Z.H. Usmonova, N.N. Xayrullayeva

CHET TILI O`QITISHNING INTEGRALLASHGAN KURSI

1-qism

Til bo'yicha bilimlarni baholash turlari va mezonlari

Usmonova Zarina Habibovna

Xayrullayeva Nodira Nematilloyevna

CHET TILI O`QITISHNING INTEGRALLASHGAN KURSI

(Til bo’yicha bilimlarni baholash turlari va mezonlari)

(1-qism)

5120100 – Filologiya va tillarni o'qitish

O‘QUV QO‘LLANMA

Buxoro – 2021

Usmonova Z.H., Xayrullayeva N.N. “Chet tili o`qitishning integrallashgan kursi” fanining “Til bo`yicha bilimlarni baholash turlari va mezonlari” moduli bo’yicha o`quv qo’llanma.

Mazkur o’quv qo’llanma “Til bo`yicha bilimlarni baholash turlari va mezonlari” moduli bo’yicha oliy ta’limning **5120100** – Filologiya va tillarni o’qitish ta’lim yo’nalishi IV bosqich bakalavriat talabalari uchun mo’ljalangan bo’lib, u “Chet tili o`qitishning integrallashgan kurs” fanidan O’zbekiston Respublikasi OO’MTVning 2018 yil 14-iyundagi 531-sonli buyrug’ining 10-ilovasi bilan tasdiqlangan namunaviy dastur asosida tuzilgan.

O’quv qo’llanmada talabalarning “Tesing and Assessment” ya`ni bilimlarni baholash turlari va mezonlari  matnlarga xos xususiyatlarni farqlay olishi va muloqot ijtimoiy-madaniy mavzulardagi so’zlarning ma’nosini bilishi va to’g’ri qo’llay olishi, integrallashgan ko`nikmalarni qo`llay olish, til ko`nikmalarini baholay olish, shuningdek, muloqot malakalarini rivojlantirish, baholash ko’nikmalarini shakllantirish hamda yozish, o`qish, eshitish va gapirish mashqlari asosida bu ko’nikmalarni  baholay olish ko`nikmalarini rivojlantirish maqsadida turli xorijiy adabiyotlarning matnlaridan parchalar berilgan, shuningdek, o`quv qo`llanma oxirida mavzular doirasida test va glossariy o’z ifodasini topgan.

Mas`ul muharrir :

Z.I. Rasulov *– Buxdu Ingliz tilshunosligi kafedrasi dotsenti , f.f.n.*

Taqrizchilar:

Q.B. Shodmonov *– Buxoro Tibbiyot instituti Ingliz tili kafedrasi professori, f.f.d*

N. F. Qosimova *– Buxdu Tarjimashunoslik va lingvodidaktika kafedrasi mudiri*

M.Q.Abuzalova *– Buxdu O`zbek Tilshunosligi kafedrasi professori, f.f.d*

Ushbu “Chet tili o`qitishning integrallashgan kursi” 1,2-qism nomli o`quv qo’llanma BuxDU Ilmiy kengashining “30” dekabr, 9-yig’ilishida yig’ilishida muhokama qilingan va nashrga tavsiya qilingan.

CONTENT

KIRISH

Mamlakatimizda “Chet tillarni o’rganish tizimini yanada takomillashtirish chora-tadbirlari to’g’risida”gi PQ 1875-sonli qarorning qabul qilinishi hozirgi kunda ta’lim tizimining barcha bosqichlarida talabalarga chet tillarini uzluksiz o’rganishni tashkil qilish, zamonaviy o’quv materiallar bilan ta’minlashni yanada takomillashtirish, shuningdek, zamonaviy pedagogik va axborot-kommunikasiya texnologiyalaridan foydalanib chet tillarni o’rganish, lug’at boyligini oshirish uchun manbalar (lug’atlar, registrlar, so’z ko’rsatkichlari va boshqalar) o’rganilgan bilim, ko’nikmalarni mustaqil ravishda amalda qo’llashga o’rgatish chet tili o’qituvchisi zimmasidagi eng katta mas’uliyat hisoblanadi.

Mazkur “Til bo`yicha bilimlarni baholash turlari va mezonlar” moduli bo’yicha mustaqil ishlash uchun tayyorlangan o`quv qo’llanma oliy ta’limning filologiya va tillarni o’qitish ta’lim yo’nalishi IV bosqich bakalavriat talabalari uchun mo’ljalangan bo’lib, u “Chet tili o`qitishning integrallashgan kursi ” fanidan O’zbekiston Respublikasi OO’MTVning 14- iyundagi 531-sonli buyrug’ining 10-ilovasi bilan tasdiqlangan namunaviy dastur asosida tayyorlangan.

