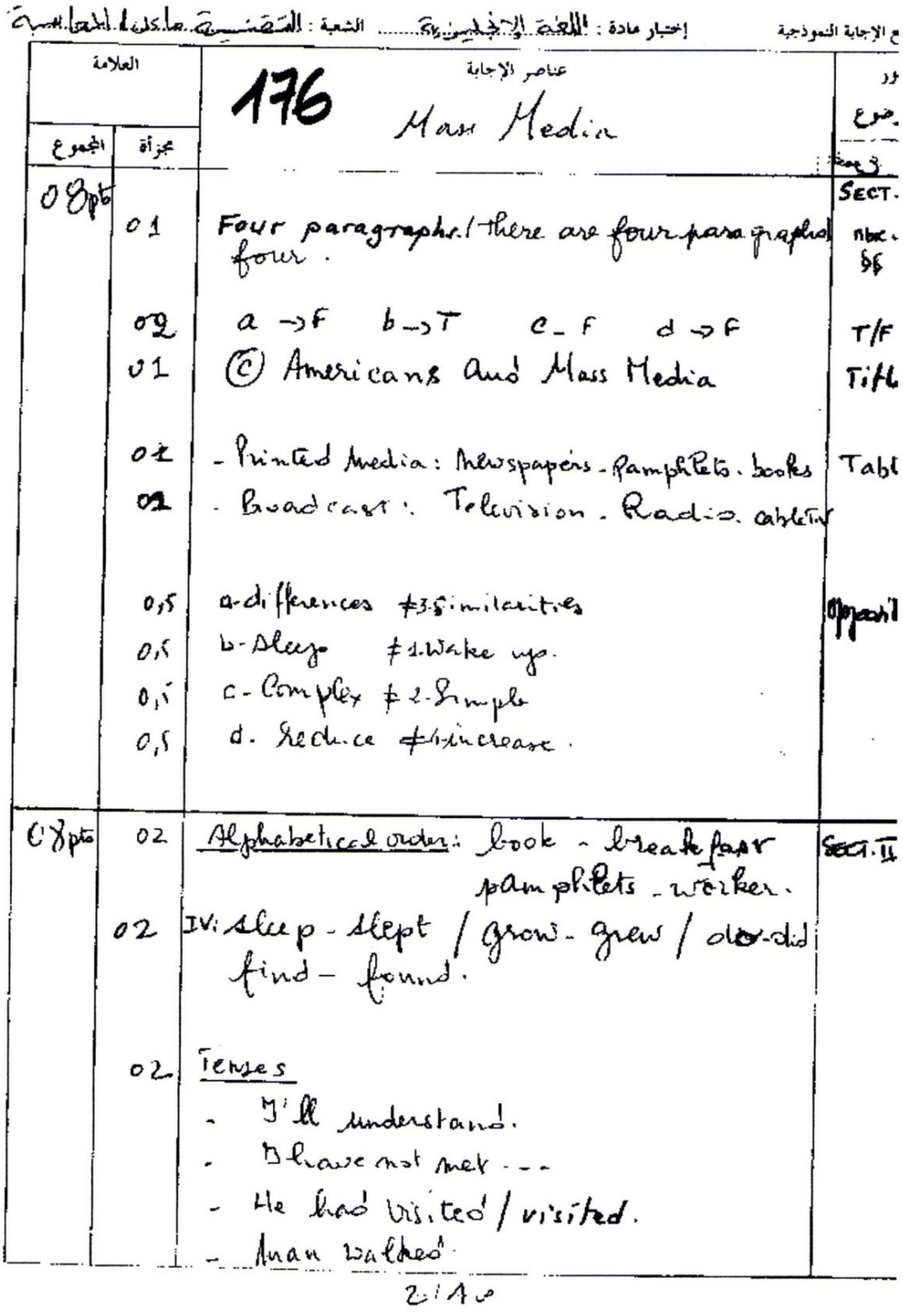
Topic two . *Write a composition of about 60 to 80 words on the following topic:*

In what way has scientific progress improved man's life ?

