

A Hands-on Approach to the Theory of TEFL Measurements

Received date: 05/11/2022 Accepted date:31/05/2023 Published date:20/06/2023

Ahmed Bechar¹ Yasser Ben Moussa²

¹Mohamed Khider University of Biskra, Algeria, **Email:** a.bechar@univ-biskra.dz

²Kasdi Merbah University of Ouargla, Algeria, **Email:** benmoussa.yasser@univ-ouargla.dz

Abstract:

Are EFL/ SLL teachers testing their students right? This legitimate question may not be popular with the *Establishment* as it questions the whole teaching-learning process. If the inflation of marks, that appears to characterize students' mark bulletins, cannot be relied upon to evaluate students' actual uptake and output, then what conclusions can we draw? Most EFL teachers rush into planning and constructing tests without having firm theoretical foundations, which eventually occasions the dreaded backwash/ washback effect. This paper intends to provide insights about in-class TEFL measurements with a view of empowering stakeholders to bridge the gap between curriculum objectives and test format and contents.

Key Words: TEFL measurements; tests; theoretical foundations; backwash/ washback

Corresponding Author: Ahmed Bechar, **Email:** a.bechar@univ-biskra.dz

1. Introduction:

In school, children who are viewed as stupid often suffer from nothing more than style mismatches that of their teacher. I know : When I took my introductory psychology course as a freshman in college, I was already eager to major in psychology. As a child, I had done **extremely poorly** on **intelligence tests** because of crippling **test anxiety**. The result was an **interest in** understanding **intelligence** that has not subsided to this day. Thus I took the introductory psychology course with every intention of using it as a launching point for college major in psychology and later career in the same field.

Those hopes came to an abrupt end when I got my grade for the course, a C. I first realized things were going badly when the professor handed us back our very first test result. It was an essay test, scored on scale from 1 to 10. For some perverse reason, the professor had decided to hand back the test booklet in descending order of score. By the time I received mine, the only people left in the room were either a whole lot taller than I was, a whole lot heavier, or both. They at least, were in college for a reason- in large part to play sports. It was no longer so clear why I was there. Indeed, I inferred that my own test booklet must have gotten out of order. But no, I received a 3 out of 10.

It turned out that what the professor expected from our essays and what I had thought he expected had **no overlap**. I thought that the

use of an essay test meant the professor wanted us to be **creative** and **go beyond the information given**. In fact, the teacher had 10 points he *wanted us to make we actually had made*. **No pretenses about creativity, no attempt even to encourage critical thinking ... Indeed, I even remember, perhaps falsely, a comment by the professor to the effect that there was one Sternberg in psychology already, and that it appeared that there would not soon be two. I got the message.** Robert Jeffrey Sternberg (1999). Thinking Styles. pp12-13

Pass-fail in common parlance refers to how successful or otherwise learners have been able to score on term tests. The claim may be put that assessment, examination, evaluation, and tests are, therefore, an inherent and integral part and parcel of the teaching-learning process. According to Atkin et al. (2001, p.7) “ Assessment is universally present in schools and a natural part of both teaching and learning”. In fact, assessment, evaluation, testing and schooling are symbiotic; they are occasioned by the circumstances. In the same vein, Soled (1995, p.1), notes “assessment, testing, and evaluation are everyday occurrences in ... education”. Hence, educational measurements of students’ achievement and progress (or otherwise) are inevitable. Future prospects significantly depend on the outcomes of these indicators. As it could be concluded from Sternberg's introspection, potential Shaws, Einsteins, Marie Curies, Russels, and Zuweils have been doomed to oblivion because of inconsistencies in educational measurements.

Amazingly, these TEFL measurement tools will remain pervasive and omnipresent as gold standards in schools as long as these are intended to hold out. These tools are, however, a double-edged sword. Oerman & Baberson (2013, p.315) report “Some critics of tests claim that tests results in emotional and psychological harm to students. The concern is that tests threaten students and make them anxious, fearful, and discouraged, resulting in harm to their self-esteem”. The social, intellectual and psychological impacts of these tools will, then, indelibly mark learners forever. A thorough theoretical analysis and empirical research need, thus, be conducted diachronically as well as synchronically to strengthen their upside and lessen their downside. This paper concerns itself with the identification of these constructs in a comprehensive theoretical framework whose premise it to make potential EFL audience (teachers, students, and novice researchers, etc.) (1) familiar and (2) aware of the complex intricacies of this enduring issue (i.e., educational measurements).