Ushbu o`quv qo’llanmada talabalarning o`z o`zini baholay olish, ijodiy fikrlash, eshitish qobilyatlarini baholay bilishning maqbul usullaridan foydalanishni o’rgatish maqsadida turli xorijiy adabiyotlardan Mayk Boyl, Ellen Kislengerlarning “Skillful Testing And Assessment” va O’Dell, F.Redman va D.H.Braunlarning “Principles of language learning (intermediate and upper)”, Virjiniya Ivens and Jeni Duleylarning “Assess and Test (full set)” kabi mashq kitoblari matnlaridan parchalar berilgan, shuningdek, matnlar asosida testlar uslubiy qo`llanma so`ngida o’z ifodasini topgan.

Ushbu o`quv qo’llanma oliy o’quv yurtlarida ingliz tili fanidan iqtidorli talabalar bilan ishlash, va ularda mustaqil ta’limni tashkil etish, zamonaviy ped texnologiyalarni qo’llashni samaradorligini oshirishga qaratilgandir. Berilgan ushbu tavsiyalardan amaliy mashg’ulotlarda unumli foydalanish va talabalarni o’zlashtirish darajasini yuksaltirish maqsadida foydalansa bo’ladi.

THEME 1: THE KINDS OF ASSESSMENT AND THEIR USAGE IN EDUCATIONAL PURPOSES.

Theoretical basis: Getting started. Learn more.

Plan:

1. ***The main features of assessment in teaching.***
2. ***The purpose of using assessment in teaching.***
3. ***The kinds of assessment and the peculiarities of implementing them in learning language.***

The aim of this study is to analyze the importance of assessment in learning a foreign language acquisition in the Secondary Education classrooms in Uzbekistan. This research also proposes possible activities to be used by new generations of English teachers in order to facilitate a linguistic and cultural immersion essential for the acquisition of the English language.

The main objective of this section is to explain the difference between assessment types and to justify their importance in a foreign language acquisition process. To do this, the section has been divided into different subsections to have a general overview about way of teaching English through them and their assessments.

Comprehending and understanding a language is necessary when students are learning a new language due to the fact that people always need to communicate and interact with others in different moments or situations in their life.

Educational assessment is the process of documenting, usually in measurable terms, knowledge, skill, attitudes, and beliefs. It is a tool or method of obtaining information from tests or other sources about the achievement or abilities of individuals. Often used interchangeably with test.[1] Assessment can focus on the individual learner, the learning community (class, workshop, or other organized group of learners), the institution, or the educational system as a whole (also

[1] National council on Measurement in Education.

known as granularity). The word 'assessment' came into use in an educational context after the Second World War.[2]

The final purpose of assessment practices in education depends on the *theoretical framework* of the practitioners and researchers, their assumptions and beliefs about the nature of human mind, the origin of knowledge, and the process of learning.

The term *assessment* is generally used to refer to all activities teachers use to help students learn and to gauge student progress.[3] Assessment can be divided for the sake of convenience using the following categorizations:

1. Initial, formative, summative and diagnostic assessment
2. Objective and subjective
3. Referencing (criterion-referenced, norm-referenced, and ipsative)
4. Informal and formal
5. Internal and external

Assessment is often divided into initial, formative, and summative categories for the purpose of considering different objectives for assessment practices.

- Placement assessment – Placement evaluation is used to place students according to prior achievement or personal characteristics, at the most appropriate point in an instructional sequence, in a unique instructional strategy, or with a suitable teacher conducted through placement testing, i.e. the tests that colleges and universities use to assess college readiness and place students into their initial classes. Placement evaluation, also referred to as pre-assessment or initial assessment, is conducted prior to instruction or intervention to establish a baseline from which individual student growth can be measured. This type of an assessment

[2] Nelson, Robert; Dawson, Phillip (2014). "A contribution to the history of assessment: how a conversation simulator redeems Socratic method". Assessment & Evaluation in Higher Education. 39 (2): 195–204. .

[3] Black, Paul, & William, Dylan (October 1998). "Inside the Black Box: Raising Standards Through Classroom Assessment."Phi Beta Kappan. Available at PDKintl.org. Retrieved January 28, 2009.

is used to know what the student's skill level is about the subject. It helps the teacher to explain the material more efficiently. These assessments are not graded.[4]

- Formative assessment – Formative assessment is generally carried out throughout a course or project. Formative assessment, also referred to as "educative assessment," is used to aid learning. In an educational setting, formative assessment might be a teacher (or peer) or the learner, providing feedback on a student's work and would not necessarily be used for grading purposes. Formative assessments can take the form of diagnostic, standardized tests, quizzes, oral question, or draft work. Formative assessments are carried out concurrently with instructions. The result may count. The formative assessments aim to see if the students understand the instruction before doing a summative assessment.

- Summative assessment – Summative assessment is generally carried out at the end of a course or project. In an educational setting, summative assessments are typically used to assign students a course grade. Summative assessments are evaluative. Summative assessments are made to summarize what the students have learned, to determine whether they understand the subject matter well. This type of assessment is typically graded (e.g. pass/fail, 0-100) and can take the form of tests, exams or projects. Summative assessments are often used to determine whether a student has passed or failed a class. A criticism of summative assessments is that they are reductive, and learners discover how well they have acquired knowledge too late for it to be of use.