2 /الصفحة

Appendix E: Model Correction of Technical Streams' BAC English Tests

June, 2001, p. 1

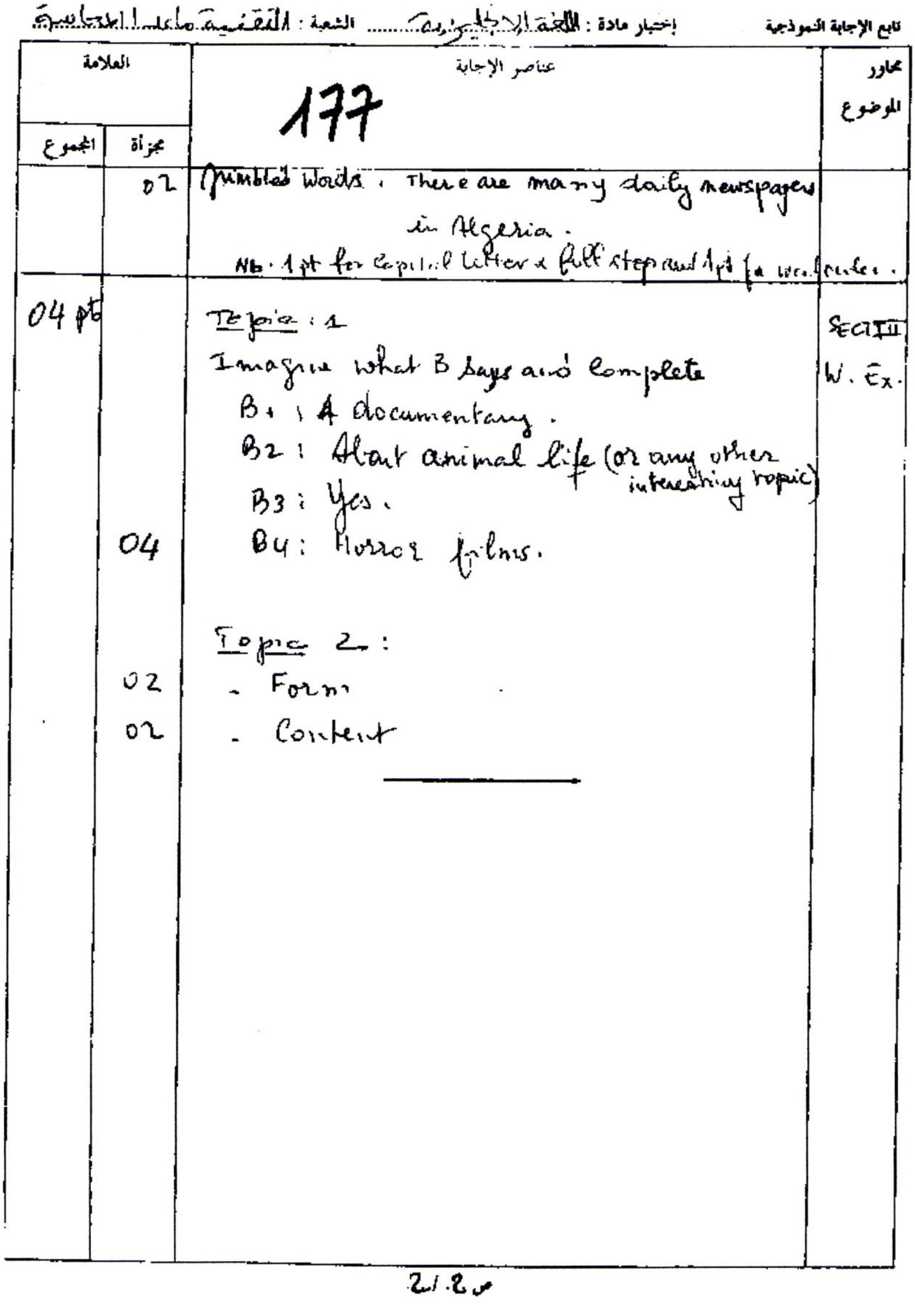


ع الإجابة النموذجية   إختبار مادة : اللغة الإنجليزية......   الشعبة : التقنيــين. هاكل. المحاسبة

العلامة		عناصر الإجابة	ور
المجموع	مجزأة	176 Mass Media	موضوع وضعية
08pts	01	Four paragraphs / there are four paragraphs four.	SECT. nbr. §§
	02	a → F b → T c – F d → F	T/F
	01	© Americans And Mass Media	Title
	01	- Printed Media: Newspapers - Pamphlets. books	Table
	01	- Broadcast: Television. Radio. cable TV	
	0,5	a- differences ≠ 3. similarities	Opposit
	0,5	b- Sleep ≠ 1. Wake up.	
	0,5	c- Complex ≠ 2. Simple	
	0,5	d. Reduce ≠ 4. increase.	
08pts	02	Alphabetical order: book - breakfast pamphlets - worker.	SECT. II
	02	IV: sleep - slept / grow - grew / ~~do~~ - did find - found.	
	02	Tenses - I'll understand. - I have not met --- - He had visited / visited. - Anan walked	

ص 1/2

June, 2001, p. 2

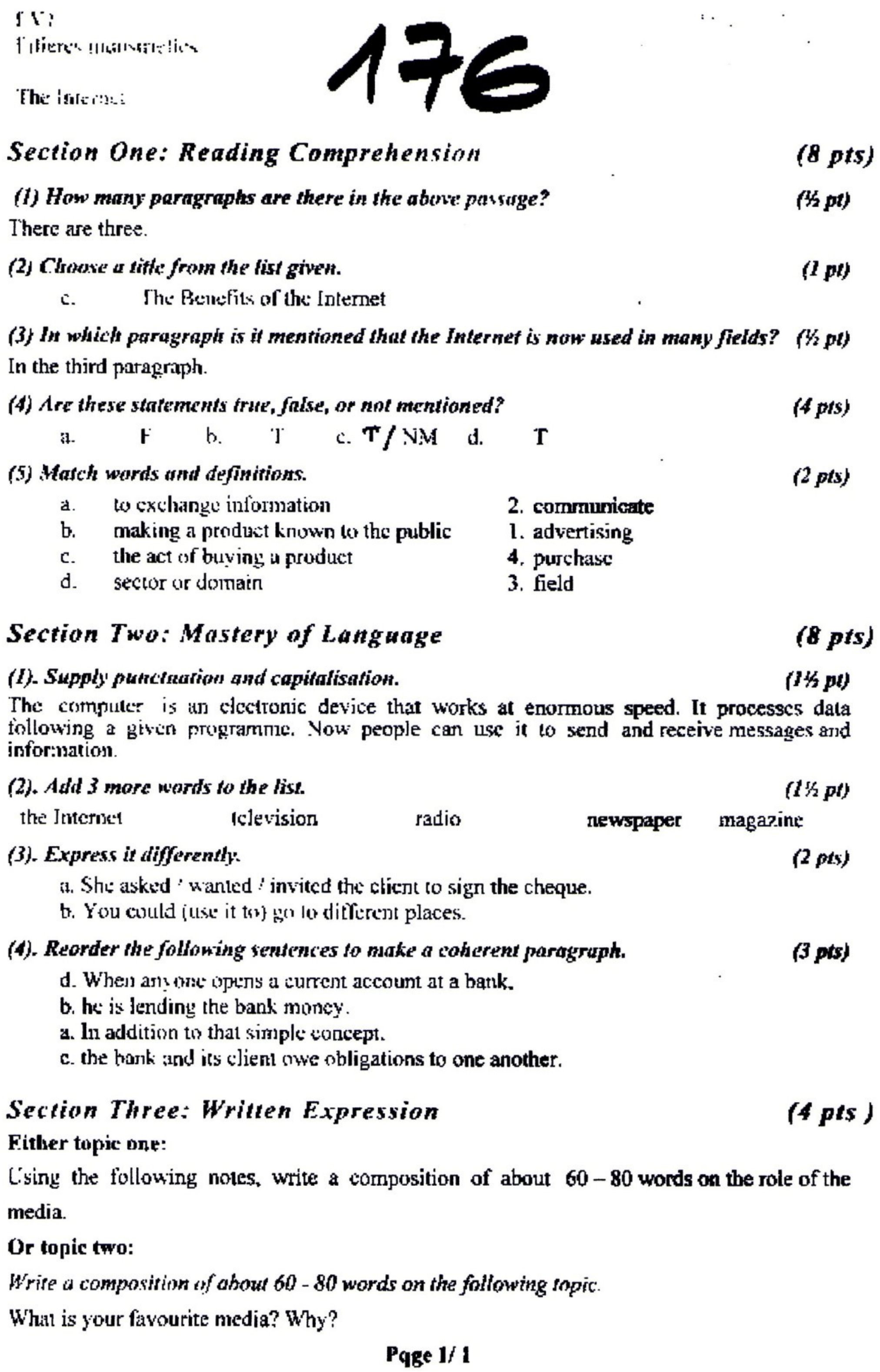


تابع الإجابة النموذجية   إختبار مادة : اللغة الإنجليزية .... الشعبة : التقنية ماعدا المحاسبة