2. Quick Overview of Assessment, Evaluation, and Testing

2.1 Assessment : Assessment derives from Latin *ad* (prefix) and *census* (to add or count) or *assidere* (sit beside). Assessment is the process of collecting information or data about learners' aptitude, behavior, progress, and performance, etc. Furthermore, assessment means in an educational context, the process of observing learning (namely, describing, collecting, recording, scoring, and interpreting

information about a learner's or one's own learning). At its most useful, assessment is an episode in the learning process; part of reflection and autobiographical understanding of progress. Traditionally, learners' assessments are utilized to determine a placement, promotion, graduation, or retention

2.2 Evaluation : Evaluation is another term that is directly attributed to Latin: *e* (prefix for out), *valu* (give value or credit) and *ation* (suffixes to make abstract nouns). Literally, evaluation indicates to come up with a judgment. Evaluation is subjective in nature unlike assessment, which basically numerical. According to Lynch (1997, p. 2) “Evaluation is a systematic attempt to gather information in order to make judgments or decisions”. Hudson (2015, np) further details the definition “[e]valuation is an **ongoing reflective judgment** on the **quality of teaching practices** and the **learning environment**” (emphasis added).

2.3 Testing : The Latin word *testum* refers interestingly to earthen pot. A test is a systematic procedure for obtaining a *sample* of learners' behavior. Longman Dictionary of Contemporary English defines testing and tests as follows “a set of questions, exercises, or practical activities to measure someone's skill, ability, or knowledge”.

3. Etymology of Assessment, Evaluation, and Testing

An all-embracing common feature among these three concepts is **systematic** in that the process of collecting data and

information on learners' achievement, aptitude, performance, and progress so that a judgment and/ or decision is made about their performance, feedback and future achievements. Whilst **assessment** is literally sitting beside someone to add up counts; it was thus used in English in the 16th century (1540) in the sense of *attributing a value to a property for tax purposes*, **evaluation** is usually understood to be *issuing judgments about students performance and/ or knowledge*. As for **test** is literally an *earthen pot in which metals were melted with the view of verifying their quality*. It was used in English in the late 16th century (1590) in the sense of a trial or examination to determine the correctness of something.

These three concepts make up three “bunked” layers where evaluation (the outer layer encloses assessment (middle layer) and test (the innermost layer) as shown below (fig.1).

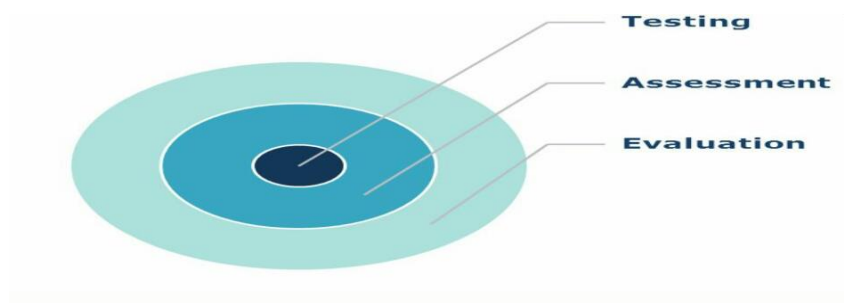


Fig.1. Measurement layers

Tests are the most practical, assessment is about all information provided by tests and other tools, and evaluation is the informed decisions that are based on test outcomes and assessment of learning conditions.

4. Definitions of Assessment and its Types

TEFL assessment seeks to determine *how well* learners are reacting to the learning of English, and it is an integral and inherent part of the quest for quality education. It provides *feedback* to students, educators, parents, policymakers, and the public about the *effectiveness of educational services* (Pelligrino, Chudowsky, and Glaser 2001, p.1). Scholars identify three (03) types of assessment (see **fig.2**) :

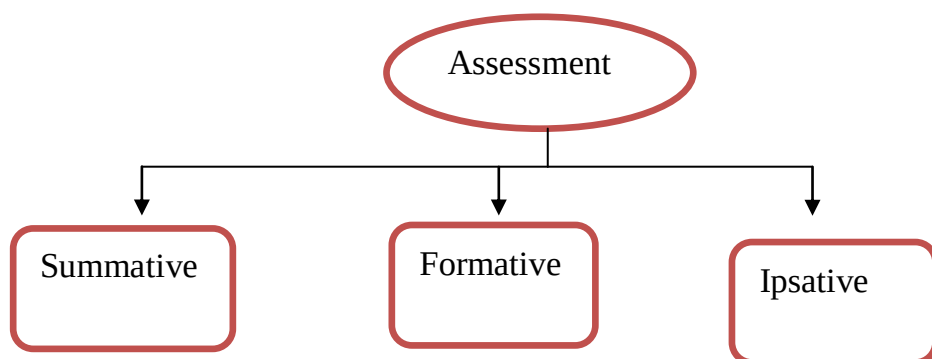


Fig.2 Types of assessment

Each of these types is explained in what follows.