- Diagnostic assessment – Diagnostic assessment deals with the whole difficulties at the end that occurs during the learning process.

Jay McTighe and Ken O'Connor proposed seven practices to effective learning. One of them is about showing the criteria of the evaluation before the test. Another is about the importance of pre-assessment to know what the skill levels of a student are before giving instructions. Giving a lot of feedback and encouraging are other practices.

[4] Mctighe, Jay; O'Connor, Ken (November 2005). "Seven practices for effective learning". Educational Leadership. 63 (3): 10–17.

Educational researcher Robert Stake explains the difference between formative and summative assessment with the following analogy:

When the cook tastes the soup, that's formative. When the guests taste the soup, that's summative.[5]

Summative and formative assessment are often referred to in a learning context as *assessment of learning* and *assessment for learning* respectively. Assessment of learning is generally summative in nature and intended to measure learning outcomes and report those outcomes to students, parents and administrators. Assessment of learning generally occurs at the conclusion of a class, course, semester or academic year. Assessment for learning is generally formative in nature and is used by teachers to consider approaches to teaching and next steps for individual learners and the class.[6]

A common form of formative assessment is *diagnostic assessment*. Diagnostic assessment measures a student's current knowledge and skills for the purpose of identifying a suitable program of learning. *Self-assessment* is a form of diagnostic assessment which involves students assessing themselves. *Forward-looking assessment* asks those being assessed to consider themselves in hypothetical future situations.[7]

Performance-based assessment is similar to summative assessment, as it focuses on achievement. It is often aligned with the standards-based education reform and outcomes-based education movement. Though ideally they are significantly different from a traditional multiple choice test, they are most commonly associated with standards-based assessment which use free-form responses to standard questions scored by human scorers on a standards-based scale, meeting, falling below or exceeding a performance standard rather than

[5] Scriven, M. (1991). Evaluation thesaurus. 4th ed. Newbury Park, CA:Sage Publications.

[6] Earl, Lorna (2003). Assessment as Learning: Using Classroom Assessment to Maximise Student Learning. Thousand Oaks, CA, Corwin Press. ISBN 0-7619-4626-8. Available at, Accessed January 23, 2009.

[7] Reed, Daniel. "Diagnostic Assessment in Language Teaching and Learning." Center for Language Education and Research, available at Google.com. Retrieved January 28, 2009.

being ranked on a curve. A well-defined task is identified and students are asked to create, produce or do something, often in settings that involve real-world application of knowledge and skills. Proficiency is demonstrated by providing an extended response. Performance formats are further differentiated into products and performances. The performance may result in a product, such as a painting, portfolio, paper or exhibition, or it may consist of a performance, such as a speech, athletic skill, musical recital or reading.

Objective and subjective

Assessment (either summative or formative) is often categorized as either objective or subjective. Objective assessment is a form of questioning which has a single correct answer. Subjective assessment is a form of questioning which may have more than one correct answer (or more than one way of expressing the correct answer). There are various types of objective and subjective questions. Objective question types include true/false answers, multiple choice, multiple-response and matching questions. Subjective questions include extended-response questions and essays. Objective assessment is well suited to the increasingly popular computerized or online assessment format.

Some have argued that the distinction between objective and subjective assessments is neither useful nor accurate because, in reality, there is no such thing as "objective" assessment. In fact, all assessments are created with inherent biases built into decisions about relevant subject matter and content, as well as cultural (class, ethnic, and gender) biases.[8]

Basis of comparison

Test results can be compared against an established criterion, or against the performance of other students, or against previous performance:

- *Criterion-referenced assessment*, typically using a criterion-referenced test, as the name implies, occurs when candidates are measured against defined (and objective) criteria. Criterion-referenced assessment is often, but not always, used to

[8] Joint Information Systems Committee (JISC). "What Do We Mean by e-Assessment?". Retrieved January 29, 2009.

establish a person's competence (whether s/he can do something). The best known example of criterion-referenced assessment is the driving test, when learner drivers are measured against a range of explicit criteria (such as "Not endangering other road users").

- *Norm-referenced assessment* (colloquially known as "grading on the curve"), typically using a norm-referenced test, is not measured against defined criteria. This type of assessment is relative to the student body undertaking the assessment. It is effectively a way of comparing students. The IQ test is the best known example of norm-referenced assessment. Many entrance tests (to prestigious schools or universities) are norm-referenced, permitting a fixed proportion of students to pass ("passing" in this context means being accepted into the school or university rather than an explicit level of ability). This means that standards may vary from year to year, depending on the quality of the cohort; criterion-referenced assessment does not vary from year to year (unless the criteria change).[9]
- *Ipsative assessment* is self comparison either in the same domain over time, or comparative to other domains within the same student.

