BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

HOW LANGUAGE ASSESSMENT WORKS

AN A TO Z OF SECOND LANGUAGE ASSESSMENT:
HOW LANGUAGE TEACHERS UNDERSTAND ASSESSMENT CONCEPTS

Edited by Christine Coombe

ISSN 2516-8649

Background to this glossary

The British Council's *How Language Assessment Works* project has been ongoing since 2014. The project is designed to create a whole series of tools for people interested in the area of language testing and assessment to increase their knowledge and understanding of the area.

Other elements of the project include:

- ***Knowledge of Assessment Animated Videos*** – a series of short non-technical overviews of a range of key concepts in the field. Those videos related to the testing of language skills also include worksheets and transcripts. The videos can be found at https://www.britishcouncil.org/exam/aptis/research/assessment-literacy

- ***MOOC (Language Assessment in the Classroom)*** – this is a four-week free course for teachers created by the British Council and launched in April 2018. It covers practical aspects of classroom test development, scoring and reporting. Find the course at: https://www.futurelearn.com/courses/language-assessment

An A to Z of Second Language Assessment:
How Language Teachers Understand Assessment Concepts

By Christine Coombe

How to cite this glossary

Coombe, C. (2018). *An A to Z of Second Language Assessment: How Language Teachers Understand Assessment Concepts*. London, UK: British Council.

Available online at www.britishcouncil.org/exam/aptis/research/assessment-literacy

This publication is part of the series, *How Language Assessment Works*.

Introducing the glossary and its creator

I am very honored to introduce this much needed and valuable glossary of terminology of language testing and assessment. The volume consists of hundreds of definitions of terms which have been authored by a whole range of people from those who have been recognized as being experts in the field to classroom teachers who deal with assessment on a daily basis but for some reason are not considered, particularly by themselves, to be 'expert'.

Other glossaries exist which have added significantly to our overall knowledge of the area of language testing. They were written for an expert or emerging expert readership by academic experts. They are technically accurate and are written in the technical language of the trade. This is all very well. What sets this particular glossary apart is the fact that the definitions have been written by practitioners, even where these are seen as experts in the theories of the field, they are actively engaged in the area of language test development. The other, and perhaps more important, distinction is the fact that they have been written with language teachers in mind. As such, they avoid what teachers perceive as inaccessible jargon while remaining technically accurate.

Christine Coombe has edited and curated an invaluable resource for both teachers (to facilitate understanding of key concepts in testing and assessment) and expert test developers and theorists (to allow them to more fully engage with, and understand, the needs of a previously neglected stakeholder group).

I am certain that this glossary will quickly come to be recognized and used by teachers, and testers, around the world. We are incredibly grateful to Christine for this significant contribution to the field of language testing and assessment.

Barry O'Sullivan
British Council, March 2018

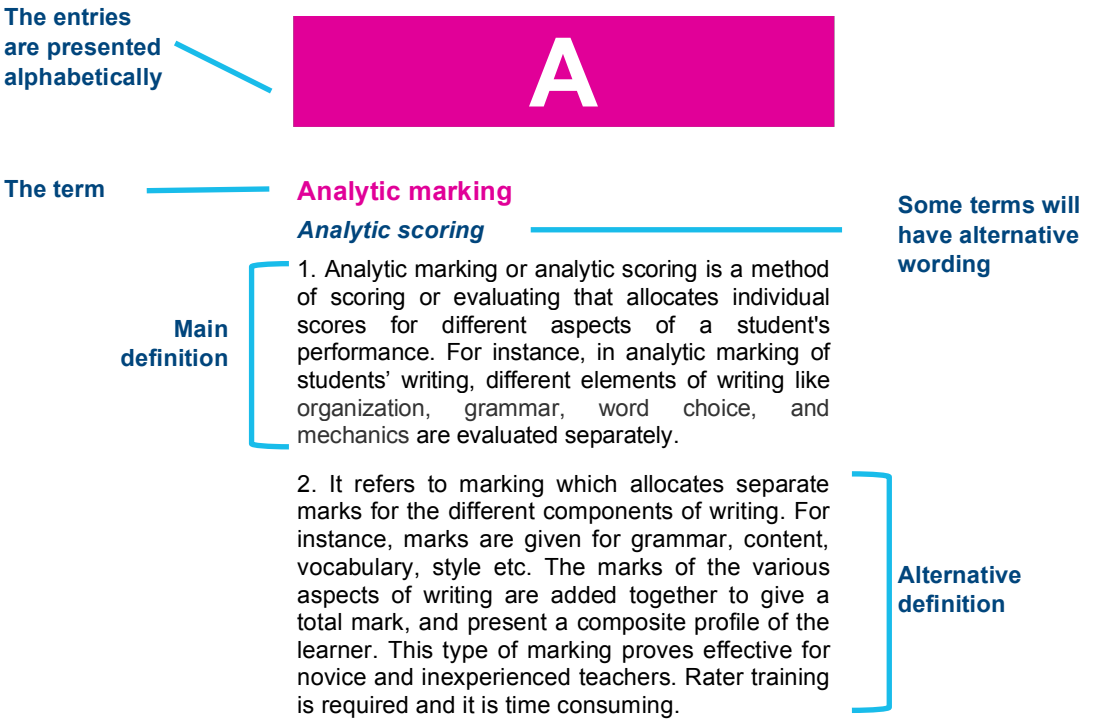
An A to Z of Second Language Assessment:
How Language Teachers Understand Assessment Concepts

By Christine Coombe

An A to Z of Second Language Assessment is an essential component of the British Council's Assessment Literacy Project and is designed for EFL/ESL teachers and those who are involved in pre-service or in-service teacher training and development. This resource, available in two formats, can be accessed online or as a downloadable PDF file. It can be used as part of a course or as a reference for those teachers who want to increase their knowledge of language testing and assessment.

Learning the terminology and jargon of the field of language assessment also means understanding the concepts represented by these terms and understanding how they are interlinked and interrelated. Written by teachers for teachers, these 285 terms and their respective definitions represent the collective knowledge of 81 teachers from 27 countries worldwide.

We have made every effort to include definitions that are readable and easy to understand, even for teachers with little or no assessment background. It is hoped that this resource will be an important one in your ongoing journey to assessment literacy.



AN A TO Z OF SECOND LANGUAGE ASSESSMENT:
How Language Teachers Understand Assessment Concepts

CONTENTS

QUICK INDEX

Click on a word or phrase to find the definition

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z

A

Academic dishonesty

Academic cheating or plagiarism

It is an act of misconduct in an academic setting. It is any type of cheating in a formal academic activity that can include copying information without proper citation, fabrication of information or data, an attempt to obtain assistance without acknowledgement, assisting others without acknowledging the assistance. As such, it is unethical behavior in an academic setting and has punishable consequences that vary in severity from one academic institution to another.

Accommodation(s)

1. An accommodation is a modification in instructional or assessment methods that allow students with physical or learning disabilities, to learn and perform tasks without undue burden. It is a way of making learning and assessment more equitable for those with disabilities.

2. Accommodation is the process of making assessment strategies and tools suitable for the characteristics of an individual in an assessment situation to make accurate interpretations of their performance.

Accountability

Assessments are typically used for the purpose of accountability: to ensure that systems, institutions, teachers and students are 'accountable' for academic achievement.

Accuracy

Accuracy is the degree of conformity to a correct or precise value, truth, fact or standard: it is the state of being correct, precise or exact.

Achievement test

It is a test that measures to what extent a student has attained skills and knowledge in a specific period of training or learning.

It is a type of criterion-referenced test that assesses the competencies gained by learners in light of prescribed or specific learning outcomes. Mid-term and final exams are good examples of achievement tests.

American Council for the Teaching of Foreign Languages (ACTFL)

Based in the USA, the American Council for the Teaching of Foreign Languages is an organization that promotes teaching and learning of all languages. In language assessment, it is particularly known for its training and certification (on-going) of oral proficiency interview testers.

Association of Language Testers in Europe (ALTE)

The Association of Language Testers in Europe is a European organization that promotes fairness and accuracy in language testing, through establishing common standards of practice in language test development and implementation.

Alternative assessment

Alternative assessment is the type of assessment that measures student performance in ways that are different from the traditional paper-and-pencil and short answer tests. Alternative assessment focuses on the individual student's overall progress. These types of assessments are not graded in the same way as standardized tests, where the number of right and wrong responses is counted. Instead, they look at the holistic performance of the students by highlighting their abilities and their overall improvement.

Some of the tools used for alternative assessment include: performance-based assessments (projects, role playing, experiments and demonstrations); open-ended questions; written essays; interviews; journals and learning logs; portfolios; self and peer assessments; authentic assessment; and others.

In order to evaluate any alternative assessment, a rubric is created as a guiding tool to help objectively evaluate the progress and performance of the students. Such rubrics help verify and identify the quality of the assignment presented. These rubrics use descriptive language in which specific words are used like "the essay contains more than three compound sentences", "the presentation uses 4–6 slides", "the interview has less than 3 grammatical errors", etc.

Alternative forms reliability

Alternative forms reliability is a means of establishing the reliability of a test through the administration of two forms of a test determined to be parallel with regard to item type, content, difficulty level, and number, along with administrative conditions that are the same in the two versions of the test.

Analytic marking

Analytic scoring

1. Analytic marking or analytic scoring is a method of scoring or evaluating that allocates individual scores for different aspects of a student's performance. For instance, in analytic marking of students' writing, different elements of writing like organization, grammar, word choice, and mechanics are evaluated separately.

2. It refers to marking which allocates separate marks for the different components of writing. For instance, marks are given for grammar, content, vocabulary, style etc. The marks of the various aspects of writing are added together to give a total mark, and present a composite profile of the learner. This type of marking proves effective for novice and inexperienced teachers. Rater training is required and it is time consuming.

Anchor item

An anchor item is an item used on multiple versions of a test or added to newly created tests to help make sure the tests are reliable (get the same results under different times or conditions). Anchor items are commonly used in high-stakes testing to ensure that tests are equivalent. Anchor items enable us to compare performance on two tests where candidates only take one test. Candidates do not need to take both.

Answer key

An answer key is a document used by practitioners to mark examinations. It has all the correct and/or acceptable answers on it.

Aptis

Aptis is a modern and flexible English language proficiency test designed to meet the diverse needs of organizations and individuals across the world. It provides reliable, accurate results about the English skills within an organization. Aptis helps companies and institutions make better decisions about recruitment, workforce development and training needs.

Aptitude test

A test which is designed to predict and measure a student's talent and ability for learning language.

Assessment

1. In language education, assessment refers to the systematic process of evaluating and measuring collected data and information on students' language knowledge, understanding, and ability in order to improve their language learning and development.

2. Assessment is the process of measuring an individual's performance on a given task in order to make inferences about their abilities. It can take different forms including tests, quizzes, interviews, written samples, observations, and so on.

Assessment for learning

Learning orientated language assessment

Assessment for learning is a type of formative assessment that creates and uses feedback to improve students' learning and performance. A student will take an assessment and use the feedback from the assessment to adjust practice and improve performance.

Assessment literacy

1. Assessment literacy is the knowledge about, and a comprehensive understanding of, students' skills and ability, interpreting the collected data from the assessments, and using these interpretations to improve students' learning and development by making appropriate decisions.

2. The fundamental know-how essential for constructing and implementing reliable test items in terms of the principles of test design, test specifications, reliability, validity and standardization. Standardized scoring or marking is also an integral element of assessment. With regards to school courses/programs, teachers need to understand the objectives of the course and align their formal and informal assessment practices to determine how far objectives are met and how teaching/learning practices may be made more effective. Therefore, not only professional assessors, but also program teachers, need to develop assessment literacy.

Assessment of learning

1. The use of a task, activity, or instrument to measure a student's level of learning or performance; also known as summative assessment, where the information generated is not then used formatively to help learning.

2. A measurement or set of strategies that provides evidence of whether the learner has successfully achieved the learning objective of a particular program.

Australian Second Language Proficiency Ratings (ASLPR)

Australian Second Language Proficiency Ratings is a scale that describes how second language proficiency develops on a scale from beginner to proficiency level, providing performance descriptions in terms of practical tasks.

Initially developed for English second language teaching, it has been adapted for English dialects in Australia, a number of other languages like French, Italian, Japanese, Chinese, Indonesian and Korean, as well as English for academic and specific purposes.

Authenticity

1. Authenticity is a feature of a test which shows to what extent a test reflects real-world situations.

2. Authenticity is the extent to which a test reflects the "real world" knowledge and skills being tested. While complete authenticity is impossible, it can be enhanced through a simulation of the skills, processes, and abilities underlying criterion behavior. Simply put, the measure of authenticity addresses the question: Does the test include situations similar to what learners will face "in real life"?

Authentic test

An authentic test is a test that involves either real-world language use (or as close as possible to this), or the kinds of language processing found in real-world language use.

B

Backwash

Washback

Backwash/washback is the result that a testing practice may have on teaching a particular program or lesson planning. It can be either positive or negative. For example, if a standardized test, based on an outdated view of language plays a very important decision-making role, teachers and learners likewise will probably decide to pay more attention to "how to pass the test" rather than "how to learn to communicate in L2". If the writing test consisted of multiple choice items rather than the skill of writing itself, this will be a clear example of negative washback.

Band score

A band score is a particular score or level awarded to language produced or a task completed by a test taker. The band score is awarded based on particular criteria set out in the band descriptors. These scores or levels are used by international institutions and government bodies to set entry requirements to courses, employment and immigration opportunities. Common examples are the setting of IELTS scores of 6.0, 6.5 and 7.0 for entry into tertiary level institutions that are usually in countries where English is spoken as a first language.

Band descriptors

Band scale descriptors

1. Band descriptors are sets of criteria or skills that a test taker can achieve or meet in test performance. Band descriptors for assessing productive skills, such as writing or speaking, often make reference to the dimensions of complexity, accuracy and fluency, as well as considering whether or not the language produced is pragmatically appropriate for the test task. Band descriptors for receptive skills, such as reading and listening, detail the micro and macro skills a test taker has including the ability to understand global ideas and specific points from the text.

2. Band descriptors are stated criteria that define the characteristics of an individual's performance on a specific task. They are used to make qualitative inferences that are reflected in different levels of performance.

Base prompt

A test prompt is input given to test takers to enable them to respond to an open-ended test item, such as writing an essay. A base prompt is the prompt expressed directly and simply, in contrast to framed or text-based prompts (Kroll & Reid, 1994).

Bell curve

A bell curve represents a normal distribution of marks where most students score in the middle, i.e., around the average (mean), some score higher and some score lower with even fewer test takers at the extremes, so it looks like a bell shape.

