

LANGUAGE TESTING TERMINOLOGY

Achievement test: A test designed to measure what a person has learned within or up to a given time in a certain program of instruction. The content of achievement tests is a sample of what has been in the syllabus. Contrast proficiency test.

ACTFL guidelines: Descriptions of language learning outcomes that tell us what the learner can do, in which contexts, and how well

ACTFL rating scale: Adapted from the Foreign Service scale, the ACTFL scale ranges from Novice (no functional proficiency), to Intermediate (survival proficiency), to Advance (limited working proficiency) and Superior (professional proficiency)

Adaptive: A form of individually tailored testing in which test items are selected from an item bank where test items are stored in rank order with respect to their item difficulty and presented to test takers during the test on the basis of their responses to previous items, until it is determined that sufficient information regarding test takers' abilities has been collected

Analytic scoring: A method of marking which can be used in tests of productive language use, such as speaking and writing. The marker makes an assessment with the aid of a list of specific points. For example, in a test of writing the analytic scale may include a focus on grammar, vocabulary, use of linking devices, etc. Contrast global scoring.

Authentic text: Text used in a test which consists of materials originally produced for a non-language testing purpose such as newspapers, magazines, etc, and not specially produced for the test.

Background knowledge: A test taker's knowledge of topic or cultural content that may affect the way the test taker responds to an item.

Band: In an item based test, a range of several scores which may be reported as a grade or band score. In a rating scale designed to assess a specific trait or ability, such as speaking or writing, a band normally represents a particular level

Bias: A test or item can be considered to be biased if one particular section of the test taking population is advantaged or disadvantaged by some feature of the test or item which is not relevant to what is being measured. Sources of bias may be connected with gender, age, culture, etc.

Communicative competence: The ability to use language appropriately in a variety of situations and settings.

Completion item: An item type in which the test taker has to complete a sentence or phrase, usually by writing in several words or supplying details such as times and telephone numbers.

Computer adaptive test: A test administered by a computer in which the difficulty level of the next item to be presented to test takers is estimated on the basis of their responses to previous items and adapted to match their abilities

Construct: A construct refers to the knowledge, skill or ability that's being tested. In a more technical and specific sense, it refers to a hypothesized ability or mental trait which cannot necessarily be directly observed or measured, for example, listening ability. Language tests attempt to measure the different constructs which underlie language ability.

Constructed response: A type of test item or task that requires test takers to respond to a series of open-ended questions by writing, speaking, or doing something rather than choose answers from a ready-made list. The most commonly used types of constructed-response items include fill-in, short-answer, and performance assessment.

Contamination effect: A rater effect which occurs when a rater assigns a score on the basis of a factor other than that being tested. An example would be raising a test taker's score on a writing test because he or she had neat handwriting.

Content validity: A conceptual or non-statistical validity based on a systematic analysis of the test content to determine whether it includes an adequate sample of the target domain to be measured. An adequate sample involves ensuring that all major aspects are covered and in suitable proportions.

Criterion-referenced scale: A rating scale that provides for translating test scores into a statement about the behavior to be expected of a person with that score and/or their relationship to a specified subject matter. Similarly, a criterion-referenced test is one that assesses achievement or performance against a cut-score that is determined as representing mastery or attainment of specified objectives.

Cut score: A score that represents achievement of the criterion, the line between success and failure, mastery and non-mastery.

Descriptor: A brief description accompanying a band on a rating scale, which summarizes the degree of proficiency or type of performance expected for a test taker to achieve that particular score.

Dichotomous scoring: Scoring based on two categories, e.g., right/wrong, pass/fail. Compare polytomous scoring.

Discrete item: A self-contained item. An item that is not linked to a text, other items, or any supplementary materials. An example of an item used in this way is multiple-choice. Compare integrative item.

Distractor: The incorrect options in multiple-choice items.

Double rating: A method of assessing performance in which two individuals independently assess test taker performance on a test. Also called back reading, or back rating.

Extended response: A form of response to an item or task in which the test taker is expected to produce (as opposed to select) a response which is longer than one or two sentences.

Genre: A type of discourse that occurs in a particular setting, that has distinctive and recognizable patterns and norms of organization and structure, and that has particular and distinctive communicative functions

Global scoring: A method of scoring which can be used in tests of writing and speaking. The rater gives a single score according to the general impression made by the language produced, rather than by breaking it down into a number of scores for various aspects of language use. Also called holistic scoring. Contrast analytic scoring.

ILR scale: A scale of functional language ability of 0 to 5 used by the Interagency Language Roundtable; the range is from 0--no knowledge of a language to 5--equivalent to a highly educated native speaker

Indirect Test: A test that measures ability indirectly by requiring test takers to perform tasks not reflective of an authentic target language use situation, from which an inference is drawn about the abilities underlying their performance on the test

Input: Input material provided in a test task for the test taker to use in order to produce an appropriate response.