المجموع	مجزأة	عناصر الإجابة 177	محاور الموضوع
	02	Jumbled Words : There are many daily newspapers in Algeria. Nb : 1pt for capital letter & full stop and 1pt for word order.	
04 pts	 04 02 02	Topic : 1 Imagine what B says and complete B1 : A documentary. B2 : About animal life (or any other interesting topic) B3 : Yes. B4 : Horror films. Topic 2 : - Form - Content	SECT II W. Ex.

ص 2/2

June, 2002



[illegible]
Filieres industrielles

The Internet

176

Section One: Reading Comprehension (8 pts)

(1) How many paragraphs are there in the above passage? ***(½ pt)***

There are three.

(2) Choose a title from the list given. ***(1 pt)***

c. The Benefits of the Internet

(3) In which paragraph is it mentioned that the Internet is now used in many fields? ***(½ pt)***

In the third paragraph.

(4) Are these statements true, false, or not mentioned? ***(4 pts)***

a. F b. T c. T / NM d. T

(5) Match words and definitions. ***(2 pts)***

a.	to exchange information	2. communicate
b.	making a product known to the public	1. advertising
c.	the act of buying a product	4. purchase
d.	sector or domain	3. field

Section Two: Mastery of Language (8 pts)

(1). Supply punctuation and capitalisation. ***(1½ pt)***

The computer is an electronic device that works at enormous speed. It processes data following a given programme. Now people can use it to send and receive messages and information.

(2). Add 3 more words to the list. ***(1½ pt)***

the Internet television radio newspaper magazine

(3). Express it differently. ***(2 pts)***

a. She asked / wanted / invited the client to sign the cheque.
b. You could (use it to) go to different places.

(4). Reorder the following sentences to make a coherent paragraph. ***(3 pts)***

d. When anyone opens a current account at a bank,
b. he is lending the bank money.
a. In addition to that simple concept.
c. the bank and its client owe obligations to one another.

Section Three: Written Expression (4 pts)

Either topic one:

Using the following notes, write a composition of about 60 – 80 words on the role of the media.

Or topic two:

Write a composition of about 60 - 80 words on the following topic.

What is your favourite media? Why?

Pqge 1/ 1

تكتب الإجابة النموذجية على هذه الورقة ولا تقبل سواها

الإجابة النموذجية لموضوع مقترح لبكالوريا دورة : 2007......

إختبار مادة : الإنجليزية .... الشعبة : التقنية ماعدا المحاسبة المدة : 2 سا........

<table>
<tr><th colspan="2">العلامة</th><th rowspan="2">عناصر الإجابة
173 – MASS PRODUCTION –</th><th rowspan="2">محاور الموضوع</th></tr>
<tr><th>المجموع</th><th>مجزأة</th></tr>
<tr><td>08 pts.</td><td>01</td><td>1/ two paragraphs / there are two paragraphs / two</td><td>SECT. I</td></tr>
<tr><td></td><td>02</td><td>2) a → F ; b → F ; c → T ; d → F</td><td></td></tr>
<tr><td></td><td>01</td><td>3) Industrial and manufactured production.</td><td></td></tr>
<tr><td></td><td>02/</td><td>4)
<table>
<tr><th>Handicrafts</th><th>Industry</th></tr>
<tr><td>pottery</td><td>mass production</td></tr>
<tr><td>Persian carpets</td><td>identical articles</td></tr>
<tr><td>Hand-made</td><td>low prices</td></tr>
<tr><td>Hand-made lace</td><td>cheap power</td></tr>
</table></td><td></td></tr>
<tr><td></td><td>02/</td><td>5)
a – manufactured → 2 made
b – production → 4 output
d – machines → 3 engines
c – satisfied → 1 happy.</td><td></td></tr>
<tr><td>08 pts</td><td>1 pt</td><td>1. He has too much work and too little time to get out with his friends</td><td>SECT. II</td></tr>
<tr><td></td><td>1 pt</td><td>2) Chinese – European – Persian – Population</td><td></td></tr>
<tr><td></td><td>01 pt
(0,25x4)</td><td>3) meant – took – rode – spoke.</td><td></td></tr>
<tr><td></td><td>02
(0,5x4)</td><td>4) have been / plays – has played / are carried / is saved.</td><td></td></tr>
<tr><td></td><td>02
(0,5x4)</td><td>5) Governments fight illiteracy through information.</td><td></td></tr>
<tr><td></td><td></td><td>B1 : Yes / Yes, it is / No / No, it isn't.
B2 : It is expensive / It is cheap.
B3 : In Ghardaïa / Any other place.
B4 : I'll keep it / I'll offer it to ..</td><td></td></tr>
<tr><td></td><td>1 pt</td><td>6/ 1 syllable : goods – hand 2 syllables : craftsman – carpet</td><td></td></tr>
<tr><td>4 pts</td><td></td><td>Topic one : form 2,5 content 1,5
Topic two : form 2 content 2</td><td>Sect III</td></tr>
</table>