Benchmark

A benchmark is a standard or a point of reference against which learning is measured or compared. It is the reference point that you measure mastery, achievement or progress against.

Bias

A test or test item is considered biased when the test taker has to have some assumed cultural knowledge outside the test to understand the question or the text. For example, if an item on a test says, “Hand me a fiver,” the test taker has to know that this is slang in British English for a five-pound note.

Blind marking

Blind marking is a process when the teacher/examiner does not know whose paper he/she is marking. Instead of names, students are often asked to write their ID information on the answer sheet/examination paper.

British National Corpus (BNC)

The BNC is a large corpus (a corpus being a systematic collection of naturally occurring language samples in spoken, written or multimodal form) of British English that is divided into sub-corpora of fiction, non-fiction, academic and newspaper genres.

The BNC allows teachers and learners to search for word frequencies, collocates and word string patterns, meaning it is a useful pedagogic and assessment building resource. Although the BNC typically acts as a reference corpus, it is currently adding new spoken language components through its ‘BNC 2014’ project.

For more information visit the BNC website at: http://www.natcorp.ox.ac.uk/news.xml?ID=BNC2014

COCA (Corpus of Contemporary American English) is a larger American English corpus that is similar to the BNC and is accessible at: http://corpus.byu.edu/coca/

Boundary effect(s)

The boundary effect occurs when a test is too easy or too difficult for a specific group of test takers, with the result that scores tend to group toward the upper boundary (too easy a test) or the lower boundary (too difficult a test).

Branching test

In this type of test, there are typically two or more sections. When a candidate completes the first section he/she is directed to a particular place in the test in order to continue. This can happen on more than one occasion.

An example of a branching test is the British Council’s Online Placement Test. Here, all candidates take the first part of the test, which focuses on grammar and vocabulary only. Depending on their performance on this part of the test, they are directed to one of three sections in part two. These three sections test reading and listening at a particular CEFR level (A, B and C). The test is designed so that the first part offers an approximate indication of CEFR level (e.g. B) and the second part fine-tunes this to a more precise indication (e.g. B1).

C

Canadian Academic English Language Assessment (CAEL)

The Canadian Academic English Language Assessment is a standardized test designed to measure the English language proficiency for those wishing to enroll in higher studies.

C-test

Cloze test

A cloze test is one where the second half of every second word in a reading passage is deleted (where there is an odd numbers of letters half + one letter are deleted). The first sentence is usually kept intact. It is frequently used as a major language testing instrument. It is extremely popular because of the ease of constructing and its high reliability and validity.

Calibration

Norming session

Calibration takes place prior to marking constructed-response items or performance assessments and includes a review of how to interpret responses. It is a process during which teachers/raters work together to have a deeper understanding of how an analytic or holistic rubric works. This session also helps to establish intra-rater and inter-rater reliability. In other words, raters use a rubric to mark several benchmark papers and through discussions, they try to make sure that the marking tool is used effectively and the bands and categories are interpreted in the same way by different raters.

Cambridge ESOL

Cambridge Assessment English (formerly known as Cambridge ESOL and before that, UCLES) is a department of Cambridge University.

Its English exams are known as the Cambridge ESOL suite exams. These include: Proficiency (CPE), Advanced (CAE), First (FCE), Preliminary (PET) and Key English Test (KET). Cambridge Assessment English also publishes other tests in collaboration with international partners like the International English Language Testing System (IELTS) test, which is developed and published in partnership with the British Council and IDP Australia.

Candidate

Test taker; examinee

A candidate is someone who sits for a certain exam in order to pass that exam or demonstrate their knowledge, skills abilities and experiences. This could be on any test, including standardized tests like IELTS, TOEFL, TOEIC, GRE, etc.

Cheating

Any attempt to get marks dishonestly which may include but is not limited to: looking at someone else's paper during a test; asking someone for an answer during a test; copying someone else's work; wearing electronic devices or headphones during a test to receive outside help; bringing a sheet into a test with writing on it; writing answers on their own arm before a test, etc.

Chi-square

A statistic to determine the difference between statistically expected and actual scores. Can be used to test for bias by investigating the proportion of different candidate type that gets an item right.

Classical Test Theory (CTT)

1. A non-parametric (i.e., for a small sample size or non-normally distributed data) statistical test to examine the relationship between two nominal variables in determining the extent to which observed frequencies depart from expected frequencies, and thus the significance of the relationship between the variables.

2. Classical Test Theory is the theory that a test taker's score on a particular occasion is a combination of the test taker's true score plus an element of error. The true score is the score that would be obtained under ideal conditions through use of a perfect test, while the element of error would be introduced by extraneous factors – such as noise in the environment, unclear test instructions, or test taker fatigue.

Classroom assessment

Classroom assessment is any assessment (usually formative) done in a classroom. Assessment FOR learning is used to evaluate the teaching-learning process. It provides feedback to the teacher about his/her teaching and to the students regarding their progress in the course. This type of formative assessment is not designed to give the students a grade but to determine if they are learning the objectives as they move through the course. Summative classroom assessments are used to evaluate students' learning at the end of a course to provide some sort of a grade. This type of assessment contributes to teaching and learning through **washback.**

Close-ended question/item

Close-ended questions/items have a limited set of responses from which a respondent may choose, for example, yes/no; true/false; ABCD.

Cloze test

Cloze deletion test

Cloze test is rooted in ***Gestalt theory*** and was first proposed by W. L. Taylor in 1953. This exercise consists of a short piece of text with certain missing words (mechanical deletion of every nth word in the text), and aims to measure comprehension by asking the test takers to fill in the blanks of the removed words in the text.

Cognitive validity

Cognitive validity is the extent to which a test is valid as a predictor of real-life language performance. For example, the cognitive validity of a speaking task in a test is a measure of how closely it elicits the cognitive processing involved in other contexts, specifically, in real-life speaking tasks.

Common European Frame of Reference (CEFR)

The CEFR operates as a proficiency framework from levels A1 (Basic user) to C2 (Proficient user) and allows proficiency in European languages to be compared. The CEFR is used for a wide range of purposes including placement and proficiency tests, as well as designing level appropriate formative and summative assessment materials. The CEFR's 'Can do' statements detail what a language learner is able to achieve at a particular level in terms of task achievement. Due to projects such as the English Profile, linguistic features that are indicative of proficiency at each level are now being attached to these statements making them much more useful. The CEFR is also influential in foreign language examinations in countries such as Japan, China and the UAE.

Recently P.R. China has developed its own national framework of reference for English language education, provisionally called China Standards of English (CSE) which will play a similar role in relation to the curriculum as envisaged for the CEFR in Europe. Its implementation has the ultimate purpose of improving English teaching, learning and assessment in China.

Communicative language testing (CLT)

Broadly speaking, the main aim of CLT is to find out what learners 'can do' with language, rather than assessing how much they know about language resources in terms of lexis, grammar or other linguistic resources. In this sense, a key feature of CLT involves students using the language for an explicit real-world purpose, for example, completing a customer survey (writing) or listening to find information about hotel facilities. With CLT, the focus is on learners' ability to use the language in meaningful contexts.

Completion item

A completion item is an objective based test item in which the learners answer in a word or two words or sometimes a phrase. It is similar to "fill in the blanks".

Composite score

An overall score derived from averaging a series of scores each of which is based on similar variables

Computer-adaptive testing (CAT)

CAT is a kind of personalized computer-based testing which matches the level of difficulty of the items with the test taker's provided responses. In other words, the correct or false response of the test taker to an item alters the difficulty of the next item which s/he receives. For instance, if a student gives correct answers to intermediate-level items, the computer continues with advanced-level items.

Computer-based testing (CBT)

Computer-assisted testing; computer-aided testing; computerized testing

Computer-based testing is a form of testing which is administered by either an offline or an online computer connected to the Internet or a network. In other words, the computer is a medium of testing instead of using pencil and paper.

Concurrent validity

Concurrent validity refers to the correspondence between the scores of two different assessments measuring the same construct. The concurrent validity of a test is established through a comparison of its scores with the scores of a similar, but usually less efficient test or other quantitative judgmental procedure, that has already been reported as a valid measure of the construct in question. A correlation between the two sets of scores indicates that the first test has concurrent validity.

Congruence (fit-to-spec)

A testing score or item that demonstrates harmony, agreement or compatibility with the test specification.

Consequential validity

Consequential validity is an aspect of validity evaluating the consequences of score interpretation and use, including societal impact. Such consequences could involve, for example, issues of bias against a specific group of test takers or unfair administrative practices.

Construct

What a test measures.

Construct validity

1. Construct validity means to what extent a test is able to measure its claims. Construct validity is fundamental to the overall validity of the test.

2. The extent to which a language test is representative of an underlying theory of language learning.

3. Cyril Weir (2013) in *Measured Constructs* describes three central aspects of construct validity. **(1) Cognitive validity:** Do the cognitive processes required to complete test tasks sufficiently resemble the cognitive processes a candidate would normally employ in non-test conditions, i.e. Are they construct relevant (Messick, 1989)? Are the range of processes elicited by test items sufficiently comprehensive to be considered representative of real-world behavior i.e., not just a small subset of those which might then give rise to fears about construct *under-representation*? Are the processes appropriately calibrated to the level of proficiency of the learner being evaluated?
(2) Context validity: Are the characteristics of the test task an adequate and comprehensive representation of those that would be normally encountered in the real-life context? Are they appropriately calibrated to the level of proficiency of the learner being evaluated?
(3) Scoring validity: How appropriate and comprehensive are the criteria employed in evaluating test output? How well calibrated are they to the level of proficiency of the learner being evaluated? Also included here is the traditional concern with reliability: How far can we depend on the scores which result from applying these criteria to test output?

Construct underrepresentation

Construct underrepresentation is a threat to validity that occurs when an assessment instrument leaves out important aspects of the construct being measured. For example, an essay test claiming to measure academic writing that only looks at grammar would be leaving out important aspects of academic writing, such as cohesion, coherence, and logic.

Constructed response question

Constructed response questions are questions that require the test taker to create or construct a response. Examples of this type of test question include short answer questions and essays. Because the rater must mark written answers, this type of question is often referred to as “subjective” because marking it requires human judgment.

Content validity

Logical validity; rational validity

1. Content validity refers to how well a test or an assessment represents all aspects of a given construct. For example, a teacher gives students a test on reading comprehension. The aim of this test is to assess students’ reading comprehension. If the test does indeed measure all appropriate aspects of reading comprehension, then it is said to have content validity.

2. The extent to which the items or tasks on a test constitute a representative sample of items or tasks of the knowledge or ability to be tested. In a classroom-teaching context, they are related to a syllabus or course.

Contract grading

A form of grading which results from negotiation and cooperation between an instructor and student, commonly resulting in the number and/or qualities of assignments needed to earn a specific grade (often letter grade).

Continuous assessment

1. Continuous assessment is a kind of formative assessment like a portfolio, in which a student produces various pieces of work over the period of the course and is evaluated by the body of work at the end. This may also be a formative classroom assessment.

2. Continuous assessment refers to periodic assessments built into a running program that serve as an indication of gradual achievement of intended learning outcomes. These periodic checks support the facilitator in determining whether learning is on track.

3. This type of assessment may be formal or informal. The most distinguishing feature is that it takes place over a period of time and uses a variety of methods for assessing learners. For example, marks of in-class tests, mid-term exams, oral presentations, quizzes, homework assignments, projects, attendance etc. Through this type of assessment, we can include affective factors such as attitude, motivation and effort, which can also be considered when arriving at a holistic picture of candidate performance.

Convergent validity

Convergent validity is one type of evidence for construct validity in which there is a strong correlation between scores for the same group of test takers on two assessment instruments measuring the same construct.

Convergence cues

A multiple-choice test item may contain multiple clues which a test-wise test taker can combine to guess the correct option. Such items would contain convergence cues.

Cambridge English Proficiency (CPE)

Certificate of Proficiency in English (formerly)

A Cambridge ESOL qualification which shows the world that one has mastered English to an exceptional level. It proves one can communicate with the fluency and sophistication of a highly competent English speaker.

Criterion-referenced testing (CRT)

CRT, as compared to NRT, is a term coined by Glaser (1963). It is designed to measure student performance and outcomes using a fixed set of previously determined and concise written criteria which students are supposed to know at a specific stage or level of their education. CRTs are used to measure whether students have acquired a specific knowledge or skill set based on the standards set by the school. They are also used to identify the shortcomings or the gaps the students may have at their level of learning.

Students pass the test or are considered proficient when they perform at or above the established qualifications. CRTs include different test types like open-ended questions, multiple-choice questions, true-false questions, and/or a combination of all. Teachers' classroom tests and quizzes can be considered CRTs because they are based on the curricular standards of the students' grade level.

Critical language testing (CLT)

Critical language testing (CLT) emerged in the late 1990s through the work of scholars such as Shohamy (2001). CLT refers to the impact and uses tests have on education and society. It seeks to examine and evaluate issues of power or the power of tests and their unintended consequences on individuals who take those tests, such as high-stakes tests.

Cut-off score

Cutting score

On a criterion-referenced test, cut-off score refers to the lowest possible score on an exam or a test in order to pass the course or to be considered as qualified. In other words, it is a pre-determined score applied to filter out unqualified students on a test or an assessment. In some tests, there are multiple cut-off scores which represent different proficiency levels, for example basic, intermediate, or advanced.

Curve grading

Grading on a curve

This method of grading is designed to establish a normal distribution of students' grades on a curve based on the average students' grades in any particular class. The term "curve" refers to the bell curve which looks like a bell shape when students' grades are plotted on it. It shows the density of the normal distribution of the students' grades. Student numeric scores are assigned to the students, which in turn are converted to percentiles. The percentile values in turn are transformed to grades by turning the scale into intervals, where the interval width of each grade designates the desired frequency for that grade. This means there are a fixed number of students who will get As, Bs, Cs, Ds and Fs. The majority of the grades will be Cs; there will be a similar number of Ds and Bs; and a similar number of students getting Fs and As. One goal of grading on a curve is to reduce or eradicate deviations in grades that can occur due to the results students receive because of having different instructors for the same course. This is also done to make sure that the students in any given class are assessed compared to other peers in the same course. Norm-Referenced Testing (NRT) is a typical example where students' grades are plotted on a curve.

D

DIALANG

1. DIALANG was developed under the coordination of Professor Charles Alderson at Lancaster University for the purpose of diagnosing test takers' language ability in 14 European languages with reference to the Common European Framework of Reference for Languages. As the first internet-based diagnostic test, DIALANG identifies learners' strengths and weaknesses in each of the dimensions assessed and provides advisory feedback to learners.