Informal and formal

Assessment can be either *formal* or *informal*. Formal assessment usually implies a written document, such as a test, quiz, or paper. A formal assessment is given a numerical score or grade based on student performance, whereas an informal assessment does not contribute to a student's final grade. An informal assessment usually occurs in a more casual manner and may include observation, inventories, checklists, rating scales, rubrics, performance and portfolio assessments, participation, peer and self-evaluation, and discussion.[10]

Internal and external

[9] Educational Technologies at Virginia Tech. "Assessment Purposes." VirginiaTech DesignShop: Lessons in Effective Teaching, available at Edtech.vt.edu. Retrieved January 29, 2009.

[10] Valencia, Sheila W. "What Are the Different Forms of Authentic Assessment?" Understanding Authentic Classroom-Based Literacy Assessment (1997), available at Eduplace.com. Retrieved January 29, 2009

Internal assessment is set and marked by the school (i.e. teachers). Students get the mark and feedback regarding the assessment. External assessment is set by the governing body, and is marked by non-biased personnel. Some external assessments give much more limited feedback in their marking. The criterion addressed by students is given detailed feedback in order for their teachers to address and compare the student's learning achievements and also to plan for the future.

In general, high-quality assessments are considered those with a high level of reliability and validity. Approaches to reliability and validity vary, however.

Reliability

Reliability relates to the consistency of an assessment. A reliable assessment is one that consistently achieves the same results with the same (or similar) cohort of students. Various factors affect reliability—including ambiguous questions, too many options within a question paper, vague marking instructions and poorly trained markers. Traditionally, the reliability of an assessment is based on the following:

1. Temporal stability: Performance on a test is comparable on two or more separate occasions.
2. Form equivalence: Performance among examinees is equivalent on different forms of a test based on the same content.
3. Internal consistency: Responses on a test are consistent across questions. For example: In a survey that asks respondents to rate attitudes toward technology, consistency would be expected in responses to the following questions:

- "I feel very negative about computers in general."
- "I enjoy using computers."[11]

Validity

Main article: Test validity

[11] Yu, Chong Ho (2005). "Reliability and Validity." Educational Assessment. Available at Creative-wisdom.com. Retrieved January 29, 2009.

Valid assessment is one that measures what it is intended to measure. For example, it would not be valid to assess driving skills through a written test alone. A more valid way of assessing driving skills would be through a combination of tests that help determine what a driver knows, such as through a written test of driving knowledge, and what a driver is able to do, such as through a performance assessment of actual driving. Teachers frequently complain that some examinations do not properly assess the syllabus upon which the examination is based; they are, effectively, questioning the validity of the exam.

Validity of an assessment is generally gauged through examination of evidence in the following categories:

1. Content – Does the content of the test measure stated objectives?

2. Criterion – Do scores correlate to an outside reference? (ex: Do high scores on a 4th grade reading test accurately predict reading skill in future grades?)

3. Construct – Does the assessment correspond to other significant variables? (ex: Do ESL students consistently perform differently on a writing exam than native English speakers?)[12]

4. Face – Does the item or theory make sense, and is it seemingly correct to the expert reader?[13]

A good assessment has both validity and reliability, plus the other quality attributes noted above for a specific context and purpose. In practice, an assessment is rarely totally valid or totally reliable. A ruler which is marked wrongly will always give the same (wrong) measurements. It is very reliable, but not very valid. Asking random individuals to tell the time without looking at a clock or watch is sometimes used as an example of an assessment which is valid, but not reliable. The answers will vary between individuals, but the average answer is probably close to the actual time. In many fields, such as medical research, educational testing, and psychology, there will often be a trade-off between

[12] Moskal Barbara M., & Leydens, Jon A (2000). "Scoring Rubric Development: Validity and Reliability." Practical Assessment, Research & Evaluation, 7(10). Retrieved January 30, 2009

[13] Committee on Standards for Educational Evaluation. (2003). The Student Evaluation Standards: How to Improve Evaluations of Students. Newbury Park, CA: Corwin Press.

reliability and validity. A history test written for high validity will have many essay and fill-in-the-blank questions. It will be a good measure of mastery of the subject, but difficult to score completely accurately. A history test written for high reliability will be entirely multiple choice. It isn't as good at measuring knowledge of history, but can easily be scored with great precision. We may generalize from this. The more reliable our estimate is of what we purport to measure, the less certain we are that we are actually measuring that aspect of attainment.

It is well to distinguish between "subject-matter" validity and "predictive" validity. The former, used widely in education, predicts the score a student would get on a similar test but with different questions. The latter, used widely in the workplace, predicts performance. Thus, a subject-matter-valid test of knowledge of driving rules is appropriate while a predictively valid test would assess whether the potential driver could follow those rules.

In the field of evaluation, and in particular educational evaluation, the Joint Committee on Standards for Educational Evaluation has published three sets of standards for evaluations. "The Personnel Evaluation Standards" was published in 1988, *The Program Evaluation Standards* (2nd edition) was published in 1994, and *The Student Evaluation Standards* was published in 2003.