ص : 01 / 01

تكتب الإجابة النموذجية على هذه الورقة ولا تقبل سواها

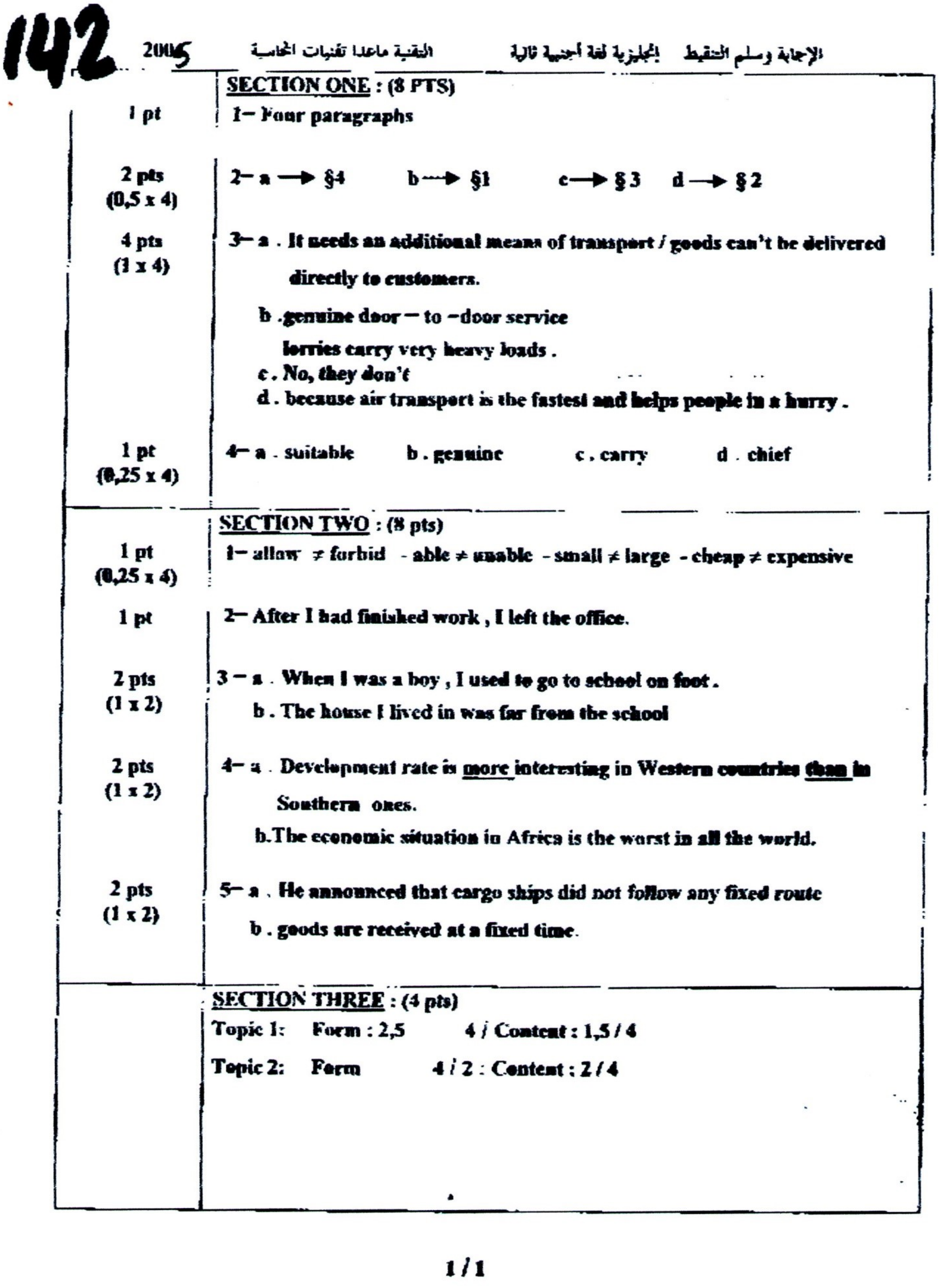
الإجابة النموذجية لموضوع مقترح لبكالوريا دورة : ....2004.....

إختبار مادة : الإنجليزية LV2...... الشعبة : التقنية ماعدا المحاسبة المدة : ..02 (ساعتان)