4.1 Summative Assessment (SA): The term summative derive from Latin (*summa : sum/ total*). Gipps (1994, cited in Sawyer, 2004, p.106) defines SA as an assessment which takes place at the end of a term or course and is used to provide information about how much students have learned and how well a course has worked. According to Avis et al. (2009, p.170-1), SA normally leads to an award of qualifications : grades, diplomas and certificates. Therefore, SA is basically an assessment of learning

4.2 Formative Assessment (FA): Formative is yet another term deriving Latin (*formare : to shape/ to mold/ to make*) : Quoting

Gipps, Sawyer also claims that FA takes place during the course of teaching and is used essentially to feedback into teaching and learning (Sawyer, 2004, p.106). Avis et al. (2009, p.170-1) note that FA is an assessment for learning- it takes place during a course or program of study, as an integral part of the learning process. Consequently, FA is an assessment for learning.

4.3 Ipsative Assessment (IA): *Oneself* in Latin is *Ipse*. Again following in Gipps’ footsteps, Sawyer (2004, p.106) defines IA as an assessment in which the student evaluates his/ her performance against his/ her previous performance. Ideally, it continues throughout the professional career.

In the following table, Sawyer (2004, p.107) lays out the types of assessment all while distinguishing their features and giving examples.

Table 1. Types of Assessments, Features and Examples.

Types of Assessment	Their Features + Examples
Summative Assessment	1. Jury/instructor evaluation at the end of the program or course ; 2. Determines how well student has learned and whether teaching is effective ; 3. Degree and course examinations, thesis/ portfolio.
	1. Instructor evaluation during course of teaching ; 2. Feedback into teaching and learning process ;

Formative Assessment	3. Grading on assignment ; 4. Feedback on coursework 5. Feedback on ipsative assessment (self-assessment statement, journal, field notes, or log).
Ipsative Assessment	1. Self-evaluation by a reflective practitioner ; 2. Evaluation of current performance against previous performance and performance of other participants ; 3. Ongoing reflection on learning ; 4. Integrates instructor and peer feedback ; 5. Formalized in self-assessment statements, journal, field notes, or log ; 6. Ideally, continues throughout the professional career.

IA can both be evaluated by the student himself/ herself as well as by instructor and peers according to Sawyer. Jarvis (2002, p.163) notes, however, “IA must be criterion-referenced but need not be assessed by any outsiders. The benchmark of achievement is oneself”. Intuitively, since IA is about assessing oneself, it is needless to be assessed by other than oneself.

5. Definitions of Evaluation and Its Types

Educational evaluation is about decisions. Lynch (1997, p. 2) identifies **evaluation** as “A systematic attempt to gather information in order to make judgments or decisions. Evaluative information can both be gathered through different methods such as observation of administration of pencil- and- paper tests”. Davidson (2005, p. xi) states that “only evaluation can tell us ‘**So what** ?’”. In other words, evaluation informs scholars and stakeholders in what way this information is important and worthy in taking and making informed decisions about students’ performance, program, and course, etc. Scriven (1991), an authority in evaluation, points out that evaluation needs to have **merit** (*intrinsic value of something ; or relative excellence of program, person, and object*) and **worth** (*the value of something to an individual, organization, institution ; or the extent to which a person, program, institution, object that fulfills a need*). Scholars have identified three (03) types of evaluation :

5.1 Summative Evaluation (SE): Morrison et al. claim that summative evaluation (SE) attempts to measure how well the major outcomes of a course or program are attained at the conclusion of instruction (Posttest) or thereafter on the job (quoted in Anfara et al., 2005, p.126). Moreover, it answers the question “How did we do ? ” Its focus is the final result (e.g., to determine the effects of Reading Program A on increasing student achievement.

5.2 Formative Evaluation (FE): FE attempts to measure the intermediate result answering the question “How are we doing ?”.