Each publication presents and elaborates a set of standards for use in a variety of educational settings. The standards provide guidelines for designing, implementing, assessing and improving the identified form of evaluation. Each of the standards has been placed in one of four fundamental categories to promote educational evaluations that are proper, useful, feasible, and accurate. In these sets of standards, validity and reliability considerations are covered under the accuracy topic. For example, the student accuracy standards help ensure that student evaluations will provide sound, accurate, and credible information about student learning and performance.

Comprehension questions:

1. What is the main features of assessment system do you know?
2. What types of assessment are there in teaching language?

3. How do you distinguish assessment types in usage?

Tests

The objectives*:- to explore the term assessment and its types;*

1.***What is multiple-choice test?***

a. An assessment instrument in which items offer the test-taker a choice among two or more listed options
b. Form of individualized written feedback about a student's performance, sometimes used as an alternative or supplement to a letter grade
c. In a writing test, a single score indicating the effectiveness of the text in achieving its primary goal
d. A test that is not limited to any one course, curriculum, or single skill in the language; rather, it tests overall global ability

2. ***What is integrative test?***

a. A test that treats language competence as a unified set of interacting abilities of grammar, vocabulary, reading, writing, speaking, and listening
b. The extent to which the linguistic criteria of the test (e.g., specified classroom objectives) are measured and implied predetermined levels of performance are actually reached
c. A test in which the absence of predetermined or absolutely correct responses require the judgment of the teacher to determine correct and incorrectanswers
d. A test that is not limited to any one course, curriculum, or single skill in the language; rather, it tests overall global ability

3. Subjective tests are…

a. Tests in which the absence of predetermined or absolutely correct responses require the judgment of the teacher to determine correct and incorrect Answers
b. Assessments that involve learners in actually performing the behavior that one purports to measure
c. Tests that aim to measure, or summarize, what a student has grasped and

typically occurs at the end of a course or unit of instruction

d. Test that are not limited to any one course, curriculum, or single skill in the language; rather, it tests overall global ability

4. ***Types of alternative assessment?***

a. Preliminary, Formative, Summative

b. Test

c. Active and passive assessments

d. Traditional and innovative

5. ***Summative test is …***

a. a test that aims to measure, or summarize, what a student has grasped and typically occurs at the end of a course or unit of instruction

b. a test that is not limited to any one course, curriculum, or single skill in the language; rather, it tests overall global ability

c. statements that describe what a student can perform at a particular point on a rating scale; sometimes also called *band descriptors*

b. attending to the end result of a linguistic action (e.g., in writing, the "final" paper, versus the various steps involved in composing the paper

6. ***Proficiency test is….***

a. a test that is not limited to any one course, curriculum, or single skill in the language; rather, it tests overall global ability

b. a test that aims to measure, or summarize, what a student has grasped and typically occurs at the end of a course or unit of instruction

c. A test that has predetermined fixed responses

d. a test in which each test-taker's score is interpreted in relation to a mean (average score), median (middle score), standard deviation (extent of variance in scores), and/or percentile rank

7. ***Communicative test is ….***

a. a test that elicits a test-taker's ability to use language that is meaningful and authentic

b. a test that is not limited to any one course, curriculum, or single skill in the language; rather, it tests overall global ability

c. A test that has predetermined fixed responses

d. a test that aims to measure, or summarize, what a student has grasped and typically occurs at the end of a course or unit of instruction

8. Which task involves deleting a number of words in the whole text, requiring the test-takers to restore the original words?

a. Multiple choice tests

b. Gap filling

c. Cloze procedure

d. Placement tests

THEME 2. SUMMATIVE AND FORMATIVE ASSESSMENT

Getting started. Learn more.

Plan:

1.The main features of summative assessment in teaching.

2.The purpose of using formative assessment in teaching.

3.Two kinds of assessment and the peculiarities of implementing them in learning language.

The CEFR distinguishes between formative and summative assessment (CEFR, 9.3.5, p. 186).

Summative assessment usually refers to a student's attainment at the end of a course, as indicated by a grade or a rank. Formative assessment involves the ongoing, informal processes of assessment related to teaching and learning. This distinction raises questions not just of the order, "What do we assess?" and "How do we assess?", but more importantly, questions about the goals and purposes of assessment, it tries to reply to the question "Why?"

Within the process of learning, formative assessment involves gathering information and providing feedback for learners and teachers alike. Such feedback is effective to the extent that those receiving it are able to make use of it, particularly being able to understand it, to take it into account, and to profit from it. To do so, people need to develop a metalanguage about assessment, which may require specific training and development of awareness, which may in turn increase their motivation. As already stated, these forms of assessment are at either ends of a continuum and are complementary.