الإجابة النموذجية 162 وسلم التنقيط

عدد الصفحات : 01

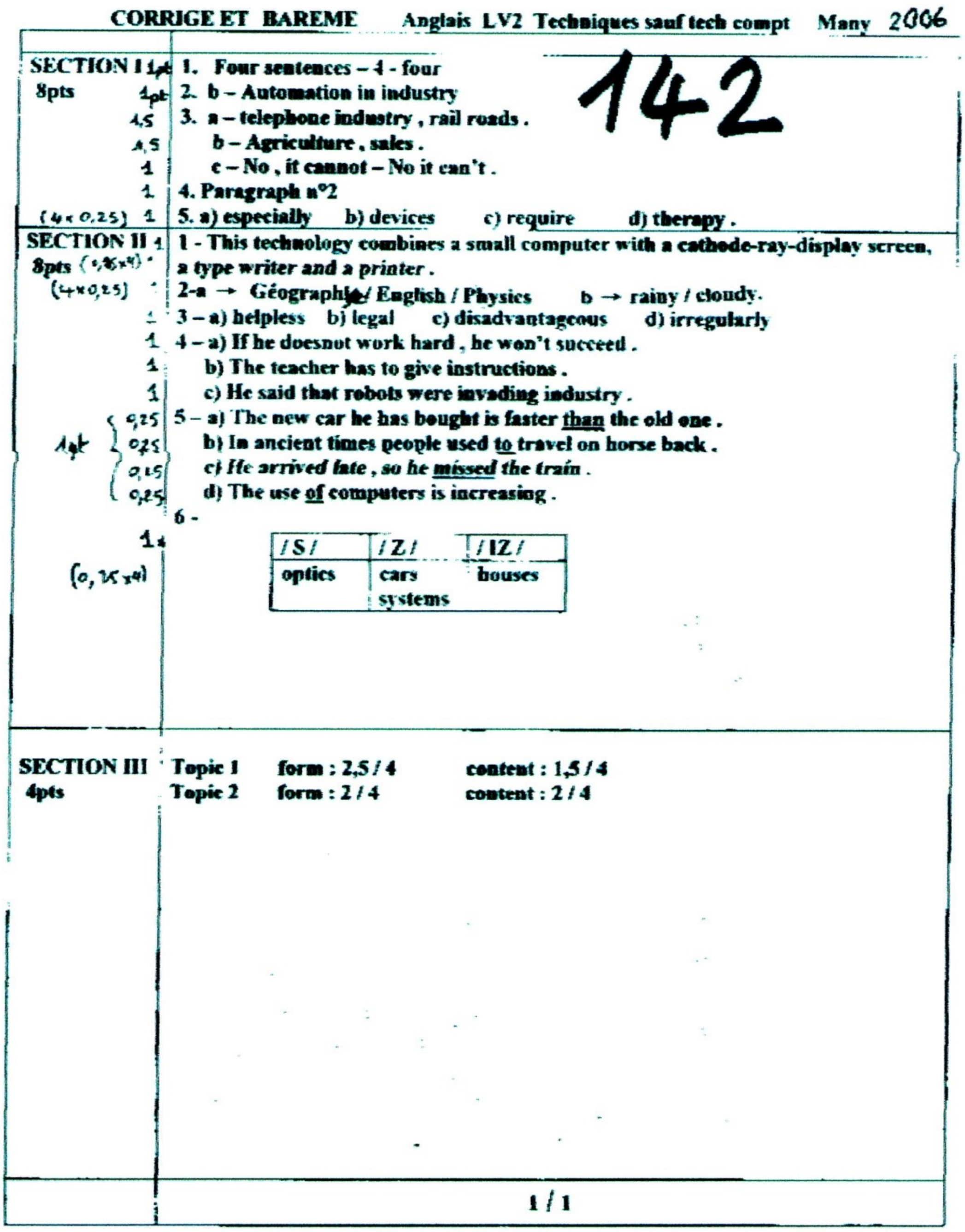
العلامة		عناصر الإجابة	محاور الموضوع
المجموع	مجزأة		
08pts.	0,5	1. There are three (03) paragraphs	Section I
	02	2. a) F - b) T - c) T - d) F	
	02	3. a) manufacturing, service industries, fire fighting b) advantages: reduce costs; improve output in quality & quantity, frees workers from repetitive tasks... drawbacks: causes serious problems for workers (unemployment).	
	01,50	4. it → automation that → robot device its → automation	
	02	5. a → 4 ; b → 1 ; c → 2 ; d → 3.	
08pts.	01	1. Unlike many countries, America does not have an official apprenticeship programme.	SECTION II
	01	2. useful - dangerous - automatic/automated - technological	
	01	3. b1: The writer said that it took/takes him a long time to dial the number/ dialing the number took/takes him...	
	01	b2: The writer asked what people could/can use the internet for.	
	01	b3: The need for specialists in electronics has been increased by automation.	
	03	4. c - a - d - b.	
04pts		Topic One: Form: 2,5/4 Content: 1,5/4 Topic Two: Form: 2/2 Content: 2/2	SECTION III



142 2005 الشعبة ماعدا تقنيات المحاسبة الإنجليزية لغة أجنبية ثالثة الإجابة وسلم التنقيط

	SECTION ONE : (8 PTS)
1 pt	1– Four paragraphs
2 pts (0,5 x 4)	2– a → §4 b → §1 c → §3 d → §2
4 pts (1 x 4)	3– a . It needs an additional means of transport / goods can't be delivered directly to customers. b .genuine door – to –door service lorries carry very heavy loads . c . No, they don't d . because air transport is the fastest and helps people in a hurry .
1 pt (0,25 x 4)	4– a . suitable b . genuine c . carry d . chief
	SECTION TWO : (8 pts)
1 pt (0,25 x 4)	1– allow ≠ forbid - able ≠ unable - small ≠ large - cheap ≠ expensive
1 pt	2– After I had finished work , I left the office.
2 pts (1 x 2)	3 – a . When I was a boy , I used to go to school on foot . b . The house I lived in was far from the school
2 pts (1 x 2)	4– a . Development rate is more interesting in Western countries than in Southern ones. b.The economic situation in Africa is the worst in all the world.
2 pts (1 x 2)	5– a . He announced that cargo ships did not follow any fixed route b . goods are received at a fixed time.
	SECTION THREE : (4 pts) Topic 1: Form : 2,5 4 / Content : 1,5 / 4 Topic 2: Form 4 / 2 : Content : 2 / 4

1 / 1



CORRIGE ET BAREME **Anglais LV2 Techniques sauf tech compt** **Many** 2006

142

SECTION I 1pt **1. Four sentences – 4 - four**

8pts 1pt **2. b – Automation in industry**

1,5 **3. a – telephone industry , rail roads .**

1,5 ***b* – Agriculture , sales .**

1 **c – No , it cannot – No it can't .**

1 **4. Paragraph n°2**

(4×0,25) 1 **5. a) especially** b) devices c) require **d) therapy .**

SECTION II 1 **1 - This technology combines a small computer with a cathode-ray-display screen, *a type writer and a printer*.**

8pts (0,25×4) 1 (4×0,25) 1 **2-a → Géographie / English / Physics b → rainy / cloudy.**

1 **3 – a) helpless b) legal c) disadvantageous d) irregularly**

1 **4 – a) If he doesnot work hard , he won't succeed .**

1 **b) The teacher has to give instructions .**

1 **c) He said that robots were invading industry .**

1pt { 0,25 **5 – a) The new car he has bought is faster <u>than</u> the old one .**

0,25 **b) In ancient times people used <u>to</u> travel on horse back .**

0,15 ***c) He arrived late , so he <u>missed</u> the train .***

0,25 **d) The use <u>of</u> computers is increasing .**

1 (0,25×4) **6 -**

/S/	/Z/	/IZ/
optics	cars systems	houses

SECTION III **Topic 1** **form : 2,5 / 4** **content : 1,5 / 4**

4pts **Topic 2** **form : 2 / 4** **content : 2 / 4**

1 / 1

Appendix F: Raters' Questionnaire

Section 1. Qualities of raters

Item 1. In your point of view, on what criteria do the educational authorities appoint teachers for the rating process?