Therefore, the purpose of FE is to measure the improvement of a course or program (e.g., To determine that each major component of Reading Program A is working as designed).

5.3 Diagnostic Evaluation (DE): The identification through actual examination like what physicians do to find out about a disease is what comes to be termed diagnosis. Kumari & Srivastava (2005, p. 2) define diagnostic evaluation (DE) as “being used when the teacher decides that information is needed about student knowledge or skills prior to deciding on the most effective instruction”. In wide brief, DE then informs instruction.

6. Definition of Testing and its Types

Testing is one way to examine learners. Other major tools that can be used to measure learners’ achievement, performance, and progress are observation, questionnaires, observation schedules, interviews guides, opinionnaires, and surveys. Testing and tests cannot, therefore, be confused with assessment and evaluation. Allison (1999, np) identifies testing in the following terms: “Formal modes of assessment that are officially scheduled, with clearly delimited time on task and strict limitations on available guidance”. Popham, for instance, considers a test “a formal attempt to describe a student’s status with respect to educationally relevant variable” (2003, p.5).

Testing comes to be defined by the Free Dictionary as “A series of questions, problems, or physical responses designed to

determine knowledge, intelligence, or ability. Wikipedia definition of testing is as follows : A **test** or an **examination** (or "exam") is an assessment intended to measure a test-taker's knowledge, skill, aptitude, physical fitness, or classification in many other topics (e.g., beliefs). A tentative composite definition of the tests may be rating of learners' aptitudes, knowledge, progress, and performance in publically announced schedules, limited in time, according to certain regulations, scored items, and well-planned feedback.

7. Effects of Testing on TEFL

Testing tools could be likened to Magnetic Resonance Imaging (MRI) on account of the detailed description. Both are enormously revealing and rewarding. Koretz (2008, p.1) observes "Testing has an enormous impact on the practice of education, and it looms large in the minds of countless families as they decide where to live and whether to use public schools. Testing is revealing to all stakeholders (namely, teachers, administrators, parents, and policy makers, etc.). Basanta (1995, p.56) outlines the benefits of testing :

1. Testing tells teachers what students can or cannot do.
2. Testing tells students how well they are progressing.
3. Testing can help identify areas of remedial work.
4. Testing will help evaluate the effectiveness of the program, coursebook, and methods.

8. Categories of Tests

Four categories of tests serving different objectives are recognized by educational researchers. These are achievement, aptitude, placement, and proficiency tests.

8.1 Achievement Tests (AcT) : According to Wright (2008, p.322), the term achievement came into English (1650) from agriculture and since until modern era of testing in 1910. Achievement tests are standardized test designed to efficiently measure the amount of

knowledge and /or skill a person has acquired, usually as a result of classroom instruction such as the procedures a statistical profile used as measurement to evaluate student learning in comparison with a standard or norm. According to Hughes (2003 , p.13), AcT are “directly related to language course, their purpose being to establish how successful individual students, groups of students, or the courses themselves have been achieving objectives”. Kaplan and Saccuzzo (2008, p.309) define (AcT) an “attempt to assess what a person has learned following a specific course of instruction”.

Here is an illustration from the *Metropolitan Achievement Test* (MAT), qualified a state-of-art achievement test in the USA. The example given here tries to measure achievement in reading by evaluating vocabulary, word recognition, and reading comprehension :

A. Jennifer _____ to play house.

Pick the word that best completes the sentence

A. Wood B. book C. likes D. hopes

B. Paris is in _____ .

A. England B. Spain C. Canada D. France

8.2 Aptitude Tests (ApT) : They are standardized tests that measure learners’ ability to acquire knowledge and skills. Kaplan and Saccuzzo (2008, p.309) identify ApT as the “*attempt to evaluate a student’s potential for learning rather than how much a student has already learned*”. Moreover, ApT evaluate a wide range of experiences obtained in a variety of ways ; in the same breath , they evaluate the effects of unknown and uncontrolled experiences.

In *The Aptitude Test Workbook*, Barrett and Barrett (2004, p.4) give the following examples :

A. Which word has the closest meaning to Seek ?

a) Lose b) Pursue c) Hide d) Cover

B. Which is the one odd out ?

a) Fur b) Hair c) Feathers d) Pile

8.3 Placement Tests (PlaT) : They are tests that help teachers or administrators determine learners' knowledge/ language level, used for creating classes with distinct levels. Oxford Advanced Learner's Dictionary (OALD) defines PlaT by pointing out to the objective "to find out the appropriate level for students in a course or programme of study". In the same vein, Merriem-Webster Dictionary identifies PlaT as "A test usually given to a student entering an educational institution to determine specific knowledge or proficiency in appropriate courses or classes".