The step from formative assessment to self-assessment is short. It is not coincidental that the CEFR puts "Assessment by others and self-assessment" as the final set of pairs among the different types of assessment (CEFR, 9.3.13, pp. 191–192). From a perspective of effectiveness, self-assessment plays a considerable role. To do self-assessment, learners need to have suitable tools at their disposal. The assumption that rating on a scale and rating on a checklist (CEFR, 9.3.9, p. 189) are complementary is fully justified as shown in Chapter 3 of the CEFR and in other respects in the ELP (European Language Portfolio).

	Pre-assessment	**Formative Assessment**	**Summative Assessment**
What is it?	Assessment that is used to collect information about	Assessment that gathers information about student learning.	Assessment that shows what students have learned.
When is it used?	Before a lesson or new unit of study.	During a lesson or unit of study.	At the end of a lesson or unit of study.
Why is it used?	To determine the readiness level of students and to inform	To track students' progress and to make changes to instruction.	To provide evidence of what students learned.

Formative assessment is a process that uses informal assessment strategies to gather information on student learning. Teachers determine what students are understanding and what they still need to learn to master a goal or outcome. Strategies used to gather formative assessment information take place during regular class instruction as formative assessment and instruction are closely linked.

Most formative assessment strategies are quick and easy to use and fit seamlessly into the instruction process. The information gathered is never marked or graded. Descriptive feedback may accompany a formative assessment to let students know whether they have mastered an outcome or whether they require more practice.

Formative assessment strategies are used throughout a unit of study. They are linked to the instruction and focus on discovering what students know and need to know about the end goal or outcome. Teachers use formative assessment during the learning process and use the information to make adjustments to their instruction to better satisfy learner needs. Using formative assessment over the course of a unit will provide teachers with information on the learning processes of their students. Teachers can use one assessment strategy, change or adapt the instruction, and then reassess using the same strategy or a different one to determine if the instructional practice is impacting student achievement.

Formative assessment strategies are used to check for understanding of student learning and to make decisions about current and future instruction. Through formative assessment, teachers can discover the rate at which students are learning, the current knowledge of students, what information or skills students still need to learn, and whether the learning opportunities they are providing for students is effective or if they need to change or adapt their instruction. Results of formative assessment drive instruction. If students are doing well and progressing as expected, teachers continue with their current instruction practices. If students are not progressing as expected and are missing key information or skills, teachers plan other learning opportunities to help students attain the information or skills they need to be successful.

During a unit on measurement in math, teachers may set up demonstration stations for students to show what they have learned using standard measures studied throughout the unit. As students participate in the demonstration stations, teachers focus on the process the students are using to attain a solution, as well as the solution itself.

Deciding on what type of formative assessment strategy to use will depend on a number of factors. Teachers need to determine what aspect of student learning they want to measure. They then need to consider the learning preferences of their students. Formative assessment strategies can be given to students individually, as partners, in small groups, or as a class. The type of grouping used for the formative assessment will also influence the choice of strategy. Teachers should not rely on one type of assessment strategy. A variety of individual and group formative assessment strategies should be used. Individual strategies allow teachers to get a clear picture of each student and their understanding of the concept or skill being measured. Group strategies provide teachers with general information about student learning that can be used to plan instruction. Students can also use formative assessment information to make changes to their learning.

Teachers use formative assessment information to assess how their current instructional strategies are working with their students. If there are students who are struggling, teachers may need to work individually with a student, present information other ways, or adapt their current instructional strategy. Students who have appeared to master the outcome or goal being formatively assessed, may need to be further assessed or have learning opportunities planned that challenge them and are designed at their level of understanding. Teachers are also able to identify misunderstandings students may have and adapt their instruction accordingly.

Students can use formative assessment information to determine what they need to do to achieve the goals or outcomes of the unit. Students may need to adapt or to change their learning to master curriculum outcomes. If students are not achieving at an expected rate, they can look at the strategies they are using for learning and decide whether they need to change their current learning strategies or adopt new ways of learning. The information provided by formative assessment strategies can also be used to help students reflect on current learning goals or set new goals.

Formative assessment strategies for teachers:

ABC Brainstorming; Analogies; Checklists; Choral Response; Cloze

Procedure; Concept Maps; Conferences; Computer Surveys; Demonstration Stations; Discussions; Double Entry Journals; Drawings; Email Questions; Examples/Non-Examples; Exit Cards; First of Five; Four Corners; Graffiti Wall; Graphic Organizers; Individual Whiteboards; Inside-Outside Circle; Learning Logs; List Ten Things; Matching Activities; Observations; One Minute Essays; One Minute Fluency; One Sentence Summaries; Open-Ended Questions; Paper Pass; Peer-Assessments; Placemats; Problem Solving; Questionnaires; Questioning; Quick Writes; Reflection Journals; Repeat Pre-assessments; Response Cards; Self-Assessments; Sentence Prompts; Show of Hands; Student Composed Questions; Teach a Friend; Think-Pair-Share; Three Facts and a Fib; Three Minute Pause; Three Things; Thumbs Up, Thumbs Down; Traffic Light; Turn and Talk; Whip Around;

Formative assessment strategies for students:

Ask; Checklists; Journals; Process Exemplars; Product Exemplars; Self-Marking Quizzes; Writing Continuums.