2. This is a diagnostic online reporting tool that requires users to complete tasks and then outputs a proficiency level. It can be used with a range of European languages across language skills including reading and writing. The tool outputs a CEFR benchmarked proficiency rating e.g. B1.

Dictation

1. Dictation is an example of an integrative test and attempts to reflect how candidates process multiple aspects of language simultaneously. Dictation tests an array of areas including memory span, spelling, grammar, recognition of sound segments and overall listening comprehension skills. However, care needs to be taken while scoring dictation as all errors may not be considered equally serious.

2. An action of saying or transcribing words or the stretches of texts with the goal of listeners writing them down accurately. It is used to practice students' listening and mainly spelling skills. Dictation can be used when teaching orthographic skills and grammatical conventions of a language.

Diagnostic test

1. A diagnostic test is a test that helps both teachers and students identify or diagnose student strengths, weaknesses, and areas of difficulties.

2. Such tests are administered so that teachers can identify the weaknesses and conduct remedial teaching or identify the strengths to guide further input.

Differential item functioning (DIF)

Differential Item Functioning occurs when a test item is statistically shown to result in a particular group of test takers getting higher or lower scores than other test takers. DIF can involve item bias if the difference in scores is irrelevant to the trait being assessed, as, for example, when a test item is based on irrelevant cultural knowledge unfamiliar to a specific group.

On the other hand, if a group of test takers receives higher scores because of more extensive study of the trait being assessed, then DIF would not involve bias.

Digital assessment methods

Digital assessment describes any assessment that is conducted via digital means on computers, PCs, laptops, tablets, smart phones, or any other devices that connect assessments, learners, and instructors on the cloud. Such assessments can be used in face-to-face classrooms, as well as in online synchronous and asynchronous instruction scenarios. Examples of such assessments range from short polls and quick surveys that immediately display class results for all to see to longer assessments, essays, and all other forms of creations by students (project-based learning such as crossword puzzles or other products), as well as collaborative products generated by learners on the cloud (such as Google Drive). There is an infinite number of platforms that can be utilized for digital assessments and new ones are created and released daily.

Direct test

A test in which the test is an authentic task in which the students demonstrate they have learned course objective(s) successfully. A direct test for a language learner to write an essay will be to write an essay. This is opposed to an **indirect test** which might require the student to take a multiple-choice test identifying the required parts of an essay.

Discrete-point test/item

1, A discrete-point test refers to the testing of one element or item at a time, commonly seen in tests of individual grammatical items, words or language functions.

2. This term refers to a component in testing when only one element is being tested at a time. It is not integrated within other items that are being tested.

Discrimination

1. The ability to ascertain different levels of performance by different levels of test takers.

2. Discrimination refers to the idea that a test item should discriminate between strong and weak test takers. In other words, you would expect strong test-takers to answer a difficult test item correctly, whereas you would expect weaker test-takers to answer that test item incorrectly (see **Discrimination Index**).

3. Refers to a test's fundamental property of differentiating or capturing the range of individual test-takers' language abilities – in other words, separating high-ability from low-ability test-takers.

Discrimination Index (DI)

1. A measurement of how well a test item can differentiate between the performance of those who do well on a test (high scorers) and those who do poorly on a test (low scorers). Range is -1 to +1. The higher the positive correlation, the better the test is doing its job.

2. The Discrimination Index is the measure of how well a test item distinguishes between strong and weak test takers. DI is calculated using point-biserial correlation which compares the test takers' performance on an item (i.e. whether they got it right or wrong) with their overall test scores. DI is measured between -1.0 to 1.0. Deciding on an acceptable DI for an item is largely dependent on how easy or difficult the item is. For most testing purposes, a DI of 0.2 and above is acceptable. However, you would not expect an item that is very easy or very difficult to have a high DI, so you might accept a lower DI for those types of items. When a test item has a low DI, or even a negative DI, it should be examined and possibly discarded (also see **Discrimination**).

3. Refers to a measure of the extent to which test-takers known to be high on the attribute being measured exhibit responses to a test item that differ from those known to rank low on the attribute. The index is a measure of the discriminating power of the item.

Dispersion

Dispersion refers to the spread of test scores of a test-taking group (such as men, women, undergraduate or graduate students, first language or second language test takers, etc.). One of the most widely used statistical indices for dispersion is *Standard Deviation* or SD. Other indices include *Variance* (which is Standard Deviation squared) and *Range* (which is the difference between the highest score and lowest score).

When the dispersion is small, that is the SD is small, we can say that the test-taking group is homogenous in the ability or subject area tested. When the dispersion is large, that is the SD is large, we can say that the test-taking group is heterogeneous in the ability or subject area tested. Statistical software such as SPSS produces dispersion indices in its Frequency or Descriptives procedure.

2. Refers to how spread out a set of data is (e.g., test scores). Measures of dispersion can provide information about variations in the data (e.g., range and standard deviation).

3. Hatch and Lazarton (1991) suggest that in norm-referenced tests, a small SD indicates that the true variation in a population has not been measured. Since the SD is an indicator of reliability, the higher the SD the more reliable the test is likely to be.

Distractor

A multiple-choice item consists of a stem and options. One of the options is the key/correct answer, the others are distractors – the aim of which is to "distract" the test taker's attention from the key. Distractors must be plausible.

Distribution (see also dispersion)

A distribution is the spread and pattern of scores or measures in a group of test takers.

Dynamic assessment (DA)

1. Dynamic assessment is an approach to assessment in which learners are provided with fine-tuned assistance whenever the need arises. It is derived from Vygotsky's sociocultural theory of mind. According to Sternberg and Grigorenko (2002, cited in Poehner, 2008, p.13), dynamic assessment, "takes into account the results of an intervention. In this intervention, the examiner teaches the examinees how to perform better on individual items or on a test as a whole".

2. A dialectical mode of assessment that, instead of obtaining a static measure of a learner's proficiency, incorporates mediation in the assessment process with the goal of revealing the range of the learner's present and potential abilities in support of his/her development.

E

English Language Test for International Students (ELTiS)

This academic test for secondary level or high school students aims to assess the listening and reading comprehension skills of students whose first language is not English. ELTiS scores facilitate the decision-making process for selection and placement of students into overseas language programs and schools. The scores are generally used for the placement of exchange students who need to be able to function in English in an American school environment. The test comprises two sections – listening and reading. The listening section is 25 minutes and the reading section is 45 minutes; the total test time is 70 minutes. The test consists of multiple choice questions.

Entrance test

An entrance test is usually a high-stakes test which applicants take as a requirement of an educational institution. The purpose of an entrance test is to give decision-makers an overall understanding about the test taker's knowledge and ability to use the language.

Empirical validity

Empirical validity refers to a type of validity which is calculated for a recently developed test by correlating its results with those of an already established test. There are two approaches to the measurement of empirical validity: concurrent, and predictive. In the concurrent approach, our newly-developed test and an already established test are administered at the same time to the same group of test takers to find the degree of correspondence among their scores. In predictive validity, there is a time gap between the collection of the two sets of scores. For example, a test which claims to measure scholastic aptitude might be administered to a group of test takers at the beginning of the academic year, and later on, its results will be correlated with the final achievement scores of the students at the end of the academic year to see how well our test could predict the test takers' academic achievement and performance.

Essay question/tests

A performance based test which enables evaluation of the writing skill. Candidates are asked to respond to a prompt and have considerable freedom of response. It is relatively easy to construct essay questions but scoring may be subjective.

English for Specific Purposes test

ESP / LSP test

1. An ESP test refers to an assessment of English for Specific Purposes (ESP). The defining feature of an ESP/LSP test is arguably its subject-specific content, which arises from an analysis of specific target language use situations involving a comprehensive and fine-grained needs analysis of a particular occupational or academic group. Typical examples are OET (Occupational English Test) designed for health professionals, and ELPAC (English Language Proficiency for Aeronautical Communication) designed for air traffic controllers.

2. ESP testing refers to measuring ESP assessment. ESP testing may be for placement, achievement, and/or proficiency purposes. In planning the test, the test developer must consider ESP content material taught, instructional methods used, intended learning outcomes, test specifications, instrument development techniques and test item strategies, test reliability, validity and/or practicality, to create the test.

Equated forms

Test equating

Equated forms are two (or more) forms of the same test (i.e., built to the same specifications) that have been related through a statistical analysis called **test equating.** Test scores from equated forms are considered to be equivalent, even though the individual items on each form differ.

Equivalent forms reliability

Equivalent forms are alternative forms/versions of a test that are similar in all important properties. Evidence needs to be collected to show that different forms/versions are comparable in test content, and that the factor loadings and measurement errors are approximately the same across different forms/versions. Equivalent forms reliability, or parallel forms reliability, can be used to measure test reliability, which is the consistency and freedom from error of a test, as indicated by correlation coefficient obtained from responses to two or more equivalent forms/versions of the test.

E-rater

E-rater® is an automated essay scoring program offered by Educational Testing Service (ETS). This program was first developed by a team of experts in **educational assessment** and natural language processing under the supervision of Jill Burstein in 1999. The program is now used by ETS as the second rater in scoring the writing component of TOEFL IBT.

Error

Refers to a deviation from established accuracy. It may be the result of oversight or insufficient understanding.

Error count marking

Error count method refers to the method of marking compositions or a piece of writing by counting the total number of errors made and subtracting them from the total number of words. It is a mechanical procedure of marking which may fail to accurately reflect learners' strengths and weaknesses in writing.

Error recognition item(s)

Items in a text selection of which at least one has a syntax and/or logic error. It is the job of the learner to identify the incorrect element of the phrase or written text.

Ethics

ILTA Code of Ethics

Ethics in language testing and assessment refers to a system of established and acknowledged principles that guide the construction, development, administration, management and evaluation and use of language assessments in any context.

Educational Testing Services (ETS)

ETS is the world's leading private non-profit educational testing and assessment organization. It was founded in 1947 and its headquarters are in New Jersey, USA. ETS has developed a range of standardized tests mainly in the United States for K–12 through higher education. It also develops and administers international tests such as the TOEFL (Test of English as a Foreign Language), TOEIC (Test of English for International Communication), and the Graduate Record Examination (GRE) for applicants who intend to study and work in an English-speaking country.

It has established more than 9,000 test centers in over 180 countries. ETS has also conducted the US National Assessment of Educational Progress (NAEP), which is referred to as the Nation's Report Card.

European Association of Language Testing and Assessment (EALTA)

EALTA is an organization in Europe dedicated to language testing and assessment. Its mission is to improve, share, and promote better understanding of language testing and assessment among language testers in Europe and beyond.

Evaluation

1. Evaluation refers to the process of utilizing assessment results to gauge and support learning, teaching and program development. Evaluation can take place at four levels:

- **Learner feedback** that measures learner reaction to the program
- **Learner learning** that measures learner progress following the learning (through assessments)
- **Learner behavior** that measures observable skills acquired following learning
- **Learning results** that measure the impact of skills/knowledge acquisition on the environment in which learners function post learning.

2. Evaluation is the principled and systematic collection of information for the purposes of decision making about the worth of something. Evaluations provide useful information about manifold areas and the evaluation of programs and curricula are indispensable in the current educational climate as we need evidence of explicit outcomes and achievement of aims.

3. The systematic gathering of information for decision-making related to learning, instruction, or a program.

Examinee

Test taker; candidate

A person/test taker who is examined in a test by examiners.

Examiner

A person whose job is to inspect, set and mark exams of test taker's knowledge or proficiency. In an oral interview test, an examiner can also play the role of an interlocutor who interacts with examinees.

External validity

Criterion-related validity

A statistical measure of the extent to which a test is close to its established criterion measure; also known as criterion-related validity.

Exit test

An exit test is designed to evaluate if students have demonstrated adequate knowledge, skills and/or ability upon completing a program so as to facilitate decisions on graduation or certification.

F

Face validity

Logical validity

1. Face validity is a non-technical term that refers to the degree to which people subjectively perceive a test as being fair, reliable and valid.

2. The extent to which a test appears to users to be an acceptable measure of their ability.

3. Face validity is one dimension of validity which assesses whether the test looks right to the test taker and test user. It pertains to the surface features of the test. In other words, does the test appear to be acceptable to the candidates and other stake-holders? There are no statistical calculations associated with this type of validity.

Facility value

Item discrimination value

Facility value is a numerical representation of how easy or difficult a multiple choice question (MCQ) item is for a given pool of students. To calculate the item facility value, the number of students answering the item correctly is divided by the total number of test takers. The range of the facility value is 0-1. The closer the value is to 0 the more difficult the item is. The closer the value is to 1 the easier the item is.

Fairness

Relates to validity during test construction, consistency of implementation during test administration, and consequences for use of test scores for individuals or groups, examiners' judgments or evaluations, and society at large.

First Certificate in English (FCE)

The Cambridge English: First Certificate (FCE) equates to a CEFR B2 level and indicates that the holder is able to study or work in an English-speaking country. It is the third level of the Cambridge English Main Suite exams. The test consists of 4 computer or paper-based exams which in total last about 3.5 hours. The exams are Reading and Use of English, Writing, Listening and Speaking. The 75-minute Reading and Use of English exam has 7 parts and uses varied texts to test control of grammar and vocabulary. The 80-minute Writing exam requires the writing of two different genres (e.g. a letter and an essay). The 40-minute Listening exam has 30 questions spread over 4 parts, each of which uses a different genre of spoken English. The 14-minute Speaking exam has 4 parts and is conducted in pairs or threes. A group of three will extend the length of the exam.

Feedback

Feedback is the teacher's or peer's opinion (usually justified through rubrics) about the learner's written or oral performance on either a traditional test or an authentic task. The teacher can choose to give either written feedback or oral feedback during the teacher-student conference. Feedback is an essential part of assessment and sometimes can play a more positive role in students' learning than marks, in that they are provided in a stress-free atmosphere and are aimed at helping students find their strengths and weaknesses.

Field testing

Pilot testing

1. In some testing programs, a distinction is made between pilot testing and field testing. Field testing might mean piloting a test on a larger scale. The purpose of field testing is to evaluate the feasibility or practicality of the operational use of a test/item. In computer-based assessments, field testing is also conducted as a stress test.

2. Pilot testing refers to a small-scale trial of a test instrument for addressing potential issues with the test, for example, test items, instructions, and format.