a- Their experience in teaching □

b- Their experience in teaching third year level □

c- Their expertise in rating □

d- There are no requirements for the appointment of raters □

Item 2. Suppose that you are responsible for the selection of raters, on what criteria do you base your choice?

a- Experience in teaching □

b- Experience in teaching the third year level □

c- Expertise in rating □

d- Other factors □

Item 3: Do you think that raters' educational or cultural background can affect their scoring behavior?

a- Yes, I think so □

b- No, I do not think so □

Item 4: Do you think that rates' judgment in general can bear elements of subjectivity?

a-I agree □ b-Do not agree □

Item 5: According to you, do experienced and novice raters employ the same scoring strategies?

a- Yes they do □ b- No, they do not □

- If no, novice raters are, according to you, significantly more lenient in their judgment than expert raters?

a- More lenient □ b- Not more lenient □

Section Two .The Rating Process

Item 6: Operational scoring starts…………

As soon as raters meet □ On the second session of the first day □ On the second day □

- If operational scoring is delayed to the second session or to the second day, what is the first session devoted to?
- Explanation and analysis of the scoring guide □
- Refining the scoring guide □
- Drafting a new scoring guide □

Item 7: Discussio in the first session aims at……

- obtaining a satisfactory level of agreement □
- agreeing on the same scoring techniques □
- other purposes □

Item 8: In your point of view, the scoring guide is indispensible to….

novice raters □ expert raters □

Item 9: In the pre-scoring session, sample scripts are…………

- blindly single-rated by the chief examiner □
- blindly double-scored by pairs of raters □
- scored collectively by all the participants □

Item 10: In the pre-scoring session, the sample papers represent the ……….

problematic scripts □ consensus scripts □ randomly-chosen scripts □

Item 11: Once live scoring is under way; do you discuss with table leaders or the chief examiner the difficulties that might encounter you during the correction of test takers' papers?

Certainly □ Not necessarily □

Section Three . Rater Training

Item 12: Have you attended a seminar, a colloquium, or a meeting on rating?

Yes, I have □ No, I have not □

Item 13: Do you think that introducing raters to the assessment without any type of training can affect the consistency of their scoring?

agree □ do not agree □

- If so, can training sessions determine whether a rater can participate satisfactorily in the rating process?

Agree □ Do not agree □

Section Four: Rater Reliability

Item 14: According to you, rater consistency can be understood as …………

intra rater reliability □ inter rater reliability □th types of reliability □

Item 15: According to you, variability between raters could be understood in terms of………..

severity □ leniency □

Item 16: Can judges' severity or leniency be modified by training?

Sure □ Maybe □ Do not think so □

Item 17: Can the consistency of your scoring be affected by the succession of the number papers that you are supposed to correct each day?

Yes □ Yes, to some extent □ No, not at all □

Section Five: Methods for Solving Rater Discrepancies

Item 18: In the BAC exam, scripts are blindly…

single-rated □ double-rated □

Item 19: How much tolerance for discrepancies between raters is allowed in the BAC exam?

One mark □ Two marks □ Three marks □ Four marks □

Item 20: In the case of adjacent agreement, how will the final score be computed?

I consider the high mark □ The low and the high marks are averaged □

Item 21: What happens in the case of disagreement between the first and the second raters?

a- The two raters discuss the issue and assign a consensus score □
b- A third rater is brought in to resolve the discrepancy □
c- Other solutions □

- If a third rater is brought in, how to compute the final score

a- Considering the expert's score □
b- Averaging the three scores □
c- Averaging the two closest scores □

Item 22: Does the chief examiner communicate to discrepant raters the amount of variability which they have done?

Yes □ No □

Section Six . Rating Scales

Item 23: Does the scoring guide include a rating scale?

Yes □ No □

Item 24: In the lack of rating scales, how do you score the writing tasks?

a- Depend on my own judgment □
b- Rate the script on several aspects □
c- Read the script and assign a holistic score □
d- Other techniques □

Item 25: If two raters assign the scores, included in the table below, to the same script, will their rating be considered as identical or variable?

a) Identical □ b) variable □

Exam Sections	Script 1	Script 2
Reading	06/08	05/08
Mastery of Language	05/08	02/08
Written Expression	00/04	04/04
Final Score	11/20	11/20

Section Seven: Incorporation Automated Scoring

Item 26: What is your point of view on the incorporation of automated scoring in the BAC English tests?

Promising □ Threatening □

- If promising, which tasks can, in your opinion, be better scored by the computer?

Yes-no questions □

Matching activities □

Phonetics □

Grammar □

Others □

Item 27: Do you think that computerized scoring can soon be operational in the BAC Exam?

Yes, I think so □ I do not think so □

Section Eight: Test tryout

Item 28: Has the Ministry of Education piloted a draft sample of the BAC English test in your school?

Yes □ No □

- If so, how often has that happened?

…………………………………………………………………………………..

Item 29: Do you think that test tryout provide more efficient information on item difficultly and discrimination indices than the information provided by teachers' expertise?

Agree □ Do not agree □ Do not know □

Appendix G: The Interview

Purpose of the Interview : Investigation of inter rater and intra rater reliability in Eloued BAC Exam rating Centre (2015).

Interviewer: MrNaoua Mohamed

Interviewee: The Chief Examiner of English language test rating committee in Eloued BAC Exam Rating Centre (2013). He has been invited to oversee the scoring process of English language tests for at least five BAC sessions. His rating expertise has developed from his experience as a rater, and then from his numerous appointments as a chief examiner.

First, let me express my deep thanks and acknowledgments for your cooperation in the administration of the questionnaires in Eloued BAC Exam Rating Centre, and also for agreeing to this interview. My questions intend to investigate the inter-rater and intra-rater reliability of the scoring process in the committee that you have already overseen. These questions will involve the following points:

The Category of Raters Participating in the Rating Process.

The Rating Process

- The Pre-rating Stage (The Standardization meeting)
- Live Rating
- The Type of Scoring
- Monitoring Raters' Marks
- Agreements and Discrepancies
- Operational Scores
- Method for Resolving Rater Discrepancies
- The Post Scoring Procedures
- The Analysis of Discrepancies
- Rater training
- The Incorporation of Automated Scoring in the BAC rating centers
- The Post Scoring Procedures

Q 1: As far as I know, this is not the first time in which you chair a BAC English test rating committee.