Formative assessment:

- Requires students to take responsibility for their own learning
- Communicates clear, specific learning goals
- Focuses on goals that represent valuable educational outcomes with applicability beyond the learning context
- Identifies the student's current knowledge/skills and the necessary steps for reaching the desired goals
- Requires development of plans for attaining the desired goals
- Encourages students to self-monitor progress toward the learning goals
- Provides examples of learning goals including, when relevant, the specific grading criteria or rubrics that will be used to evaluate the student's work
- Provides frequent assessment, including peer and student self-assessment and assessment embedded within learning activities.
- Includes feedback that is non-evaluative, specific, timely, and related to the learning goals, and that provides opportunities for the student to revise and

improve work products and deepen understandings

- Promotes metacognition and reflection by students on their work

By contrast, another type of assessment, formative assessment, takes place before or during the instruction with the explicit purpose of eliciting evidence that can be used to improve the current learning. One widely accepted definition of formative assessment describes it as a classroom-based process in which students and teachers collect evidence of learning in order to understand current learning progress and to make adjustments to learning or to teaching as necessary Such adjustments could include the development quality formative assessment, the focus will always be on promoting learning by targeting teacher (and peer) support for specific student needs. For formative assessment to be effective, classroom practices that assume students simply learn what the teacher presents to them must be interrupted and replaced with a process that tailors support to student learning needs. For example, when teachers become more aware of students' learning progress, and in some cases their struggles in learning, the next step in the process requires action from the teacher to help students either overcome the struggles or reach even higher. That action may require the teacher to change future lesson plans to spend additional time on those areas with which students are struggling or with those students who are struggling, and the additional collection of follow-up evidence to determine whether the action taken was successful.

Comprehension questions:

1.What is the main peculiarities of summative assessment in learning language?

2. What is the main features of formative assessment in teaching?

3. Are there any differences between summative and formative assessment?

Tests:

The aim of test is (or should be) to help language learner to cope with assessment in teaching.

1. According to the Glossary for Educational Reform, what assessment are defined by three criteria?

a. They are used to determine whether students have learned what they were expected to learn or to level or degree to which students have learned the material.

b. They may be used to measure learning progress and achievement and to evaluate the effectiveness of educational programs. Tests may also measure student progress toward stated improvement goals or to determine student placement in programs.

c. They are recorded as scores or grades for a student's academic record for a report card or for admission to higher education.

2. At the district, state, or national level, ……………. tests are an additional form of summative assessments. The legislation passed in 2002 known as the No Child Left Behind Act mandated annual testing in every state. This testing was linked to federal funding of public schools.

a) Placement

b) Formative

c) Summative

d) Diagnostic

e) Standardized

3. ... refers to the activities required by students during the conduct of a course. It takes place within the normal teaching period and contributes to the final assessment.

a) Continuous assessment

b) Formative assessment

c) Summative assessment

d) Self assessment

4. It is used in various stages throughout the language course to determine learner's progress up to that point and to see what they have learnt.

a) Progress testing

b) Proficiency testing

c) Achievement testing

d) Self assessment

5. ... a process in which you make a judgment about a person or situation, or the judgment you make. Fill in the gaps.