Richards et al. (2008, p.111) in *Placement and Evaluation Package Interchange (3ed.)* give the following example :

1. (6pts) Complete these sentences. Use the correct form of the verbs in the box.

Do	have	watch	go	rain	wear
----	------	-------	----	------	------

A.A(ndy) : What are you doing, Kids ?

B(illy) : _____ TV.

8.4 Proficiency Tests (ProT) : They are tests that measure learners' language background and skills (often used a placement test). Hughes (2003, p.11) observes " Proficiency tests are designed to measure people's ability in language, regardless of any training they may have had in that language". Hughes further details "the content of ProT is based on specification of what candidates have to be able to do in the language in order to be considered proficiency" (2003 , p.13). For instance, ProT would help choose a translator/ interpreter for

Halliburton in Hassi Messaoud, or to test an Algerian student's English with the view of deciding whether it is satisfactory for further studies at Oxford University.

In *How to Prepare for the California State University Writing Proficiency Exam*, Obrecht and Ferris (2005, p.5-6) propose the following topics :

8.4.1 Structured Question

We all know at least one person among acquaintances or family who behaves in an unusual matter or whose lifestyle does not conform to be accepted form of behavior. Write a well-organized essay in which you do the following :

- 1. Identify and describe an acquaintance or relative you consider unconventional.*
- 2. Explain what you believe to be unusual about this person.*
- 3. Explain why you believe this person behaves the way s/he does*

As you can see, this question is beautifully organized for you. Be sure to answer it just as it is formatted, section by section.

8.4.2 Unstructured Question

It may be a good intellectual exercise to compare the possible reasons that motivate prolife and prochoice advocates to argue for their respective beliefs.

As for the pre-tertiary levels, Green et al. (2007, p.5) provide the following example for the candidates for *California High School Proficiency Exam* (CHSPE) :

Write a short (1 or 2 page) persuasive essay.

Some people believe that young people under the age of 18 should not be allowed to drive. Do you agree or disagree ? Write a letter to the editor of your local newspaper to convince people to agree with your proposition.

9. Standardized Tests

Like millions of other students around the world, Algerian students sit three standardized tests : The Certificate of Primary Education or *sixième* (previously *Fundamental Education*), The Certificate of the Middle School Education (**BEM**, formerly **BEF**), and the **Baccalauréat** or BaAC for short. Standardized tests, particularly the BAC, are *high stake* tests as students' future prospects depend on the success or failure in these. High school standardized tests are known differently: In Britain, it is *A-Level*; Italy, it is *Maturità*, in Germany *Abitur*, in Poland *Egzamin Maturalny*, and the Israelis have the *Bagrut*.

9.1 Definition of Standardized Test (ST)

According to Wright (2008, p.322), the term In his book *Standardized Testing in Canada* (1994), Traub defines this type of

testing as follows “Standardized Test (is a test that is) administered and scored in the same way whenever and wherever it is used, so that scores of anyone taking it can be compared” (p.6). Differently stated, a standardized test such as the Bac is seated throughout Algeria at the same time using the same scoring schemes and cut scores. It noteworthy to mention, however, that standardized tests (ST) lend themselves to achievement tests. On score of that, they are usually called by their acronym SAT or CAT in the USA.

10. Typology of Standardized Tests (ST)

Educational test developers identify two types of ST : Norm-Referenced Tests (NRT) and Criterion-Referenced Tests (CRT).

10.1 Norm-Referenced Tests (NRT)

Literally, a norm refers to "an accepted standard" according to Cambridge Dictionary. Colburn (2003, p. 42) states that “NRT scores are ranked and compared to each other”. In other words, “ In NRT, the fraction of students who pass the test will be the same-regardless of how the individual students perform”. For instance, in the tests organized to hire new teachers in middle and high schools or in Magister contests (concours / مسابقات), candidates are chosen from the top list no matter what they score and in accordance with the number required to hire.