A: Yes

Q 2: How often have you already been appointed in this position?

A: Four times

Q3: When and in which centers have you previously worked?

A: Ghardaia (2007) Eloued (2008 /2009 /2010 /2013)

Q 4: Do you think that rating at Guémar (Eloued) center meet the requirement of scoring conditions?

A: Yes

Q 5: What type of problems that usually encounter raters?

A: No serious problems.

Q 6: Now let us turn to the raters themselves, would you please inform us of the exact number of assessors who participated in the scoring process this year?

A: 63

Q 07: Has this participation been limited to raters from the 'wilaya' of Eloued; or it has extended to raters from other 'wilayas'?

A: Raters were limited to the wilaya of Eloued.

Q 8: Do the heads of rating committees have a given role in the appointment of raters? Or the latter are exclusively appointed by local departments of education.

A: They are exclusively appointed by local departments of education.

Q 9: As a chief examiner and according to your previous experience, on what grounds are raters appointed in the assessment process? In other words, are they chosen because of their expertise in rating or their experience in teaching examination levels?

A: There are no specific requirements in the appointment of raters

Q10: Do you agree on the fact that raters' experience is important for the scoring process?

A: I totally agree

Q 11: Then, in your point of view and for ensuring more reliable scoring what percentage should expert raters form?

A: They should, at least, form two thirds of the whole number of raters.

Q 12: Would you please inform us of the number of committees and specialties that have been rated under your supervision this year?

Specialities	Number of Copies		Specialities	Number of Copies
Lit &phil	4800		Math	119
F.L	558		Math .T	460
EM	1500		Exp.Sci	5120

The Rating Process

Q 13: What do you devote your first meeting with raters to?

A: Distribution, explanation, discussion and refinement of the scoring guide; the first day is wholly devoted to the standardization of the rating procedures.

Q 14: When do you exactly engage in live scoring?

A: On the second day.

Q 15: What method or technique do you use in order to standardize raters' marks?

A: Sample papers are scored by all raters; who then engage in general discussion to reach agreement on a given model of rating.

Q: 16: On what grounds do you choose sample scripts?

A: We pick them out randomly.

Q 17: What is the type of raters who usually engage in this discussion?

A: All types of raters

Q 18: Does the use of discussion as a form of consensus allow the opportunity for one type of raters to dominate the other type?

A: I agree.

Q 19: According to your experience, what type of raters who usually dominate the discussion session?

A: The raters with the highest level of expertise in scoring.

Q 20: Do you think discussion dominance can affect rating consistency?

A: No

Q 21: Supposing that there are some extreme differences amongst raters in the pre-scoring or the standardization meeting, how do you resolve these discrepancies?

A: These discrepancies are settled by discussion method.

Q 22: 0n what criteria are raters divided into teams?

A: According to their level of expertise

Q 23: After the division of raters into teams, on what criteria do you appoint the team or table leaders?

A: According to their level of expertise

Q 24: Once live scoring is underway, do you hold meetings with the team leaders to standardize raters' marks?

A: Yes

Q 25: Do team leaders communicate your directions to raters?

A: Yes

Q 26: Is it useful to have debriefing sessions regularly?

A: No

Q: 27: Now let us turn again to the scoring guide, what procedures of scoring does the guide propose, I mean objective or subjective scoring?

A: It includes the two types.

Q: 28: Does the guide include a rating scale that specifies the scoring of the writing skill?

A: No.

Q 29: According to your supervision of the rating process, do raters assign a single score to the written tasks, or do they give different marks that are finally combined into a composite score?

A: They assign a single score.

Resolving Rater Discrepancies

Q 30: In scoring the BAC English tests, what does agreement mean, does it require the two raters to assign the same score?

A: No. It requires them to assign adjacent scores.

Q 31: What is the extent to which scores can be considered adjacent?

A: When they are 4 or less than 4 points apart.

Q 32: In case there are adjacent agreements, how to report the operational score?

A: By averaging the two scores

Q 33: Since adjacent scores can extend to 04 points apart, does the scoring guide consider the operational scores included in the table below as discrepant or adjacent?

	Rater One	Rater Two
Reading	05	0.5
Section Two	06	02.5
Written Expression	00	04
Final Score	11	07

A: Adjacent and need to be averaged

Q34: In your opinion, does this type of scoring reflect pupils' language ability?

A: yes.

Q35: What is the exact number of discrepant scores this year?

A: More than 160

Q 36: In case of discrepancies, do you invite the original raters to discuss and reach agreement?

A: No. A third rater is brought in.

Q 37: On what criteria are third raters or adjudicators selected?

A: No special criteria. Adjudicators are chosen because they do not live far away from rating centers

Q 38: Once a third rater is brought in, what is the model that you apply to resolve these discrepancies?

A: The third rater's mark is averaged with the closest mark.

Q 39: Is it possible for you to identify the raters who have assigned significant discrepant scores?

A: yes

Q 40: Do you communicate to these raters the number of discrepant scores they have assigned?

A: No

41: Do you agree on the fact that the identification of raters who show significant variations can contribute to reducing rater differences?

A: No

Q 42: Is the record of discrepant raters evaluated by the Educational authorities?

A: No

Q 43: Do the Educational authorities call the raterswho displayed significant variations to training sessions?

A: No

Q 44: Based on your experience, what do you think of introducing raters to the assessment without any training?

A: Live scoring is the only opportunity for practicing in double rating.

Q 45: Now let me ask you about the post scoring procedures; have you been invited to attend a meeting that was devoted to the analysis of raters' discrepancies?

A: No

Q 46: Are you in favor of holding seminars or meetings to study the source of raters' discrepancies?

A: Yes

Q 47: Do you think that the recommendations of such meetings can be used as feedback in rater training?

A: Yes

Q 48: What is your point of view concerning the incorporation automated scoring in the BAC English tests?

A: I consider it threatening

Q 49: Do you think computerized scoring will soon be operational in the BAC Exam?

A: I do not think so…

I am deeply grateful to you for your cooperation. Thank you again for taking the time to discuss so many aspects in scoring English tests in the committee that you have overseen.