Flesch-Kincaid Readability Index

The Flesch-Kincaid Readability Index is an online tool for testing or measuring readability level (see websites below). In other words, it shows how difficult a reading passage is. There are two kinds of readability tests: The Flesch Reading Ease and the Flesch-Kincaid Grade Level.

http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.php

http://www.readabilityformulas.com/flesch-grade-level-readability-formula.php

Fluency

Fluency is an assessment category or criterion most commonly associated with features of speech delivery, such as delivery rate, number of pauses and repetitions, and frequency of breakdowns and self-corrections. In the assessment of writing skills, writing fluency is most often measured by the composition rate, along with other indicators, such as the number of words correctly spelled, sentences written, holistic rating, and so on.

Forced-choice item

A forced-choice item refers to any items that require a test taker to choose between response options, such as true/false and multiple-choice questions.

Formal assessment

Formal assessment is part of a formal course design or evaluation scheme, which takes the form of written tests, student products marked with a rubric, or standard-based performance assessments. Formal assessment may contribute towards a grade. It is contrasted with **informal assessment,** which can take place continually throughout a course and depends on the informed judgment of the teacher, who evaluates and gives feedback spontaneously or in situations prompted by the student, such as during classwork, or outside the classroom.

Formative assessment

Formative assessment generates data on student learning while it occurs. In contrast to summative assessments or standardized testing capturing the point of knowledge or mastery after a lesson or unit, formative assessment is conducted during the learning process. Formative assessment is a way of framing learning activities such that they generate observable and measurable data for teachers and learners alike. During formative assessments, learners can discover what they have mastered already and which areas they need to improve.

At the same time, teachers can discover which knowledge, skills and abilities should be clarified and strengthened. There are multiple ways of conducting formative assessments. Examples include: asking for feedback by show of hand or digital clicker/polling systems exhibiting class results; exit slips; and observation of student problem-solving activities in the classroom; as well as more sophisticated ways of generating products using higher-order thinking skills on Bloom's taxonomy.

Framed prompt

Kroll and Reid (1994) classified writing prompts into three types: **base, framed and text-based**. Base prompts state the entire task in simple and direct terms, whereas framed prompts present the writer with a situation that acts as a frame for the interpretation of the task.

Task 1 on the IELTS General Training writing test is an example of a framed prompt. Here is another example:

On a recent flight, British Airlines lost your luggage. Write a complaint letter to the General Manager, Mr Brown, telling him about your problem. Be sure to include the following:

- *your flight details*
- *a description of the lost luggage and its contents*
- *what you would like Mr Brown to do.*

Free writing

It is special writing strategy when writers write on a specific topic or a prompt for a period of time without paying attention to mistakes, organization and other conventional writing rules. The goal is to start writing, to generate ideas, to develop some thoughts and/or break the psychological barrier against writing. It is also used as a prewriting technique before a specific writing task. Free writing can be a useful tool to help students get rid of the fear of writing in another language, it can also be used as a planning strategy for a more structured writing task.

Frequencies

1. The count of the number of times that a particular feature appears in the performance of someone being assessed.

2. Refers to the number of occurrences (frequency counts) of a particular item in a data set, for example, counts of categories or responses, or how often specific response options occur in the pool of test takers.

G

Gap fill item

A gap fill item is an objective test format in which candidates are required to fill in blanks with word or phrases in a sentence or a text. When options are provided, candidates can choose the correct or most appropriate option for a gap.

Gatekeeping exam

A gatekeeping exam refers to a high-stakes exam for screening or selective purposes. The exam either opens or closes the door for examinees to opportunities of learning, working or immigration. A gatekeeping exam has important educational and social consequences on individuals and the society.

General Certificate of Secondary Education (GCSE) (UK)

GCSE is an academic qualification awarded to students in secondary education in England and Wales, who are assessed in a number of subjects over two years.

Generalizability

1. The extent to which the results of an assessment in one area can predict the results in another.

2. Generalizability is extending results/conclusions from findings of a research study. These findings can be extended from the sample and apply to the whole population.

Generalizability theory

Generalizability theory is a statistical framework to assess reliability of measurement under certain conditions. It is helpful in evaluating assessments' reliability.

Global error

1. Global errors are communicative errors which prevent the learner from comprehending some aspects of the message. Wrong order of the main constituents, missing, wrong selection, and misplacing of connectors are identified as global errors.

2. A global error is an error which impacts on the organization or coherence of a written sentence or spoken utterance and is one which causes comprehension strain for the reader or listener. An example of such an error might be word order or paragraph cohesion.

Grade inflation

Score inflation

Grade inflation is to continually grant increasingly higher academic grades to students who would have previously received lower grades, resulting in "intellectual bankruptcy".

Grading on the curve

Curve grading

Curve grading is the method of grading designed to establish a normal distribution of students' grades on a curve based on the average students' grades in any particular class. The term "curve" refers to the bell curve which looks like a bell shape when students' grades are plotted. It shows the density of the normal distribution of the students' grades. This works when numeric scores assigned to the students are converted to percentiles. The percentile values in turn are transformed to grades by turning the scale into intervals, where the interval width of each grade designates the desired frequency for that grade. This means that there are a fixed number of students who will get As, Bs, Cs, Ds and Fs. The majority of the grades will be Cs; there will be a similar number of Ds and Bs; and a similar number of students getting Fs and As. The goal of grading on a curve is to reduce or eradicate deviations in grades that can occur due to the results students receive from different instructors for the same course. This is also done to make sure that the students in any given class are assessed compared to other peers in the same course. Norm-Referenced Testing (NRT) is a typical example where students' grades are plotted on a curve.

Guessing factor

Guessing

The guessing factor represents construct-irrelevant and seemingly random behavior of candidates when responding to test items. The most typical example of the guessing factor is in multiple choice listening or reading tests. When a candidate does not know the answer to a question, they simply pick an option at random in order to increase their odds of earning a mark. Guessing factors can be accounted for in some measurement models (e.g., Item Response Theory). It is important to note that all guessing is not truly random, and guessing may be used in coordination with other test-taking strategies.

Guided writing

Guided writing is the type of writing where the teacher works with his/her students on a specific writing task, supporting them to improve and scaffolding their writing towards more independent writing. Guided writing personalizes learning by enabling the teacher to adapt teaching to the specific needs of the group and facilitating the teaching and learning of each individual student. The teacher observes, monitors and responds to the needs of each student within the group by giving them individual guidance and immediate feedback on their achievements, discussing areas for improvement. The teacher provides the students with specific concepts and strategies to help them in their writing. Students experience continuous and sustained attention to their writing by producing short, but complete pieces of writing. Guided writing can be fully utilized by providing the learners with the language they need to complete the task together with the teacher. In a guided writing lesson for example, the teacher may say "what do you think about…," "can we say …," or "how about if we say…"

H

Halo effect

1. The halo effect refers to a type of undesirable rater behavior in which a candidate's performance in one dimension influences a rater's judgments on other dimensions. For example, a rater may be thoroughly impressed by the accuracy and complexity of grammar in a candidate's essay and end up giving undeservedly high marks in vocabulary, content, and organization.

2. The halo effect happens when you are influenced by looking at someone's appearance, fame, reputation, attitude and behavior which impresses you without you knowing them. For example: you will talk to some prestigious company representative rather than an unknown new company representative. The reason is because you have more confidence (without any facts) with the big well-known company as they already have a good reputation in the community.

High-stakes testing

1. High-stakes testing is used to make important decisions with consequences for one or more stakeholders (students, educators, schools, etc.) in the assessment process. An admissions test of a university is an example of a high-stakes test.

2. High-stakes tests have immense power over candidates' lives because the results of these tests have serious consequences. These are tests which hold the key for entrance into college/university programs. For instance, IELTS is an example of a high-stakes test which is a prerequisite for overseas university programs.

Holistic marking

Impressionistic marking; impression marking; global marking

1. Holistic marking is a type of assessment practice that is used to evaluate productive skills. This form of marking uses integrated criteria to reflect the overall effectiveness of written or spoken discourse. It may also reflect the impact that a speaker or writer has on the listener or reader. Holistic scores can be used together with analytic scores in order to arrive at an accurate and informed decision.

2. Holistic marking is also called impressionistic or global marking. Generally speaking, it is related to the grading of writing and speaking performances where markers award a mark based on their overall impression of the performance. Marking is swift and perceived to be reliable if two or more markers rate the same script. It is effective and suitable for large classes where examiners must mark a large number of scripts.

I

Impact

Impact refers to the various ways in which test use affects society, an education system, and the individuals within these. Impact operates at two levels: a macro level, in terms of the societal or educational system in general; and a micro level, in terms of the individuals who are affected by the particular test use.

International Development Program (IDP Australia)

IDP Australia is an international education organization based in Australia which, besides offering educational services to international students in higher education, is co-owner of the high-stakes English language test – IELTS. IDP Australia partners with Cambridge English Language Assessments and the British Council for IELTS tests.

International English Language Testing System (IELTS)

IELTS is a test that is owned by three partners: the British Council, IDP Australia and Cambridge English Language Assessment. The test has four parts: Listening, Speaking, Reading, and Writing. IELTS has two separate modules: Academic and General. The academic module is for candidates entering educational institutions. The general module is for candidates applying for residency in or immigration to Canada or the UK. IELTS is a high-stakes test accepted by thousands of universities and workplaces around the world.

Impression marking

Impressionistic marking; holistic marking; global marking

The learner's overall performance is marked without focusing on any specific aspect or criterion. This approach to marking is often based on a single rating scale aiming to provide an approximation of the learner's ability in, for example, placement or progress tests where errors of judgment can be rectified at a later stage.

Impure test item

In discrete-point testing, good items test only one thing. An impure test item tests more than one thing.

Incident report

Most standardized assessments (e.g. TOEFL, IELTS) feature an incident report procedure within their application standards through which invigilators or proctors can report situations or events that may compromise the integrity of the assessment. Some examples of incidents that might occur include disruptive behavior from a candidate, or irregularities in the handling of materials or times of the test. Incident reports are usually submitted in a paper format that is meant to be filled out and signed by invigilators/proctors and those responsible for the test, then added to the testing package and sent back to the evaluating agency.

Incremental validity

Incremental validity refers to the collection of evidence to determine if a new assessment will provide more information than measures that are already in use. If a new assessment does not provide any new information than the current or simpler measures are already providing, the new assessment is considered to be unnecessary.

Informal testing

Informal testing can refer to the application of a testing procedure in settings that do not conform to the regular (or formal) testing procedures. The purpose of informal testing is not the formal evaluation of a candidate. For example, some schools apply sample papers of certificate exams to students prior to their official test date as preparation, so that students acquire experience with the test's tasks, formats and conditions, as well as a sense of their possible performance in the formal test.

Integrative testing

Integrated testing

Refers to a testing format requiring test takers to combine various skills in answering test items or performing test tasks.

Interlocutor

1. In a speaking test, the interlocutor is the interviewer who interacts with the candidate(s) for the full duration of the test. Interlocutors need to follow a standardized format in their speech and behavior, and are nearly always the assessors too, following set criteria and descriptors laid down by the specific tests they are administering.

2. The interlocutor is someone who responds to, or interacts with, examinees during a speaking test.

Internal validity

A measure of the extent to which the results of an assessment can be trusted to be determined by known variables and not unknown variables

International Language Testing Association (ILTA)

ILTA is a professional association in the field of language testing and assessment.

Inter-rater reliability

1. When the observed behavior or performance in a test is used as data to evaluate the performance of the students, the observation must be reliable. This reliability could be achieved by having two or more examiners to observe the same performance or grade the same test and use the same scoring guide for evaluation. Their scores must be highly correlated to ensure that they are evaluating the construct in the same way.

2. Inter-rater reliability refers to the level of agreement between two or more independent raters in their assessment of test takers' performance.

Interview

1. An interview is a test task used in speaking assessment. It can be used as a noun or a verb to indicate a conversation or the action of conversing with someone on a specific topic. The interviewer has a set of questions and proceeds in a structured, semi-structured or unstructured manner to solicit responses from the interviewee(s) with the goal of learning about their opinions/views/facts/statements about some topics/questions. In the TESOL context, interviews can be used as one of the tools to develop or assess speaking skills in another language. The learners can be given cards with interviewer and interviewee roles, a set of sample questions and answers and clear directions on how to role-play the interview.

2. An interview is a formal discussion between two or more people about a certain topic. The interviewer asks a range of prepared and follow-up questions while the interviewee responds with his/her own views. The interviewer should be unbiased in asking questions and responding to interviewees. The interviewer's role is to be engaged in the conversation and gather as much information from the interviewee as possible, using both open and closed-ended questions.

Interviewer bias

Interviewer bias happens when an interviewer leads the interviewee into giving responses based on what the interviewer wants to hear. Such questions affect the outcome of the interview negatively and may also distort the reliability of the results.

Intra-rater reliability

1. Refers to the internal consistency of the rater. In other words, the rater must be trained to use the rubric effectively to rate similar written or oral performances of the same or different students in the same way.

2. Refers to the extent to which a rater is internally consistent in using a rating scale by comparing two independent assessments of the same test performance by the same rater.

Invigilation/invigilator proctor (US)

1. Someone who proctors a test, keeping an eye on test takers during a test to prevent cheating or other improper behavior.

2. The person responsible for observing the administration of an assessment. In a formal proficiency assessment, an invigilator/proctor is not necessarily the test's administrator, but they generally must be people who are not involved with the testing group on a teaching or personal capacity. An invigilator or proctor must be trained on the testing procedures, and can report incidents and instances of breach of protocol to an administrator.

Item Response Theory (IRT)

IRT is an approach to test item analysis based on the relationship between test taker performance on test items and underlying test taker ability.

Item analysis

Item analysis involves statistical analysis of characteristics of administered test items, for example, **item difficulty** and **item discrimination**.

Item bank

An item bank is a large collection of test items classified by characteristics such as content or item characteristics (e.g., discrimination and difficulty). Item banks can be used in large-scale computer-adaptive tests or in classroom tests, as in when teachers are provided with a bank of test items by a textbook publisher classified by content, item type and difficulty.

Item difficulty

Item facility

Item difficulty is a statistic indicating how difficult (or easy) a test item is for a specific group of test takers. Item difficulty is usually expressed on a scale from 0% to 100%. If an item has a difficulty index of 90%, it indicates that 90% of the test takers correctly answered the item, and thus the item was relatively easy for that particular group of test takers.

Item discrimination

Item discriminability

Item discrimination indicates how effectively a particular test item distinguishes between low-performing and high-performing test takers in terms of the trait being measured by the test item. Analysis of item discrimination can reveal how effectively an item functions in a particular administration of a test. For example, if an item is poorly constructed with nonsensical options or unclear instructions and thus has low item discrimination, it would not effectively distinguish between test takers with high levels and ones with low levels of the trait being assessed.