10.2 Criterion-Referenced Tests (CRT)

Deriving from Ancient Greek *kritēs* meaning to judge, criterion refers to the act of evaluating or judging. For instance, to set out to join a British university, students ought to take the International English Language Testing System (IELTS), students are taking a CRT because they are judged according to certain criteria (i.e., vocabulary, grammar, etc.). CRT are, therefore, “ones in which graders compare performance to preestablished criteria” Colburn, 2003, p.42). In different terms, “If a teacher tells students they must pass 12 questions out of 20 in a quiz, then s/he is using a CRT” (Colburn (2003, p. 42).

11. Features of Useful Tests

Validity, reliability, and practicality are the three main components that need be taken into consideration in developing *useful* tests. A useful test may be understood to be the one which yields *accurate* information on learners' *actual* learning curve.

11.1 Validity : A valid educational tool need be accepted as well grounded and true. Gipps (2011, np), validity “is the extent to which a test measures what it was designed to measure. If it does not measure of it purports to measure, then its use is misleading”. It is noteworthy to mention that “ validity is relative to purpose- a test may be valid for one purpose but not to another ” Orlich et al. (2009, p.322).

11.2 Reliability : A reliable educational measurement tool is dependable and trustworthy. Schmitt (2000, p. 166) defines test reliability as “[...] the stability or consistency of a test’s behavior over

time”. Wright (2008, p.121) expresses the same idea in different terms “Reliability expresses the extent to which test scores provide consistent and dependable information”. In other words, the same test-taker would ideally get the same results in the same test if taken in a different date or on a different occasion.

Test-takers often complain from long tests; they frequently brandish the standard of defeat “the allotted time was not sufficient”. Orlich et al. (2009, p.322.) requests stakeholders to pay heed to “One rule for you to remember is that reliability increases with test length”. Indeed, the longer the test, the chances are that test takers could score higher than expected.

11.3 Practability : For a test to be practical, it needs to be meet physical circumstances. Fulcher and Davidson (2007, p.15) point out that practicality “is concerned with the test implementation rather than the meaning of test scores”. Differently stated, practicality takes into consideration the circumstances in which a test is administered. Davies (1999, p.148) explains “The term practicality covers a wide range of issues, such as the cost of the development and maintenance, **test length**, ease of marking, time required to administrator the test (individual or group administration), ease of administration (including availability of suitable interviewers, and raters, availability of appropriate room or rooms) and equipment required (computers, language laboratory, etc.)”.

12. Theoretical Requisites of Tests

Each of the *theoretical requisites* as they come to be called by Basanta (1995) has its ramifications. In the following section a more comprehensive approach to define the features discussed upon with the view to help you both construct and administer “useful tests”. The following tables will certainly make it clearer for to single out the similarities and differences.

12.1 Test Validity

In their book *Teacher-Made Assessments : How to Connect Curriculum, Instruction, and Student Learning* (2008), Christopher Gareis and Leslie Grant look at test validity as jewelers look at a diamond. Jewelers consider the four (04) Cs : **Cut, Color, Carat Weight**, and **Clarity**. In like manner (more or less), educationalists and test designers look at validity as being four-Cs-dimensional : Construct, Content, Criterion and Consequential Validity.

Table1. Types of validity (adapted from Gareis &Grant, 2008,p. 40-42)

Type of Validity	Definitions	Questions
<i>Construct (or Face) Validity</i>	Construct related evidence of validity is concerned with how accurately an assessment aligns with theoretical framework of	Can we infer students’ knowledge and/ or or skills in this subject area from the assessment ? (Are you aiming at the right target ?)

	intended learning outcomes, standards, or objectives of the instructional unit.	
<i>Content (or Sampling) Validity</i>	Content-related evidence of validity is concerned with ensuring that an assessment adequately samples the standards or objectives of an instructional unit and conversely that it does not assess learning outcomes that were not intended for nor taught to students.	Does the assessment adequately sample the intended learning outcomes ?
<i>Criterion Validity</i>	Criterion-related evidence of validity is concerned with how accurately an assessment equates with another assessment that is intended to measure the same learning outcomes, standards, or objectives. It could be determined when 2	Does the assessment measure intended outcomes of learning that are also measured on some other assessment ? Are there items on the assessment with no corresponding intended learning outcomes ?

	related assessments occur in proximity to each other.	
<i>Consequential Validity</i>	Consequential-related evidence of validity is concerned with appropriateness of the intended and unintended outcomes that ensue from an assessment.	What are the consequences of using the assessment for decision regarding student’s learning ? For better or worse, are there any unintended consequences of this assessment for students ?