RÉSUMÉ

Pour l'optimisation de l'enseignement d'anglais au niveau secondaire, le Ministère de l'Éducation en Algérie a fixé plusieurs objectifs qui s'adapteraient aux besoins de chaque spécialité. Dans les spécialités de la technologie, les programmes ont été conçus pour permettre aux apprenants d'utiliser cette langue pour des objectifs académiques ou professionnels spécifiques, ou pour leur permettre d'avoir accès à la documentation scientifique et technologique et poursuivre en conséquence leurs études ultérieures. Afin de savoir dans quelle mesure ces objectifs ont été atteints, on a eu recours aux tests, aux évaluations et aux statistiques qui indiquent, selon les chiffres publiés par le Centre d'Orientation d' Eloued (2001-2006) et l'Office National des Examens et Concours (ONEC), que les résultats du baccalauréat d'anglais obtenus par les élèves des branches de technologie à Eloued les classent au bas de la liste derrière toutes les autres spécialités de l'enseignement secondaire. Vu que ces apprenants étudient dans les mêmes établissements, utilisent les mêmes manuels et reçoivent des cours dispensés par les mêmes enseignants, la présente étude tente de mettre l'accent sur les examens du BAC Anglais à propos duquel nous avons formulé quelques hypothèses relatives à leur structure, leurs contenus, leurs degrés de représentativité des programmes scolaires et leurs corrections. Ces hypothèses ont été vérifiées à travers les données que nous avons recueillies par des outils méthodiques de l'approche descriptive : le questionnaire, l'entrevue et les sources documentaires. Le questionnaire a été distribué à une population de 63 correcteurs des examens du BAC au Centre de Correction à la wilaya d'Eloued. De même qu'une interview a été réalisée avec le chef du même comité. Quant aux données des sources documentaires, elles consistent des copies des examens du BAC Anglais, les résultats obtenus par les apprenants à ces examens, ainsi que le programme scolaire d'enseignement des cours des branches de technologie. L'analyse et le traitement de ces données s'inscrivent dans le cadre du modèle argumentatif de Toulmin (1958, 2003) dont la conclusion vient d'infirmer les explications négatives déjà données aux résultats des apprenants, les décisions basées sur ces interprétations ainsi que les hypothèses allant dans le sens que les apprenants ne sont pas capables de maitriser cette langue. L'objectif principal de cette étude est d'identifier les facteurs responsables de la sous-performance des apprenants de la filière de technologie en anglais ; et ce, à travers l'analyse et l'évaluation des examens : quant à leurs structures et le degré de leur conformité avec les programmes. Ce qui permet de proposer, à la lumière des résultats de cette analyse, un ensemble de recommandations destinées à améliorer le processus de l'évaluation et des examens en anglais au Baccalauréat.

Mots Clés: Construction - Évaluation - Examens - Technologie - validité

الملخص

سطرت وزارة التربية الوطنية في الجزائر عدة أهداف لتعليم اللغة الانجليزية في التعليم الثانوي تتلاءم مع متطلبات كل شعبة أو تخصص. ففي شعب التكنولوجيا مثلا، سطّرت أهداف كي تمكّن التلاميذ من استعمال هذه اللغة لأغراض أكاديمية أو وظيفية محددة أو للاطلاع على محتويات المجلات والوثائق الأكاديمية المتخصصة في ميدان العلوم والتكنولوجيا. ولكي يتم التعرّف على مدى تحقيق هذه الأهداف على أرض الواقع، يتطلب منّا الأمر اللجوء إلى الفحص والتقييم. ففيما يخصّ شعب التكنولوجيا بولاية الوادي، تشير الاحصائيات التي ينشرها مركز التوجيه المدرسي وكذا الديوان الوطني للامتحانات والمسابقات للسنوات 2001-2006 أن نتائج اختبارات اللغة الانجليزية الخاصة بامتحان البكالوريا ترتب هؤلاء التلاميذ في مؤخرة القائمة خلف جميع شعب التعليم الثانوي الأخرى. ونظرا لأن هؤلاء التلاميذ يدرسون في المؤسسات نفسها ويستعملون الكتب نفسها ويدّرسون تقريبا من طرف الأساتذة أنفسهم، ارتأت هذه الدراسة أن تركز على اختبارات اللغة الإنجليزية في امتحان البكالوريا نفسها وذلك بإثارة عدّة فرضيات تخص هذه الاختبارات من حيث بنيتها، محتواها، مدى تمثيلها لجميع محاور البرنامج الدراسي أو عمليه تصحيحها. ولفحص مدى صحّة هذه الفرضيات تبنت هذه الدراسة المنهج الوصفي بأدواته المعروفة كالاستبيان، المقابلة والبيانات الوثائقية، حيث تم توزيع الاستبيان على مجتمع الدراسة المتكون من 63 مصححا لاختبارات اللغة الإنجليزية بمركز التصحيح لولاية الوادي كما تم إجراء المقابلة مع رئيس لجنة التصحيح بالمركز نفسه. وتشمل البيانات الوثائقية نسخا من اختبارات اللغة الانجليزية، النقاط التي تحصل عليها التلاميذ في هذه الاختبارات، وكذا البرنامج الدراسي لشعب التكنولوجيا. وأثبتت نتائج تحليل البيانات عدم صحة التفسيرات التي أعطيت لنتائج التلاميذ والتي تفيد بأنهم غير قادرين على استعمال هذه اللغة وكذا عدم صحّة القرارات الناتجة عن تلك التفسيرات والآثار السلبية الناجمة عنها. ولمعرفة الأسباب الحقيقية الكامنة وراء تدنّي نتائج هؤلاء التلاميذ في هذه المادة، ارتأينا تحليل وتقييم هذه الاختبارات من حيث بنيتها ومدى توافقها مع البرنامج الدراسي من حيث المحتوى والتوزيع المتساوي للوحدات، ومن ثمّ اقتراح حلول تمكن من تحسين عملية التقييم الخاصة باللغة الانجليزية في امتحان البكالوريا.

الكلمات المفتاحية: البنية – الاختبارات – التقييم – التكنولوجيا – المصداقية