Indigenous assessment criteria

During the rating scale development process, assessment researchers recommend clarifying ***the subjective judgments*** of the relevant rubric users regarding the quality of knowledge and usage of a language in a target context. The technical term 'indigenous assessment criteria' is defined by Knoch (2009) as "[T]he criteria that experienced professionals use for evaluating communicative language use" (p.22).

K

Key

Answer key

1. The key of a test item refers to the correct or acceptable answer. The term is typically used in the context of selected-response items (such as multiple choice, matching, and true/false) where there is a clear correct or acceptable answer called the key which appears a choice along with incorrect or unacceptable choices (called distractors). When a test taker chooses the key as his/her answer to an item, he will receive a point or score; if he/she chooses a distractor instead, he/she will not receive a point or score. In rare cases, items may have two keys to indicate one fully correct or acceptable answer that will receive full points (say, 2 points) and one partially correct or acceptable answer that may receive partial points (say, 1 point). Clearly identifying the keys for all selected-response items in a test will help teachers or graders score items faster and with few errors.

2. Documentation of the correct responses for a test or assessment measure

3. The key is the correct answer to a particular test item. The **answer key** is all the correct answers to a particular test.

L

Language Assessment Quarterly (LAQ)

Language Assessment Quarterly is an international publication focused on issues to do with language assessment. This peer-reviewed publication publishes advanced theory, research, and practice in theoretical issues, empirical research and professional standards.

Language for Specific Purposes Testing

LSP Testing; ESP testing

ESP testing refers to measuring English for Specific Purposes (ESP) assessment. ESP testing may be for placement, achievement, and/or proficiency purposes. In planning the test, the test developer must consider ESP content material taught, instructional methods used, intended learning outcomes, test specifications, instrument development techniques and test item strategies, test reliability, validity and/or practicality, to create the test.

The test is created based on identified intention. Hence, the testing process demonstrates the person's degree of ESP content or skill acquisition; thus, inferring how much information or what aptitude (ability) is acquired. For example, the ESP testing may show that a person is able to successfully write up and negotiate a job employment contract or carry out a Skype conference call in English.

Language Testing

Language Testing is a peer-reviewed international journal that publishes original research articles which contribute to methodological innovation and the practical improvement of language testing and assessment in first and second language. It provides a forum for the exchange of ideas and information among the people working in the fields of first and second language testing and assessment.

Language Testing Research Colloquium (LTRC)

LTRC is an annual conference that began in 1979 and is attended by active researchers to discuss the issues and problems in language testing and assessment. The International Language Testing Association (ILTA) was formed as a part of this conference held in 1992 in Vancouver. In the initial years, the conferences were held in English-speaking locations, focusing primarily on testing issues in the English language. In the past 20 years, however, the conferences have been held in non-English speaking locations and have had presentations on assessment issues in other languages as well.

Local error

1. Local errors refer to linguistic, morphological, lexical, syntactic, and orthographic errors. Grammatical errors belong to local errors. Local errors usually only affect a single element of a sentence, but do not prevent comprehension of the message.

2. A local error is an error which affects a single element of a sentence or spoken utterance that is unlikely to cause the reader or listener great difficulty in comprehension. An example might be errors in spelling or errors in using word order inflections.

Localized test/localization

A test that is designed to cater for the local needs of the test population. This may mean choosing appropriate cultural topics and making sure the processes of test design, piloting, administration and scoring reflect local needs and expectations. In more recent localization movements, this has also involved localization of language use in context to include the spread and changing shape of English in countries that use English as an official language.

Low-stakes testing

Low-stakes tests are tests which do not have a serious effect on student lives. For example, scores on class tests or class quizzes or homework assignments. Students are not severely penalized for the consequences of this type of test

M

Machine-scored test

Machine scored tests refer to tests that can be read and scored by machines. While research into machine scored testing goes back to the 1920s, it really took off in the 1950s with the advent of multiple choice tests that could be read and scored by optical scanners. The cost savings involved with such tests made it feasible to mount large-scale, multiple choice test programs from the 1960s on. Machine testing is an approach to testing and assessment that remains in use in many academic institutions around the world to this day.

Marking

Grading

Marking refers to the process of assigning numerical or otherwise systematic values to the performance of a candidate in an assessment. It is also known as grading. Marking can be done by comparing a candidate's responses to those in an answer key and calculating a result according to a raw percentage or other criteria defined for the test, such as differently-weighted questions or sections.

Mark-remark reliability

The degree to which a completed assessment will receive identical scores when marked again at another time.

Marking scale

Rating scale

1. A marking scale generally contains a series of bands/levels, accompanied with longer or shorter descriptions of what language performances are like at a certain band/level. It is usually used in the scoring process to acquaint raters with the key features of the language performance at each level. Based on an understanding of typical performances at each level, raters are expected to be able to assign test takers' performances into different levels accurately and consistently.

2. Identifies the range of values (e.g., numbers, letters, qualitative terms) that may be assigned as grades and an interpretation for different values or sub-ranges within the scale, such as what must be achieved for a pass or a high distinction.

3. A marking scale refers to a range of data measurements by numbers or nominal labels, which are assigned by the educator or adopted from an education system or country grading model, to be used to measure the value amount of a product. The content of the product is compared to the level descriptors in the range and assigned a value. For example, an essay may receive a numerical grade of 80/100 or a nominal label of B to show that the product is average in content, in accordance with the level descriptors.

Marking scheme

Marking scheme (Markscheme) refers to a marking key containing correct and acceptable answers and mark allocation for answers for a question paper. It is a specific guide for examiners in relation to a particular subject area/exam they are marking. Sometimes it may specify more than one correct answer. The purpose is to provide guidance to markers to be uniform and fair in scoring scripts.

Mastery testing

Testing that focuses on identifying a student's ability to demonstrate a minimum set of requirements that are based on an acceptable standard often determined by an end user (organization, institution).

Matching item format

Matching items are a common objectively-scored item in language testing. They are essentially an extended form of MCQs that draw on a student's ability to make connections among ideas, vocabulary and structure. Matching questions present the students with two columns of information. Students must find the matches between the two columns. Items in the left-hand column are called premises or stems, and the items in the right-hand column are called options. The premise column is typically numbered while the option column is lettered. There should be more options to choose from than premises.

Mean

In statistics, the mean is a measure of central tendency. There are several different types of mean used in statistics. The arithmetic mean, or simple mean, is the sum of sampled values divided by the number of samples.

Measurement

Measurement refers to the practice of quantifying a physical feature or an attribute. For example, a person's height (feature) can be measured using centimeters, and the same person's anxiety or introversion (attribute) can be measured by using tests that assign a score to each of the intended traits. Assigning test scores to someone's vocabulary knowledge or reading proficiency are both examples of measuring an attribute.

Measures of central tendency

The measure of central tendency is a single value that summarizes a whole set of data. This single value represents the middle or center of its distribution. The three main measures of central tendency are mean, mode, and median.

Median

The median is the middle number in a data set wherein the numbers contained in that set are in order. For example, if the data set is: 2, 3, 5, 6, 7, 7, 9. The median is 6.

Michigan English Language Assessment Battery (MELAB)

The MELAB evaluates the advanced-level English language competence of adult non-native speakers of English. The test consists of written composition, listening comprehension and questions to assess verbal ability, which include grammar, close reading, vocabulary and reading comprehension.

Mis-keying

This is the act of entering data incorrectly into an electronic device or on an Optical Mark Reader (OMR) sheet. This practice is done by both test takers and test administrators.

Modern Language Aptitude Test (MLAT)

A test of language aptitude, or one's likely success in learning a foreign language, especially listening and speaking.

Mode

The mode is the value that occurs most often in a data set. For example, in the data set: 2, 3, 5, 6, 7, 7, 9. The mode is 7.

Moderation

1. In some tertiary level programs, moderation is an exercise where teachers of a common course come together to blind-mark exam papers with sample benchmarks ranging from high, average and low-level performance. Based on the use of a standardized rubric, teachers then discuss the grades they have independently awarded to each paper and come to a consensus on a final grade to be awarded respectively. This process helps ensure that assessors make no assumptions on the criteria and that marking guidelines are clearly followed while promoting consistency.

2. The moderation process is one where the main objective is to reach agreement on test quality related issues (i.e., on the quality of items in a test or on ironing out large differences in individual marks awarded to scripts in writing tests through such activities as double marking or marking arbitration).

Multiple-choice item/question (MCI / MCQ)

1. A multiple-choice item is a question or element within a test that provides several options representing plausible correct responses. Learners are to select from these options the one(s) they believe to be correct.

2. A multiple-choice item consists of a stem and options. One of the options is the key/correct answer, the others are distractors – the aim of which is to "distract" the test taker's attention from the key. Distractors should be plausible.

Multiple-measures assessment

Multiple-measures assessment refers to the idea that it is better to use multiple ways of assessing students, rather than just one way. Using multiple types of assessments, such as multiple tests, assignments, projects, and portfolios, produces richer and more varied data about student performance, than can be obtained from just a single assessment type.

Multi-trait scoring

1. Assesses the performance of a test taker on a variety of points of interest (traits) for that particular task. For example, you might assess someone's oral presentation for grammatical accuracy, coherence of information, pronunciation, and use of supporting visuals.

2. Commonly referred to as analytic scoring, with ratings based on several important aspects or specific features of production quality (e.g., writing, speaking), as opposed to assigning a single overall (holistic) rating. From the users' perspective, one benefit of multi-trait scoring is its usefulness in generating diagnostic feedback to inform teaching and learning.

Multi-faceted Rasch analysis

Many facet Rasch analysis

Multi-faceted Rasch analysis is a data analysis method which represents the extension of the basic or dichotomous Rasch model, and which can incorporate different facets in a measurement situation in its analysis. It is typically used to examine the various issues (e.g., rater severity, rating scale) in rater-mediated assessment such as speaking and writing.

N

National Assessment of Education Progress (NAEP)

The National Assessment of Educational Progress (NAEP) is the largest nationally representative assessment of what American students know and can do in various subject areas. NAEP is a congressionally-mandated project administered by the National Center for Education Statistics (NCES), within the Institute of Education Sciences (IES) of the U.S. Department of Education. The first national administration of NAEP occurred in 1969.

Sometimes referred to as The Nation's Report Card, NAEP results are designed to provide group-level data on student achievement in various subjects. There are no results for individual students, classrooms, or schools. NAEP reports results for different demographic groups, including gender, socioeconomic status, and race/ethnicity.

Needs analysis

In the context of second language assessment, needs analysis is the intentional process of collecting data to support curriculum design. Drawing on the Council of Europe threshold level English projects, the principle of needs analysis in effective second language teaching has been applied in English for Specific Purposes (ESP) programs but is essential for all English language instruction that prepares learners for the future.

No Child Left Behind (NCLB)

No Child Left Behind was an educational policy enacted into law in the United States during the administration of George W. Bush. It included a series of requirements and guidelines for school districts and states receiving federal funds which sought to promote accountability in education. Among its measures, it consolidated in most states a system of formal high-stakes testing in elementary, middle and high school, whose results were analyzed to account for the progress and proficiency of students, including English language learners who are labelled under NCLB Title III as Limited English Proficient.

Norm-referenced testing (NRT)

NRT, as compared to Criterion-referenced Testing (CRT), is a term coined by Glaser (1963). It is a test designed to use a bell-curve to rank test takers. This bell curve has a few people scoring on the low section of the curve, with the majority of students clustering in the middle of the curve, and a few people scoring on the high end of that curve. NRTs compare a student's performance in comparison to the performance of others in areas of knowledge. This knowledge may contain something that the learner has not learned or knew about before and not aligned to predetermined standards. It measures how, and to what extent a student is performing ahead of or behind the norm. The items used in NRTs vary in difficulty and are chosen in a way that discriminates between high and low achievers.

IELTS, SAT and GRE exams are good examples of NRTs.

O

Objectively-scored test items

A kind of test which does not involve any personal judgment from the corrector(s). Test correctors therefore do not have to give personal opinions on the answers. Examples of this type of test item are multiple choice tests or true/false test items (compared to essays or open-ended answers where the corrector plays more of a role in judging the quality of students' answers).

Observed score

1. Observed score is a concept which comes from Classical Test Theory. In this approach, the score we assign to a test taker is not, in fact, his/her true score, but a score which is a combination of his/her true ability, and some error of measurement (Observed Score= True Score + Error Score). As a result, in test design, care must be taken to reduce the amount of error (error score) so that the observed score gets as close as possible to the true score.

2. Refers to a test taker's actual test score. The observed score can be conceptualized as the sum of the true score and the error of measurement. When an observed score (the actual test score) is known, confidence intervals can be placed around it to estimate the probability of obtaining a certain true score (i.e., the probable range of a test taker's true score, given the observed score). (Also see **True score**).

Open book test

An open book test is one where a student is allowed to consult references or their course materials to help them complete the test. In the context of language testing, this might be most effective in writing tests or in-class projects.

Open-ended question/item

An open-ended question / item allows for a respondent to produce a free-form answer.

Oral Proficiency Interview (OPI)

Oral Proficiency Interviews are a frequent component of foreign/second language proficiency assessments, which are meant to serve as a measure of a candidate's ability to engage in oral interaction with another person. Depending on the purpose, target, and design of the assessment, it can include a number of tasks or segments measuring a candidate's ability to engage in polite conversation, speaking at length about a topic familiar to them, providing their opinion or impressions when prompted, engaging another person in an exchange of information. The ACTFL OPI is a prominent example of this kind of assessment.

Interviews are carefully designed to be marked according to specified criteria, and follow strict protocols in their application, recording, and marking.

Optical Mark Reader (OMR)

Optical Mark Readers are devices that can identify and mark candidate's responses in a standardized test using special paper forms, in which students mark their responses by darkening a designated space, most often with a pencil. These spaces are arranged in rows, which designate the range of possible answers. Darkened spaces indicate the answer(s) selected by the candidate. This allows the OMR to calculate a student's results according to a pre-configured key.

OMR tests create issues in assessment, such as properly filling the spaces and smudges on the paper, which may create false readings.

P

Pen-and-paper test

1. A pen-and-paper test is one that is administered in a paper format and students respond by writing their answers on the test paper.

2. Pen-and-paper tests are often discussed in assessment in contrast with other testing formats available, such as computer-based tests, or other types of assessment, such as performance tasks, portfolios, or continuous assessment.

Parallel forms

Two or more forms of a test designed to measure the same construct(s), developed under the same test specifications (e.g., number of items, type of items). Ideally, parallel forms could be used interchangeably because the test takers would obtain the same score regardless of the form they take. Having parallel forms help programs and testing companies to have test security.