Integrative item/task: Used to refer to items or tasks which require more than one skill or sub-skill for their completion. For example, reading a letter and writing a response to it. Compare discrete item.

Inter-rater reliability: The degree of agreement between two assessments of the same sample of performance made at different times by the same rater.

Intra-rater reliability: The degree of agreement between two assessments of the same sample of performance made at different times by the same marker. In other words, the degree of agreement among raters or scorers of a test.

Item (also, test item): Each testing point in a test which is given a separate score or scores. Examples are: one gap in a cloze test; one multiple choice question with three or four options; one sentence for grammatical transformation; one question to which a sentence-length response is expected.

Key: The correct option or response to a test item.

Language proficiency: A degree of skill with which a person can use a language, such as how well a person can read, write, speak, or understand a language. This can be contrasted with language achievement, which describes language ability as a result of learning in a particular program or according to a particular syllabus.

Model answer: A good example of the expected response to an open-ended task which is provided by the item writer, and can be used in the development of a mark scheme to guide markers.

Multiple-choice item: A type of test item which consists of a question or incomplete sentence (stem), with a choice of answers or ways of completing the sentence (options). The test taker's task is to choose the correct option (key) from a set of possibilities. There may be any number of incorrect possibilities (distractors).

Open-ended question: A type of item or task in a written test which requires the test taker to supply, as opposed to select, a response.

Operational validity: The extent to which tasks, items or interviewers on a test perform as they are supposed to and function to create an accurate score in a real world setting, as opposed to a setting involving an experiment, a simulation or training.

Options: The range of possibilities in a multiple-choice item or matching tasks from which the correct one (key) must be selected.

Performance test: A test in which the ability of candidates to perform particular tasks, usually associated with job or study requirements, is assessed. Performance tests use 'real-life' performance as a criterion.

Polytomous scoring: Scoring an item using a scale of at least three points. For example, the answer to a question can be assigned 0, 1, or 2 points. Open-ended questions are often scored polytomously. Also referred to as *scalar* or *polychotomous scoring*. Compare *dichotomous scoring*.

Proficiency test: A test which measures how much of a language someone has learned. Proficiency tests are designed to measure the language ability of examinees regardless of how, when, why, or under what circumstances they may have experienced the language. Contrast *achievement test*.

Prompt: In test of speaking or writing, graphic materials or texts designed to elicit a response from the test taker.

Pronunciation: The ability to produce consonants, vowels and stress like most native speakers of the language.

Rater: Someone who assigns a score to a test taker's performance in a test, using subjective judgment to do so. Raters are normally qualified in the relevant field, and are required to undergo a process of training and standardization. Also referred to as *marker*, *scorer*, *assessor*, and *examiner*.

Rater effects: A source of error in assessment, the result of certain tendencies of raters such as harshness or leniency, or a prejudice in favor of a certain type of test takers, which affect the scores given to the test takers.

Rating: Assign a score to a test taker's response to a test. This may involve professional judgment, or the application of a rating scheme which lists all acceptable responses.

Rating scale: A scale consisting of several ranked and structured categories used for making subjective judgments. In language testing, rating scales of assessing performance are typically accompanied by band descriptors which make their interpretation clear.

Rating scheme: A list of all the acceptable response to the items in a test. A rating scheme makes it possible for a rater to assign a score to a test accurately. Also referred to as a *scoring rubric* or an *answer key*

Reverse Design: Using targeted functions (or functional outcomes) as a starting point to design assessments, and then design curriculum to prepare students to perform the targeted functions, measured by the corresponding assessments

Rubric: The instructions given to test takers to guide their responses to a particular test task. A scoring rubric consists of the instructions given to raters to guide their scoring.

Short answer item: An open-ended item for which the test taker is required to formulate a written answer using a word or a phrase.

Specifications (also, test specifications): A description of the characteristics of a test, including what is tested, how it is tested, and details such as number and length of forms, item types used

Stem: Part of a written prompt, usually an incomplete sentence or direct question for which the completion or correct response has to be supplied or selected from options.

Target Functions (Functional Outcomes): Descriptions, usually in reading, listening, writing and speaking, of what learners can do with the language in real-life situations at the end of a sequence of instruction.

Task type: Test tasks are referred to by names which tend to be descriptive of what they are testing and the form they take, e.g., multiple-choice reading comprehension.

Test-retest reliability: An estimate of the reliability of a test determined by the extent to which a test gives the same results if it is administered at two different points in time and assumes no change in ability or proficiency between the first and second administrations.

Validity: The degree to which a test measures what it is supposed to measure, or can be used successfully for the purpose for which it is intended. A number of different statistical procedures can be applied to a test to estimate its validity. Such procedures generally seek to determine what the test measures, and how well it does so.