According to the aforementioned authors’ light, construct validity (CV) is usually considered the most fundamental and encompassing of the four types of validity. Unlike Basanta, they consider face validity and construct validity as the same term. For Basanta (1995, p.56), face validity (FV) necessarily requires the approval of laymen ; she notes “If a test is appealing to laymen - students, administrators, etc. – it has face validity”. Differently stated, if the format, layout, organization, scoring scheme, quality of print (e.g., no blotches or dark spots), legibility of the printing typography (e.g., typeface: alphabet, boldface, italics, numerals, and punctuation), clarity of images (or diagrams, etc.), and quality of paper are

satisfactorily represented in the test paper to the liking of laypeople, consequently, we may judge the test as having face validity.

12.2 Test Reliability

When you can trust the results of tests, then tests have achieved reliability. Basanta explains “Tests are reliable if their results are consistent, i.e., if administered to the same students on another occasion, they would obtain the same results” (1995, p.). Ideally, if the tests are administered to the same students at other times, the same results are obtained. Other factors, however, intervene not to obtain the same results such as forgetfulness. It is noteworthy to mention that reliability increases or decreases according to:

Table 2. Test reliability

<i>Environment</i>	Cold, Heat, arrangement sitting, wall color, etc.
<i>Pedagogy</i>	Type of tasks, layout, clarity of instructions, etc.
<i>Biology</i>	Hunger, thirst, lack of sleep, etc.
<i>Psychology</i>	Anxiety, anger, fear, happiness, etc.

What is more, Bach (2005, p. 126) notes three important types of reliability : **testee-**, **tester-**, and **test-**related reliability. While testee-reliability concerns the obtaining of the same results by the same student at other occasions even though the circumstances are different, tester-related reliability concerns the attribution of the same scores by different scorers to the same test-taker’s answers, and the test-related reliability means the consistency in results if taken at other times.



12.3 Test Practicality

It is interesting to mention that practicality is as an integral part of test construction and administration as the other qualities (i.e., reliability and validity). It affects the test and test agents as much as the aforementioned qualities. On account of that test developers and administrators ought to give it due consideration and seek to establish a balance between the qualities. Bachman and Palmer (1996, p.35) state “In designing and developing a test, we try to achieve the optimum balance among the qualities ..”. Practicality is then at the fulcrum of the testing process all the more as it “pertains primarily to **the ways** in which the test will be implemented, and to a large degree, whether it will be developed and used at all”. In different words, the surrounding circumstances may favor or disfavor the implementation and use of a given test. By way of illustration, in case EFL learners are on strike, it is advisable not to force them to sit an examination, otherwise the scores will certainly demonstrate a lot to be desired and students will be emotionally offended.

Practicality refers to the balance between available resources and required resources. In some instances, there are more test takers than rooms available. The shortage in these material resources directly affects academic integrity (such as cheating, plagiarizing, and copying whole passages. Intuitively, the more the rooms are crowded, the more there are chances for cheating on the examination. Bachman and Palmer (1996, p.35) propose this fraction

to explain the relationship between the required resources in the design and development and use of the test and the resources that will be available for these activities:

Available Resources Practicality = $\frac{\text{Available Resources}}{\text{Required Resources}}$ (If practicality is ≥ 1, the test development and use are practical) Required Resources
--

As for resources, three (3) types are outline below :

12.3.1 Human Resources : Human resources count for example test writers, scorers/ raters, administrators, clerical and technical support personnel

12.3.2 Material Resources: Material resources refer to physical equipments such as rooms, equipment, paper, chairs, desks, light, etc.

12.3.3 Time Resources: The temporal aspect is decisive in TEFL measurement tools construction. This includes among other issues time to develop the test from beginning to end and time on individual tasks (correcting and reporting marks).

13. Washback Effect

Backwash and washback are two terms that come in the literature on educational measurements. Although both terms refer to the same phenomenon, it appears that backwash is a favored term in dictionaries such as Cambridge, Collins Dictionary, Merriam-Webster, and Oxford Learners' Dictionary. Backwash is a term borrowed to the sea waves backward movement, which may account

for the preference. In his Foreword to *Washback in Language Testing: Research Contexts and Methods*, Alderson (2004, np) notes that in Edinburg and Lancaster universities the tradition is to use *washback* whilst Arthur Hughes at the University of Reading the preference is for *backwash*.