Parallel forms reliability

Refers to one of the primary classifications of psychometric reliability, assessed by sequentially administering parallel test forms to the same sample of test takers. The correlation between scores on the two test forms is the estimate of the parallel forms reliability. Error variance is related not to the transient error occurring between time one and time two (i.e., test-retest reliability), but instead to variance between test form one and test form two.

Pass score

Pass/fail

A pass score is a favorable result based solely on the meeting of a designated set of standards to an acceptable degree, generally as described in a rubric/marking scale or other instrument for enumerating criteria and recording the candidate's performance. Pass scores are generally awarded for assessments which are meant to be holistic in nature, and do not reflect a learner's specific performance in any particular criteria. Oftentimes, pass scores are accompanied by other descriptive documents that explain the criteria that were met in general, as well as the specific performance of the learner.

Peer assessment

The assessment by language learners ('peers') of each other's work to a standard provided by a teacher or as determined by the learners. Peer-assessment in language education is commonly used in pairs or small groups to encourage learners to develop the ability to critically review their own and others' language and to foster autonomous learning skills.

Pearson Test of English (Academic)

The Pearson Test of English is a four-skill English language proficiency assessment developed to match both the Common European Framework of Reference scale, as well as Pearson's own Global Scale of English, which is purportedly more granular in its levels and descriptors.

Percentile score

Percentile rank

Percentile rank, also known as percentile score, indicates the position of a score within a given set of scores, as compared to all other scores in the set. One way to calculate the percentile rank is to use the formula: (F/n) 100, where *F* denotes cumulative frequency and *n* is the total number of test takers.

Performance testing

A kind of test in which the test taker's performance (not the competence) is evaluated. This is to test how well a test taker performs. This is usually associated with a form of oral speaking presentation or reading aloud, although other productive tests like writing tests can also be included in this criterion. Performance may include pronunciation, intonation, clarity of ideas, gestures/ facial expressions in speaking, or spelling, ideas, organization, coherence in writing.

Preliminary English Test (PET)

PET is a Cambridge English Language Assessment examination for General English assessment, administered by the British Council. This test assesses communication in English in practical, everyday situations. It claims to provide a good foundation to study for a professional English qualification. The level of qualification is Intermediate, equal to B1 on the CEFR.

Pilot testing

Field testing

Pilot testing is an evaluation of a test group or groups in order to verify and compare the instrument's validity for determining actual ability or knowledge. A pilot test's results can be compared to candidate results in other similar assessments or measures with established validity (which includes measures of reliability) to see how closely they match. They can also be used to obtain data on candidates' experience and the security and effectiveness of procedures.

Pilot tests are generally completed during the process of developing an assessment, and they should ideally inform potential changes to the test's format, question design, and application and marking procedures before it can be used more widely.

Placement testing

Placement testing is most likely to occur when the level of a student is unknown and an institution must determine which course level the student should be enrolled in. A test is likely to have a series of tasks/questions which correspond to various language levels. Traditionally, such tests start with low-level test items and become gradually more difficult to correspond to course sequence.

Plagiarism

Plagiarism is an act which breaches academic honesty whereby a person takes someone else's work, thoughts, expressions, or ideas and presents them as one's own original work without crediting the source.

Point biserial

Point biserial correlation is similar to the correlation coefficient. This is used to measure the strength and direction of the relationship between two variables where one is dichotomous. We typically look for a positive point biserial correlation with item analyses.

Pop quiz

A pop quiz is the unannounced or otherwise spontaneous application of a test to a group of students, generally on content or skills which are currently being studied. Pop quizzes are used by some teachers to promote continuous study, as well as a type of informal continuous assessment. Due to their stress-inducing nature and sometimes insufficient preparation, student results in these quizzes may not accurately reflect their understanding or mastery. As part of assessment for learning, however, pop-quizzes can be used to help students review their own performance to become conscious of the current state of their learning.

Portfolio

A set of student classwork, homework, and language samples that is collected, typically by the student with teacher guidance, to demonstrate a student's progress throughout the course of a class. Portfolios are useful in helping students document their own progress and in helping teachers to assess student progress.

Portfolio assessment

A portfolio assessment is a collection of student's academic work which is placed in a folder, either paper-based or digital. It may be a presentation of student's best work or a collection of all work to show the progress of course goals over time. It is one form of alternative assessment and is student-centered. The teacher needs to spell out the criteria for portfolio assessment in advance.

Population validity

Refers to a type of external validity, specifically the extent to which the sample of test takers used and, by extension, the results of the study can be extrapolated to a target test taker population or else can be considered generalizable to a defined population of test-takers.

Post-test

Pre-test

A measurement of the learning received during a course by comparing what students knew before starting the course through a pre-test and what developments have been achieved after the class experience through a post-test. More specifically, the tests indicate how the students are learning in the course. The data may be used to target students requiring extra help, may identify gaps in course design that need to be addressed, and teaching and learning methods that need to be changed or developed.

Power test

Refers to a kind of test which attempts to assess the different levels of ability of the candidates. Candidates are provided adequate time to complete all the items of the test. However, it may still not be possible for them to complete all the items on the test as the items are graded and become increasingly difficult.

Practicality

1. Practicality can refer to a criterion for selecting specific types of assessments; concretely to the teacher or institution's ability to administer the assessment within the constraints of time, space, staffing, resources, government/institutional policies, and candidates or parents' own preferences, among others. For example, formal oral interviews can be omitted as part of a course's evaluation scheme due to practicality, favoring instead portfolio-based speaking assessments.

2. Practicality is an integral part of the overall quality of language assessments. Practicality refers to resources available to developers and users of language assessments in the processes of developing, administering, scoring, and using their assessments. Resources include human resources, material and financial resources, and the time set aside for assessment activities. The degree of practicality can be measured by the difference between the resources that will be required in the development and use of an assessment and the resources that will be available for these activities.

Predictive validity

The extent to which a test predicts performance in an external situation or future performance (e.g., performance in a job or academic setting).

Presentation

Presentation can refer to a specific type of class or alternative assessment task, where students present orally on a given topic according to the objectives and criteria set out by the teacher, for example, as the final stage of a project-based unit or lesson.

It can also refer to a stage in the Presentation–Practice–Produce lesson scheme, in which information and target language input are provided, which are later followed by structured practice and later an oral or written product.

A third definition refers to how information is provided to a specific audience, such as instructions or prompts in a test.

Pre-test

Post-test

A measurement of the learning received during a course by comparing what students knew before starting the course through a pre-test and what developments have been achieved after the class experience through a post-test. More specifically, the tests indicate how the students are learning in the course. The data may be used to target students requiring extra help, may identify gaps in course design that need to be addressed, and teaching and learning methods that need to be changed or developed.

Primary trait scoring

Primary trait analysis is a technique for identifying particular aspects of student language for feedback or assessment. After specific traits on which to focus are specified, criteria are established for each trait. In primary trait analysis, secondary traits are ignored, which allows teachers and students to focus their attention.

Proctor/proctoring (US)

Invigilator/Invigilating (UK)

A proctor or invigilator is the person responsible for observing the administration of an assessment instrument, such as a pen-and-paper test. In a formal proficiency assessment, an invigilator/proctor is not necessarily the test's administrator, and they generally must be people who are not involved with the testing group on a teaching or personal capacity. An invigilator or proctor must be trained on the testing procedures, and can report incidents and instances of breach of protocol to an administrator.

Testing procedures often make recommendations or requirements on the number and ideal profile of proctors/invigilators.

Progress test

A progress test is a type of test meant to be applied after instruction or sufficient practice on specific content or a given skill within a course. They are often created in advance and scheduled at specific points in a course's timetable.

Program evaluation

A program evaluation refers to an assessment of an instructional program, such as a language school or a degree-granting unit within an academic institution. Evaluations are usually conducted by reviewers from outside the program who review evidence provided by the program in relation to agreed-upon standards or objectives. Often programs will engage in a self-assessment as a way of creating and organizing evidence for the reviewers. Reviewers will typically then conduct a "site visit" to verify the evidence and conduct additional investigations as they see fit. Program evaluations may be used for the purposes of accreditation, stakeholder accountability, or routine review.

Project

Projects are an example of an alternative assessment method. When doing projects, students work together to achieve a common goal, purpose or concrete outcome (i.e., produce a brochure, write a report, make a display or create an article etc.) Educators believe they are a very effective means of mixing classroom practice with assessment. While some projects can last the length of the class, others may be accomplished within days, weeks, or months. Some of the most important considerations to keep in mind when using projects as alternative assessments are that they must be closely tied to curricular objectives and outcomes; they should have clear language and academic goals; and they should have unique qualities that excite the students to higher levels of thinking. Projects can come in many forms, and students can work as individuals or in groups. They can be based on reading, writing, listening, speaking, or integrate all four language skills.

Prompt

In writing or speaking, as well as any problem-oriented content area, a prompt refers to the bit of information provided for students to respond to in an assessment or in class tasks. It often provides background information, specific instructions and marking criteria, as well as any other relevant information the students can need to successfully attempt the task.

An insufficiently clear or ambiguous prompt often results in inadequate responses, so care must be applied in their design.

Psychometrics

Psychometrics is the field of study, in which psychological measurements are done to find individual differences among the learners who possess, various levels of knowledge, abilities, attitudes and personality traits.

Q

Qualitative test data

Refers to data not in numerical form (e.g., questionnaire data from open-ended questions and interview data). Qualitative data may also be quantified or transformed into numerical form.

Quantitative test data

Refers to data in numerical form, obtained through frequencies, counts, and measurement, such as test scores, rating scales, or fixed responses from questionnaires.

Quiz

A quiz is a brief written test, generally part of continuous assessment, and which is meant to provide teachers and learners with a sense of their progress towards stated content and skill mastery goals. Quizzes are generally teacher-made, individual, pen-and-paper and time-constrained, and can be scheduled or spontaneous within the course's timetable (also see **Pop quiz**).

R

Rational deletion cloze

A cloze test is one where a decision has been made to delete particular words or parts of speech in order to measure the ability of a test taker to fill in those particular items accurately. For example, one might delete vocabulary being assessed, or definite and indefinite articles, or present tense verbs…or a combination of these in order to assess all of these features.

Rater

An individual who is tasked with making an evaluation of a test taker's spoken or written language performance, and which normally involves using some form of agreed upon and defined rating scale.

Rater training

The training of raters for their work of evaluating performance to minimize potential rater bias and enhance rater reliability.

Rating scale (UK)

Rubric (USA)

A rubric or a rating scale is a numerical scale used to evaluate a test taker's spoken or written performance. Modern approaches to rating scale development emphasize including a descriptor that defines performance at each level on the scale. There are two types: holistic and analytic.

Holistic or impressionistic marking is one in which a level on a numerical scale is allocated according to the rater's impression, without a set of descriptors to constrain or define that impression.

Analytic rubrics include several categories, such as organization, content, language, and others, and a detailed description is provided for each category and each band.

Analytic scales are considered to offer greater reliability by offering more measurement points, and offer greater depth of feedback useful for diagnostic uses. However, they have a trade-off in terms of practicality as holistic scales are generally quicker and more efficient for raters to use. Modern holistic rating scale approaches often include a hybrid in which guided descriptors will cover several categories but are included at each level on the scale, rather than separating them out into different scales for each category.

Raw score

A raw score is the total score based on the number of correct answers.

Readability

1. A measurement of the appropriacy of a text's level of difficulty for a reader; it is often determined by the complexity of the vocabulary and syntax.

2. Readability refers to the difficulty level of a reading text which includes the semantic and syntactic structures and presentation of the text (i.e., legibility). Readability is dependent on characteristics of the readers and the texts.

3. Readability is a measure of the ease with which a reader can understand a written text. It depends on a number of factors including vocabulary complexity, sentence length, text coherence and organization, etc. Various formulas have been developed, among which the Flesch-Kincaid and Flesch-Kincaid Grade Level formulas are among the most popular. Coh-Metrix is an automated tool which provides a range of measurements that have often been used in language research for providing more detailed information on text readability and text cohesion.

Recall item

A test taker is asked to retrieve from memory previous information presented and articulate it to the examiner. The information that the test taker is able to retrieve and articulate is considered "recall items".

Recognition item

A test taker is asked to indicate in some form or another whether an item tested is familiar to them. For example, the test taker may need to select the appropriate tense from a list of conjugated options. These are items that the test taker may "recognize" but may or may not be able to produce independently.

Reliability

Reliability refers to the consistency of scores or test results. The questions to be raised are: 1) Are test results dependable and trustworthy? 2) If a student took the same test the following day would the test results be the same? Reliability is a fundamental criterion of a good test. It is regarded as an attribute of validity. Reliability for selected response items is measured through indices such as Cronbach's Alpha.

For spoken or written performances marked by human raters, two types of reliability are crucial for arriving at reliable scores: 1) Inter-rater reliability: this refers to the consistency of marking between two or more raters. 2) Intra-rater reliability refers to the consistency of marking within the marker.

Response option

In selected-response type items, a response option refers to all the possible answer alternatives that the test taker is presented with.

Retired test

1. Refers to a test form no longer in use for its original purpose. Retired tests are sometimes published as practice material or used for research purposes in cases where test security is not a primary concern.

2. An evaluation or assessment that has been withdrawn from a systemic bank of tests. Sometimes retired tests can be used as practice test materials.

Rubric (USA)

Instruction (UK)

A rubric is a grid usually used to assess someone's speaking or writing performance. A rubric has two axes: one with scores all along and the other with some criteria (categories) against which the test taker's performance will be assessed. What appears inside each cell is called the descriptor. Rubrics can be used for criterion-referenced assessment as well as norm-referenced ones. There are different types of rubrics: analytic rubric, holistic rubric, task-dependent and task-independent rubrics. In research and assessment, there are also some rubric-like instruments such as checklists and numerical rating scales.

S

Safe Assign

Plagiarism detection and prevention software that compares student submitted work against a database of previously submitted essays from within the institution, essays submitted to partner institutions, the Internet and journal articles.

Security

Refers to measures taken by test developers and/or administrators to ensure that no test-taker gains advantage by having prior knowledge of a test's content. Measures may include, for example, developing a test bank and using new or different versions of a test, safeguarding confidentiality in test development, and using a validation process.

Selected response question

Selected response questions are questions where the correct answer is pre-determined and the test taker must choose or select the response from a list of given options. Because the answers are pre-determined, these types of questions are often referred to as "objective" questions. Examples of this type of test question include true/false, multiple choice and matching.

Self-assessment

1. Self-assessment: One kind of alternative assessment which is learner focused. Learners grade or assess themselves based on specific criteria or descriptors or a self-assessment questionnaire which teachers have provided. This type of assessment can be useful for involving learners in evaluating their own strengths and weaknesses and achievement of their learning goals and objectives.

2. Self-assessment refers to the ability of language learners to assess their own performance to identify their strengths and weaknesses in the learning process. Self-assessment has been proven useful for promoting learner autonomy and is often used for formative purposes.

Sequencing task

A sequencing task is a testing item or class activity in which learners are provided discrete elements of information in disarray (for example, the events in a story, a list of functional words, concepts that appear in a text). Learners must then produce the right order of the sequence by finding the correct answer in a multiple-choice question, completing a chart, etc.

Short answer questions (SAQs)

Short answer questions are one type of test item mostly used in the testing of reading and listening comprehension. As its name suggests, a short answer question requires test takers to provide brief answers, often in words or phrases. Test takers are often required to formulate their answers in their own words. This type of test item is subjectively scored, and a comprehensive marking scheme (see **Rating scale**) should be formulated to ensure reliability. The marking scheme should include acceptable alternative answers, as well as detailed instructions about the criteria of judging grammatical and spelling errors.

Socio-cognitive validation model

Refers to a test validation approach attributed to Weir (2005) and O'Sullivan and Weir (2011) that combines social, cognitive, and evaluative dimensions of language use and links those dimensions to the context and consequences of test use. The approach examines multiple components (e.g., cognitive aspects of validity, contextual aspects of validity, scoring aspects of validity, criterion-related aspects of validity, and consequential aspects of validity) and also their interactions for evidence to justify the uses of and interpretations of a test.

Skill contamination

When assessment of one skill is influenced by another, thereby possibly affecting the performance/ assessment of the desired skill. For example, a misunderstanding of a question being asked might impact the oral performance of a candidate and "contaminate" the measurement of the examinee's speaking proficiency. However, integrated tasks which require a test taker to integrate performance across skills have also received more attention recently (see **Integrative testing**).

Specifications

Specs; test specifications

Specifications are said to be the blueprint of a test. It is a detailed description of:

a) the purpose of the test (progress, achievement, placement, other)
b) the program (duration, number of sessions per week, duration of each session, textbook/s, other instructional material, assessment tools if applicable)
c) the test takers/learners (age, prior experience, exposure to other courses, native language, other)
d) test description (skills assessed, number and type of techniques used to measure each skill, the subskills that each item tests, the weight of each item, assistance that test takers will receive and channels that may be used with some prompts or items, time allotted to each skill)
e) criterial levels of performance, setting cut-off/pass points
f) rubrics for productive skills assessed
g) answer keys for objectively-scored parts of the test
h) test administration guide.

The list looks long and overwhelming, however, in reality it is not. Any teacher has all this information in mind. All he/she has to do is put it together on paper and use and modify it for each test designed.

Test specifications help to have a more constructive approach towards test development. They also help to construct tests that meet the requirements of the curriculum, and, if appropriately written, can help reveal valuable information about the course and levels of achievement of course objectives.

Split half reliability

One of the methods of establishing test reliability is splitting the test into two parts and measuring the correlation between the scores on the two parts. It is important to remember, though, that the test can NOT be split into first and second halves. To make sure that the test is divided into two parts with reasonably equal difficulty, we need to put the odd numbered questions in one part and the even numbered questions in the other. This kind of division also takes care of an equal distribution of the skills and subskills tested.

For example, if the test consists of 10 reading items, 10 listening items, and 20 grammar items, splitting the test in this way will give an opportunity to have answers to all the skills and/or subskills tested and thus make it possible to compare the results of the two tests to establish comparability and reliability.

Stakeholder

A stakeholder in education is a person whose interests should be taken into consideration when designing a course, developing a curriculum or a syllabus, choosing teaching methods or techniques, planning lessons and/or designing assessment. Primary stakeholders are the learner, the teacher, the parent, and the administrator (school and ministry). However, there are many more potential stakeholders and recent approaches to testing validation, for example, the socio-cognitive approach (see **Socio-cognitive validation model**) have emphasized the importance of taking a wider view of stakeholders.

Standard deviation

The standard deviation is a measure used to show the amount of variation in a set of data. A low standard deviation indicates that the data points tend to be close to the mean, while a high standard deviation indicates data points which tend to be spread out over a wider range of values.

Standard error of measurement (SEM)

SEM is a measure of score accuracy, whereas reliability is a measure of the consistency of test scores. SEM provides an estimate of the accuracy of the scores. The SEM reflects the area or probable limits around the test taker's observed score within which the hypothetical "true score" for a test taker can be expected to fall.

Standardized test

Standardized tests are tests designed so that the questions, administration, marking, and interpretation of results can be conducted in a standard and systematic manner. Test takers sitting a standardized test are given the same sets of questions, answer sheets and time in which to complete the test. They also receive the same instructions in writing and, where appropriate, oral form from invigilators. Markers have a pre-specified set of correct answers and a systematic process for marking and calculating, interpreting and reporting test scores is in place. Examples of standardized tests include the IELTS and TOEFL. An important part of standardized tests is to have well-constructed test specifications (see **Test specifications**).

Standards

1. Standards are descriptions of performance and content-mastery that are meant to provide targets for instruction and reference for assessment. Standards are often linked to specific grades, levels, or assessment instruments. In their writing, standards must clearly reference skills, context, assessment criteria, measures, and degree of content mastery. They also sometimes contain behavioral and attitudinal aspects.

In recent years, a movement for standards-based education has made a deep impact in curriculum and instruction in many countries, leading to broad implications for test design, textbook content, and teacher preparation. Examples of standards include the Common Core State Standards in the United States. In language education, the widespread adoption of the CEFR and the scales within it (see **Common European Framework Of Reference**) has led to focus on explicit standards of performance for language learning.

2. Levels of performance required (proficiency standard) or a set of guidelines or principles used to guide what language testers do (e.g., a professional code of practice).

Student-designed test

A test with content that has been significantly influenced in one or more parameters by direct input from students.

Subjectively-scored test items

An item on an assessment which is graded based on subjective requirements of the grader rather than an objective key (see **Selected response** and **Grading scale**).

Summative assessment

Summative assessment refers to the processes and instruments that provide a general and final assessment of student's learning within a given course or learning unit. Due to this, summative assessment is also formal in nature, and can include instruments that measure broadly the skills and content areas developed in a course; for example, course tests, final projects, and portfolios. It is often referred to in contrast with **continuous assessment**, which is assessment produced throughout the course.

T

Take-home test

Assessment tasks which students have the opportunity to complete in their own time, without direct supervision by the teacher or an invigilator. It is advisable that take-home tests contain tasks that require applying knowledge to solve problems. For language assessment purposes, tests measuring written production (e.g. short answers, essays, letters, and others) may be assigned.

An interesting idea could be using technology in take-home tests not only for receptive but also for productive skills. For example, to reduce the time for speaking tests, the teacher could assign the students to video-record their answers to short questions or reflections on their smart phones and send the video-recorded speech to the teacher.

Target language use domain (TLU)

Target language use (TLU) domain refers to goal-oriented contexts of language use in which language users perform communicative tasks. The TLU domain is the context of actual language use in which a learner will be using the target language outside the test itself. TLU domain plays an important role in construct conceptualization and test score interpretation. An important part of language test validation is demonstrating that test tasks simulate what learners do in TLU domains. Test users, based on performance consistencies, make inferences of test takers' language ability and generalize it to TLU domains outside of the testing situation.

Task

Classroom task

Task is the central premise of task-based language teaching (TBLT) and Task-based language assessment (TBLA). TBLT is an educational framework and an approach for the theory and practice of second/foreign language (L2) learning and teaching, and a teaching methodology in which classroom tasks constitute the main focus of instruction (Richards & Schmidt, 2010). TBLA is a framework for language testing that takes the task as the fundamental unit for assessment and testing. For both TBLT and TBLA, a classroom task is defined as an activity that: (1) is goal-oriented; (2) is content-focused; (3) has a real outcome; and (4) reflects real-life language use and language need (Shehadeh, 2005; 2012). The syllabus in TBLT and TBLA is organized around activities and tasks rather than in terms of grammar or vocabulary.

Task-based language assessment (TBLA)

Task-based language assessment (TBLA) is a framework for language testing that takes the task as the fundamental unit for assessment and testing. It is a formative assessment that emphasizes assessment *for* learning rather than assessment *of* learning. That is, it is an assessment undertaken as part of an instructional course of study or program for the purpose of improving learning and teaching. Long and Norris (2000, p. 600), for instance, state that "genuinely task-based language assessment takes the task itself as the fundamental unit of analysis, motivating item selection, test instrument construction and the rating of task performance". Similarly, Weaver (2012, p. 287) explains that "[a]t its core, task-based language assessment (TBLA) involves evaluating the degree to which language learners can use their L2 to accomplish given tasks".

Teacher-made test

A test which is designed by a teacher to measure a specific lesson's objectives related to what is being taught in the classroom. This test can be written in more or less formal or informal ways, depending on the purpose of what the teacher is trying to achieve.

Test

1. A set of tasks or activities intended to elicit samples of performance which can be marked or evaluated to provide feedback on a test taker's ability or knowledge. In the context of classroom-based assessment, these might be intended to find out whether the lesson taught is understood by all, some, or few. Sometimes a test can be formal and/or informal. Tests can also be done prior, during, and after a lesson, depending on the lesson's objectives.

2. Refers to the activity of measuring samples of performance elicited by a test from a test taker. Based on a purpose or outcome criteria, the process demonstrates and reveals the person's degree of content or skill acquisition; thus, inferring how much information or what aptitude (ability) is acquired.

Test of English for Educational Purposes

TEEP (University of Reading – originally TAEP by Cyril Weir 1983-first integrated skills EAP test)

Refers to a test of proficiency specifically in English for academic purposes (including reading, listening, writing and speaking sections) for test takers whose first language is not English and who intend to pursue studies in the UK. The test was developed by the International Study and Language Institute at the University of Reading.

Test anxiety

Test anxiety is a type of performance anxiety in which the excessive stress, worry and nervousness can hinder a person's ability to perform well in a test or exam. This can be caused by a fear of failure, lack of preparation or poor test history.

Test battery

A group of tests used together for a specific measurement purpose. For example, in language testing, separate tests of reading, listening, speaking and writing are used to measure overall language proficiency.

Test development process

The process of development may vary across programs, but the general procedures are usually described as specifying the purpose of the test and inferences to be drawn; developing frameworks describing the knowledge and skills to be tested; building test specifications; creating potential test items and scoring rubrics; reviewing and piloting test items; evaluating the quality of items; and revising and refining test items. The process typically follows three essential phases (planning, design, and try-out).

In the planning phase, questions are asked that enable the developer to define the language ability to be tested, such as the characteristics of the target test takers, the purpose of the test (e.g., diagnostic, admission, exit), standards for the proposed purpose, and the test in relation to the curriculum or learning objectives.

The design phase begins with building test specifications that describe what a test is designed to measure, the skills and content areas to be tested, and the details of technical implementation (e.g., test purpose, target population, test content, test format, test duration, rating scales, and scoring method). The test specifications provide a blueprint for test writers and are the basis for creating alternative versions of the test. In the try-out phase, the draft specifications are tested and improvements made by piloting the test with target test takers and analyzing the piloting data.

Clear and explicit descriptions of test development in the literature often reference large-scale, standardized tests. This can make the process seem daunting to teachers, but many of the same procedures are equally applicable to collaboration between colleagues.

Test form

Refers to one of multiple versions of a test which are considered to be interchangeable. Multiple forms of a test are often constructed for test security purposes. For example, high-stakes standardized tests (for entrance or admissions) may require new forms for use each round/year. Test forms need to be constructed using explicit test specifications to ensure comparability, as it is assumed that different forms of a test are interchangeable in terms of difficulty and construct.

Test taker

Examinee; candidate

A test taker is a person who takes a certain test to display his/her knowledge, and skills regarding certain topics.

Testing

Testing is the process of seeing how we match up to a standard or standards. The abilities, knowledge or skills targeted by the test are known, the rubric is known, the levels are known. We have a set of clear, transparent testing guidelines to follow to see how we fit in with others taking the same test, or against a set of given criteria.

Test–retest reliability

A measure of test reliability by administering the same test to the same group of test-takers on two occasions. A correlation between the two sets of test scores provides a measure of the consistency of test scores.

Test-taking strategies

The high-stakes nature of much testing has led to an increased focus on test-taking strategies. These strategies focus on how to successfully navigate a test to produce the best outcome. One assumption is that a test taker who employs these strategies will perform better than another test taker of the same level and may even perform better than a test taker with a higher level. Examples of strategies may include: how to interpret essay questions; how to plan writing tasks; how to identify distractors in multiple choice questions; how to find relevant information in a text; and how to plan one's time. The last example shows that not all strategies need to be language specific.

Test wiseness

The art or skill of locating or guessing the right answer from clues in the test item. The clue or hint could be in the stem or the options or alternatives provided. For example, a link between a phrase or a word in the stem and options can be spotted easily by candidates. A longer response than other responses can be a giveaway for the correct answer or could be eliminated as an option in an MCQ. Teachers need to be cautious when they are writing test items, especially when writing options for MCQs.

Text-based prompt

Text-based prompts refer to written prompts which provide test takers with source texts. Such prompts provide content for test takers and involve reading and task completion. Tasks using text-based prompts as input materials may be better able to simulate real-life academic assignments and thus improve both the authenticity and validity of the test tasks. However, this needs to be undertaken in a principled way to avoid, unintentionally conflating different abilities (for example reading and writing) (see **Integrative testing** and **Skill contamination**).

Test of English as a Foreign Language (TOEFL)

The TOEFL is a standardized test of English language ability made by Educational Testing Services (ETS). Many non-native speakers take this test when applying to English-speaking universities, particularly in the United States. The test was designed to test the academic English proficiency considered necessary to study in an English medium institution. TOEFL and IELTS are two major English language tests used throughout the world by academic and professional institutions for this purpose.

Test of English for International Communication (TOEIC)

The TOEIC is a standardized test of English language ability made by Educational Testing Services (ETS). This test is designed to assess the English necessary for everyday international business. Some companies, particularly in Korea and Japan, require a high TOEIC score for career advancement.

Topic familiarity

An issue in language testing relating to learners' familiarity with particular topics that are used in input texts or task prompts in a test. Learner familiarity with topics may improve their performance in such tasks, while lack thereof may decrease it. The issue also touches on concerns about the cultural sensitivity of assessments, and points to potential issues with context validity.

Topic restriction

Topic restriction is often imposed by testing bodies or at local classroom levels where students have to answer a given topic or are restricted to choosing to answer between a choice of narrowly selected topics. This topic restriction allows responses to be more easily graded and compared in assessment research but may also be used with the intention of reducing topic bias and any possible unfair advantage by allowing students to choose their own topics.