Washback effect, which refers to imposing effects of testing on the teaching-learning process, may be positive or negative. Brown and Hudson (2002, p.48) note " [...] a washback effect can lead to very negative consequences for the curriculum, or very positive ones-depending on what the effect is and how it is used by the curriculum developers". In case test format and contents are not in line with the pre-set objectives of a course program, then this discrepancy will ensure the appearance of the washback/ backwash effect. In the same breath, Brown & and Hudson (2002, p.48) further explain "If the tests in a curriculum do not match its objectives, they are bound to create a negative washback effect on those objectives and on the curriculum as a whole". Conversely, the closer are test format and contents from a course program pre-set objectives, the lesser are the negative effects of backwash/ washback.

14. Conclusion

Measurements in the teaching of English, which are part and parcel part of the teaching-learning process, are expected to reveal how much progress that students have made in learning a foreign language, English in this case. Also, they are indicators how

successful national curriculums have been. In short, measurements are yardsticks thanks to which the success and/ or failure of the whole teaching-learning process are gauged. While assessment is supposed to detail and report learners' progress, evaluation is meant to be evidence-based judgment of learners' progress, and testing provides the rate of intake and uptake. Each type of testing, i.e., diagnostic, aptitude, achievement, etc. is bound to achieve a specific objective. Short of matching the objectives would give ample way to negative backwash/ washback.

Bibliography List :

1. Alderson, J. C. Foreword in Liying Cheng, Watanabe, Y. & Curtis, A. (2004). ***Washback in language testing***. Routledge
- Anfara, Jr., V. A., Andrews, G. P., & Mertens, S. B. (2005). The encyclopedia of middle grades education. IAP
2. Bach, S. (2005). ***Managing Human Resources***: Personnel Management in Transition. Wiley
3. Bachman, L. F & Palmer, A. (1996). ***Language testing in practice: Designing and developing Useful Language Tests***. OUP
4. Barrett, J. & Barrett, .(2004). ***The aptitude test workbook***. Kogan Page
5. Basanta, C. P. (1995). ***Coming to grips with progress testing: Some guidelines for its design***. English Teaching Forum 33 (3), 37–40. English Teaching Forum 50 (3), 37–40

6. Brown, J. D. & Hudson, T. (2002). *Criterion-referenced language testing*. CUP
7. Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. SAGE Publications
8. Davies, P. (1999). *What is evidence-based education?* British Journal of Educational Studies. 47(2), 108-121
9. Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.
10. Gareis, C. & Grant, L. (2008). *Teacher-made assessments : How to connect curriculum, instruction, and student learning*.
11. Gipps, C (2011). *Beyond testing (classic edition): Towards a theory of educational assessment*. Routledge
12. Jarvis, P. (2002). *The theory & practice of teaching*. Routledge
13. Kaplan, R. & Saccuzzo, D.P. (2008). *Psychological testing: Principles, applications, and issues* (7th ed.). Wadsworth Publishing
14. Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press
15. Kumari, S. & Srivastava, D.S. (2005). *Education Skills And Competencies*. Isha Books
16. Hudson, P. (2015). *Hudson's guide for teaching primary science*. Australian eBook Publisher
17. Lynch, B. K. (1996). *Language program evaluation*. CUP
18. Obrecht & Ferris (2005). *How to prepare for the California State University Writing Proficiency Exam*.

19. Oerman, M. H. & Baberson, K. B.(2013). ***Evaluation and testing in nursing education*** (4th ed.). Springer Publishing Company
20. Pelligrino, Chudowsky, &Glaser. (2001). ***Knowing what Students Know:*** The Science and Design of education assessment. Washington, D.C.: National.
21. Popham, W. J. (2003).***Test better, teach better: The instructional role of assessment.*** Association for Supervision and Curriculum Development.
22. Richards, J.C, Hull, J. & Proctor, S. (2008). ***New Interchange Student's Book 3*** (3rd ed). CUP
23. Sawyer, D.B. (2004). ***Fundamental Aspects of Interpreter Education.*** John Benjamins North America.
24. Scriven, M. (1991). ***Evaluation thesaurus.*** (4 ed). Sage Publications, Inc.
25. Soled, S.W. (1995). ***Assessment, testing and evaluation in teacher education.*** Ablex Publishing Corporation
26. Sternberg, R. J. (1999). ***Thinking styles.*** CUP.
27. Traub, R. (1994). ***Standardized testing in Canada.*** Canadian Education Association.
28. Wright, R. J. (2008). ***Educational assessment: Tests and measurements in the age of accountability.*** SAGE.