Transparency

Refers to making information about the assessment process and criteria accessible and available to users, including what the measures are. For example, the use of rubrics is a way of providing transparency in assessment.

Trait

1. Characteristic of an examinee's performance that you wish to measure.

2. Traits are enduring characteristics of an individual that underlie their behaviors. In language testing, the performance of language users on a test is evaluated to draw inferences about their language ability based on particular traits shown in their behaviors.

Trait scoring

1. Assessment of a test taker's performance by assessing particular characteristics of that performance. One can assess a single characteristic (primary trait) or multiple characteristics (multi-trait) of a particular performance.

2. Refers to a scoring method that enables instructors/assessors to focus on specific traits required by the communicative context implicit in a task. For example, in writing, traits can be in relation to different levels of the written discourse (e.g., word choice, organization, audience awareness) that characterize the various standards of competence central to successful task completion.

True/false question

T/F

True/False questions are second only to MCQs in frequency of use in professionally produced tests and perhaps one of the most popular formats for teacher-produced tests. They are a specialized form of the MCQ format in which there are only two possible alternatives and students must classify their answers into one of two response categories. The most common response categories are: True/false, yes/no, correct/incorrect or right/wrong. True/False questions are typically written as statements and the students' task is to decide whether the statements are true or false. To decrease the guessing factor of T/F questions, test developers often add a third response category: Not Given or Not Enough Information.

True score

True score is a hypothetical construct referring to a test taker's score if no error occurs in measurement. The concept emphasizes the fact that some error is involved in any type of measurement. It is possible to estimate an individual test taker's true score in relation to the observed score (i.e., the actual test score), which is the true score plus some measurement error (systematic) and some random error (unsystematic). Thus, the true score is a hypothetical score that is supposed to reflect a test taker's true ability. (See also **Observed score**).

Turnitin

Turnitin is an online plagiarism detector whereby students submit their essays to the website which checks the originality of their work against a vast electronic database. It helps students avoid any academic misconduct as it detects similarities by sharing an Originality Report with the student.

Test of Written English (TWE)

1. The TWE is the essay component of the TOEFL, designed for English-as-an-additional-language writers to demonstrate their ability to express ideas in standard/appropriate written English in response to an assigned topic. Each essay is scored by two or more qualified raters according to lexical and syntactic standards of English, and according to the effectiveness of the test taker's organization, development, and support of ideas with evidence or examples in writing. The TWE is only used for the institutional TOEFL nowadays.

2. The TWE is the essay component of the institutional version of the Test of English as a Foreign Language (TOEFL).

U

University of Cambridge Local Examinations Syndicate (UCLES)

1. UCLES is the acronym for University of Cambridge Local Examinations Syndicate, also referred to now as Cambridge Assessment. It is one of the world's largest assessment agencies.

2. The University of Cambridge Local Examinations Syndicate is a department of the University of Cambridge which is responsible for the designing, piloting, administration, scoring and overall continuous improvement of Cambridge examinations. Examples of English language exams for which UCLES is responsible include Cambridge English exams: FCE (First Certificate in English or simply Cambridge First), CAE (Cambridge Advanced) as well as young learner exams such as Starters, Movers and Flyers which represent different age and proficiency attainment levels.

Usefulness

In 1996, Bachman and Palmer developed a framework of test usefulness for evaluating a test which contained six categories: reliability, construct validity, authenticity, interactiveness, impact and practicality. The usefulness of a test depended on achieving the right balance of these characteristics for the intended use of the test. The inclusion of practicality as an explicit evaluation criteria to the features to be considered when evaluating a test or assessment was an important addition often overlooked in theoretical discussions of validity and validation.

V

Validity

Validity refers to 'the degree to which' or 'the accuracy with which' an assessment measures what it is supposed to measure. Since the 1980s there has been a general consensus that it is more appropriate to talk about the validity of the uses and interpretations of a test, rather than the test itself. A test could be valid for some uses for some test takers, but not for others. Various models of validity have been developed and the argument-based approach to validation and the socio-cognitive model for language test development and validation have been very influential (see **Socio-cognitive validation model**).

Validation

The process of establishing the validity of a test by gathering and evaluating the evidence for its validity and reliability.

Variance

As a means of calculating *dispersion*, or otherwise known as *variability*, *variance* tells us how far from the mean a set of scores are spread. Variance is usually employed for assessment purposes in conjunction with other statistical procedures, such as *standard deviation*.

W

Washback

Backwash

The washback effect refers to either positive or negative effects that a test can have on students or teachers' actions. For instance, the exam effects on the curriculum, the syllabus and the coursebooks, etc. (see **Impact**).

Z

Z score

Used to determine the position of a score relative to a pool of scores. A raw score is converted to a z score to calculate how many standard deviations a score is from the mean score (for a set of scores).

List of contributors

Nadia Abdallah
University of Exeter, UK
E: Na334@exeter.ac.uk

Sufian Abu-Rmaileh
United Arab Emirates University, UAE
E: sufian12000@yahoo.com

Lubna Adel
The British University in Egypt (BUE), Egypt
E: lubna03@gmail.com, lubna.adel@Bue.edu.eg

Ramin Akbari
Tarbiat Modares University, Iran
E: Akbari_ram@yahoo.com

Mark Algren
University of Missouri (Columbia), USA
E: algrenm@missouri.edu

Naziha Ali
Lahore University of Management & Sciences, Pakistan
E: nazihaali2005@yahoo.co.uk

Rosa Aronson
TESOL International Association, USA
E: raronson@tesol.org

Douglas Black
Mahidol University International College, Thailand
E: Dougblack85@gmail.com

Isaac Pérez Bolado
University of Dayton Publishing, Mexico
E: iperez@udaytonpublishing.com

James Buckingham
Sultan Qaboos University, Oman
E: auh.jimb@gmail.com

Chloe Burridge
FXPlus, Falmouth and Exeter University,
Penryn, UK
E: chloeburridge@gmail.com

Christel Broady
Georgetown College, USA
E: christel.broady@gmail.com

Christine Canning-Wilson
Boston Manhattan Group, Inc., USA
E: ccanningwilson@yahoo.com

Alissa Carter
Dubai Men's College, UAE
E: acarter@hct.ac.ae

Mojtaba Chaichi
Iran
E: humanbeing1st@gmail.com

Ying Chen
Ocean University of China, China
E: jennych2008@126.com

Christine Coombe
Dubai Men's College, UAE
E: ccoombe@hct.ac.ae

Peter Davidson
Zayed University, UAE
E: peter.davidson@zu.ac.ae

C.J. Denman
Sultan Qaboos University, Oman
E: denman@squ.edu.om

Aymen Elsheikh
Texas A & M University in Qatar, Qatar
E: elsheikhaymen@hotmail.com

Liz England
Liz England & Associates, USA
E: LizEnglandAssociates@gmail.com

Rubina Gasparyan
American University of Armenia, Yerevan
E: rgaspari@aua.am

Alsu Gilmetdinova
Kazan National Research Technical University, Russia
E: amgilmetdinova@kai.ru

Melanie Gobert
Higher Colleges of Technology, UAE
E: mgobert@hct.ac.ae

Xiaoxian Guan
East China Normal University, China
E: xxguan1@126.com

Maria Nelly Gutierrez Arvizu
Universidad de Sonora, Mexico
E: mng33@nau.edu

Doaa Hamam
Dubai Men's College, UAE
E: dhamam@hct.ac.ae

Christopher Hastings
Southwest Tennessee Community College, USA
E: christopherhastings@gmail.com

Kristof Hegedus
Euroexam International, UK
E: kristof@euroexam.org

Lana Hiasat
Higher Colleges of Technology, UAE
E: lhiasat@hct.ac.ae

Li-Shih Huang
University of Victoria, Canada
E: lshuang@uvic.ca

Ghada A. Ibrahim
Cairo University, Egypt
E: gibrahim@aucegypt.edu

Daniel R. Isbell
Michigan State University, USA
E: isbellda@msu.edu

Mary Jennifer J
National Institute of Technology, India
E: maryjennifer17@gmail.com

Wei Jie
Shanghai University of International Business and Economics, China
E: neoweism@aliyun.com

Yan Jin
Shanghai Jiao Tong University, China
E: yjin@sjtu.edu.cn

Jason Jinsong Fan
Fudan University, China
E: jinsongfan@fudan.edu.cn

Diana Johnston
Dubai Men's College, UAE
E: djohnston@hct.ac.ae

Renee Jourdenais
Middlebury Institute of International Studies at Monterey, USA
E: rjourden@miis.edu

Rubina Khan
University of Dhaka, Bangladesh
E: rkhan@agni.com

Slim Khemakhem
Higher Colleges of Technology, UAE
E: slim.khemakhem@hct.ac.ae

Mick King
Community College of Qatar, Qatar
E: micjak66@gmail.com

Georgios Kormpas
INTERLINK at YU, Saudi Arabia
E: georgekormpas@gmail.com

Antony John Kunnan
University of Macao, China
E: akunnan@umac.mo

Betty Lanteigne
American University of Sharjah, UAE
E: blanteigne@aus.edu

Stéphane Lavie
Sultan Qaboos University, Oman
E: waslavie@gmail.com

Listyani
Satya Wacana Christian University, Indonesia
E: listyani@staff.uksw.edu

Chang Liu
University of Electronic Science and Technology of China, China
E: liuchang1977@uestc.edu.cn

Steven G.B. MacWhinnie
Aomori Chuo Gakuin University, Japan
E: smacwhinnie@gmail.com

Lee McCallum
University of Exeter, UK
E: lm489@exeter.ac.uk

Barry O'Sullivan
British Council, UK
E: Barry.osullivan@britishcouncil.org

Sriganesh R
National Institute of Technology, India
E: mail@sriganesh.biz

Arifa Rahman
University of Dhaka, Bangladesh
E: arifa73@yahoo.com

Nick Rea
Chartered Management Institute, UK
E: Nick.rea@managers.org.uk

Hayo Reinders
Unitec, New Zealand and Anaheim University, USA
E: info@innovationinteaching.org

Dudley Reynolds
Carnegie Mellon University in Qatar, Qatar
E: dreynolds@cmu.edu

Christine Sabieh
Notre Dame University, Lebanon
E: csabieh@ndu.edu.lb

Pushpa Sadhwani
Higher Colleges of Technology, UAE
E: psadhwani@hct.ac.ae

Sabeena Salam
Dubai Pharmacy College, UAE
E: s_sabeena@yahoo.com

Amira Salama
The American University in Cairo, Egypt
E: amirasalama@aucegypt.edu

Ashok Sapkota
Tribhuvan University, Nepal
E: assapkota@gmail.com

Ramy Shabara
The American University in Cairo, Egypt
E: ramy.shabara@aucegypt.edu

Ali Shehadeh
UAE University, UAE
E: ali.shehadeh@uaeu.ac.ae

Belinda Southby
Higher Colleges of Technology, UAE
E: belinda.southby@hct.ac.ae

Dara Tafazoli
University of Cordoba, Spain
E: dara.tafazoli@yahoo.com

Farhad Tayebipour
Majan University College, Oman
E: tayebipour@yahoo.com

Aidan Thorne
British Council Khartoum, Sudan
E: aidan.thorne@gmail.com

Reza Vahdanisanavi
Islamic Azad University, Iran
E: Vahdani.reza@gmail.com

Hua Wang
Shanghai Jiao Tong University, China
E: nnhuaw@163.com

Jie Wang
East China University of Science and Technology, China
E: jiewangcathy@126.com

Li Wang
Xi'an International Studies University, China
E: orley@126.com

Cyril Weir
CRELLA University of Bedfordshire, UK
E: cyrilweir@gmail.com

Eileen N. Whelan Ariza
Florida Atlantic University, USA
E: eariza@fau.edu

Beth Wiens
Zayed University, UAE
E: Beth.Wiens@zu.ac.ae

Zhongbao Zhao
Hunan Institute of Science and Technology, China
E: Michaelzhao998@hotmail.com

Shaoyan Zou
Qingdao Agricultural University, China
E: amandazsy@163.com

References

Bachman, L. & Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Baker, R. (1997). Classical test theory and item response theory in test analysis. Special Report No. 2. *Language Testing Update.*

Glaser, R. (1963). Instructional Technology and the Measurement of Learning Outcomes, *American Psychologist*, Vol 18: 519–521.

Hatch, E. & Lazaraton, A. (1991). *The Research Manual: Design and Statistics for Applied Linguistics,* New York, NY: Newbury House.

Knoch, U. (2009). Collaborating with ESP Stakeholders in Rating Scale Validation: The Case of the ICAO Rating Scale. *Spaan Fellow Working Papers in Second or Foreign Language Assessment. 7*, 20–46.

Kroll, B. & Reid, J. (1994). Guidelines for designing writing prompts. *Journal of Second Language Writing, 3*(3), 231–255.

Long, M. H. & Norris, J. M. (2000). Task-based language teaching and assessment. In M. Byram (Ed.), *Encyclopedia of Language Teaching* (pp. 597–603). London: Routledge.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5–11.

O'Sullivan, B. & Weir, C. (2011). Language Testing and Validation**.** In B. O'Sullivan (Ed.) *Language Testing: Theory & Practice* (pp.13–32). Oxford: Palgrave.

Poehner, M. (2008). *Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting L2 Development*. Berlin: Springer Publishing.

Richards, J. & Schmidt, R. (2010). *Longman Dictionary of Language Teaching and Applied Linguistics* (4th ed.). London: Longman.

Shehadeh, A. (2005). Task-based language learning and teaching: Theories and applications. In C. Edwards & J. Willis (Eds.), *Teachers Exploring Tasks in English Language Teaching* (pp. 13–30). London: Palgrave Macmillan.

Shehadeh, A. (2012). Task-based language assessment: Components, development, and implementation. In C. Coombe, P. Davidson, B. O'Sullivan & S. Stoynoff (Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 156–163). Cambridge, England: Cambridge University Press.

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. London: Longman.

Weaver, C. (2012). Incorporating a formative assessment cycle into task-based language teaching in a university setting in Japan. In A. Shehadeh & C. Coombe (Eds.), *Task-based Language Teaching in Foreign Language Contexts: Research and Implementation* (pp. 287–312). Amsterdam, The Netherlands: John Benjamins.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Oxford: Palgrave.

Weir, C.J., Vidakovic, I. & Galaczi, E.D. (2013). *Measured Constructs: A History of Cambridge English Examinations*, 1913–2012. Cambridge, UK: Cambridge University Press.