a) Assessment

b) Marking

c) Testing

d) Teaching

6. Find the right types of assessment.

a) Diagnostic and summative

b) Summative and formality

c) Validity and reliability

d) Summative and validity

7. Assessment for learning is … .

a) Formative

b) Summative

c) Natural

d) Alternative

8. Evaluation that is most often undertaken at the end of a project:

a) Summative evaluation

b) Follow up evaluation

c) Summary evaluation

d) Diagnostic evaluation

9. The main goal of summative assessments is to …

a) Check understanding

b) Monitor learning

c) Evaluate learning

d) Get ranked

10. The main goal of formative assessments is to …

a) Monitor learning

b) Evaluate learning

c) Grade harshly

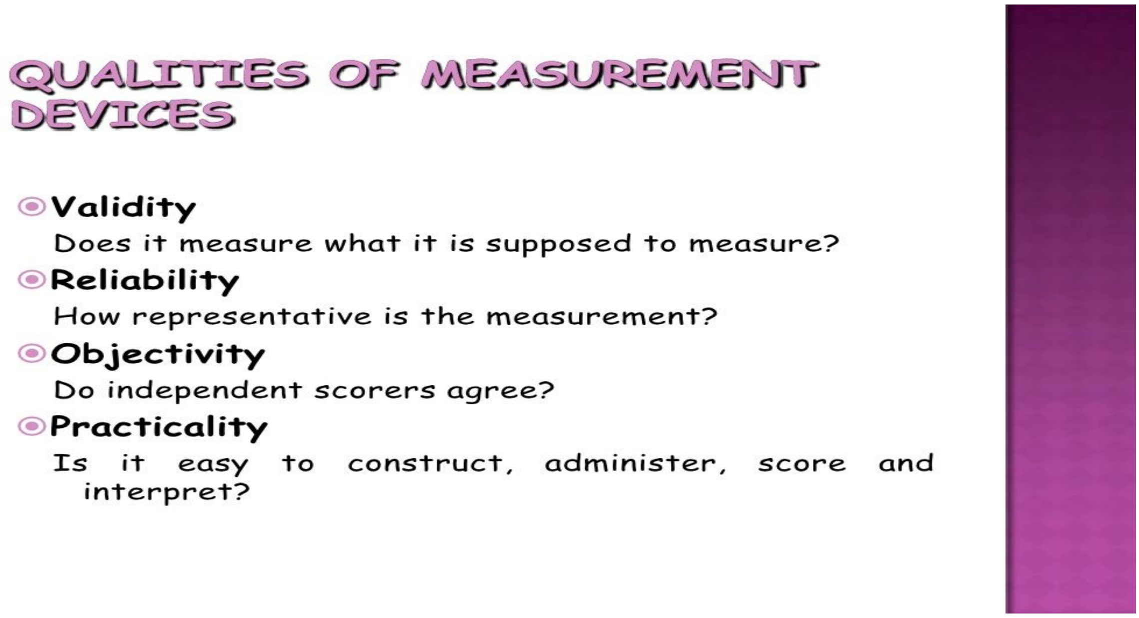
d) Grade fairly

THEME: 3. PRACTICALITY, RELIABILITY AND VALIDITY

Getting started. Learn more.

Plan:

1. The main importance of practicality in testing.
2. The main features of validity in testing.
3. The role reliability in testing the students.



QUALITIES OF MEASUREMENT DEVICES

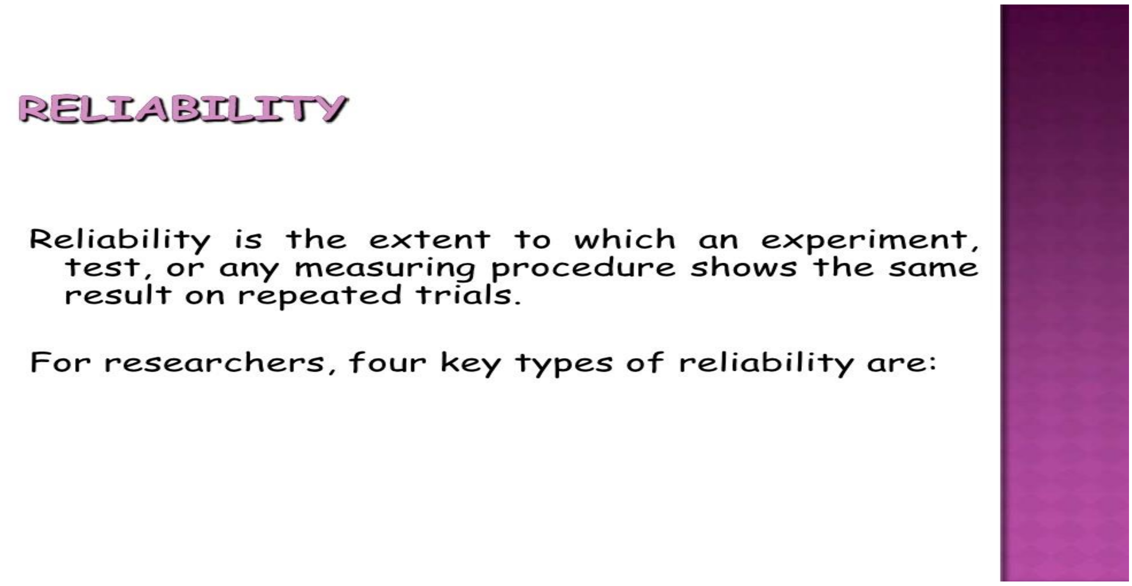
- **Validity**
  Does it measure what it is supposed to measure?
- **Reliability**
  How representative is the measurement?
- **Objectivity**
  Do independent scorers agree?
- **Practicality**
  Is it easy to construct, administer, score and interpret?

Practicality

Every good assessment has to be practical. In an ideal world all assessments would be identical to what the target task is. If you are testing an English as a second language learner and their ability to provide customer service in English while working in a hotel, the ideal way to test and see if a learner can actually do that task is to actually have the learner go to a hotel and work with customers; however, this isn't very practical.

The first reason why this isn't practical is that if a student messes up, the business could really suffer. This however, wouldn't be an issue if the target task was something less risky. If the target task is using a foreign language to buy an item and you have access to speakers of that language, a student could go and try doing the actual task. If they fail the task, there really isn’t much of a risk of losing anything.

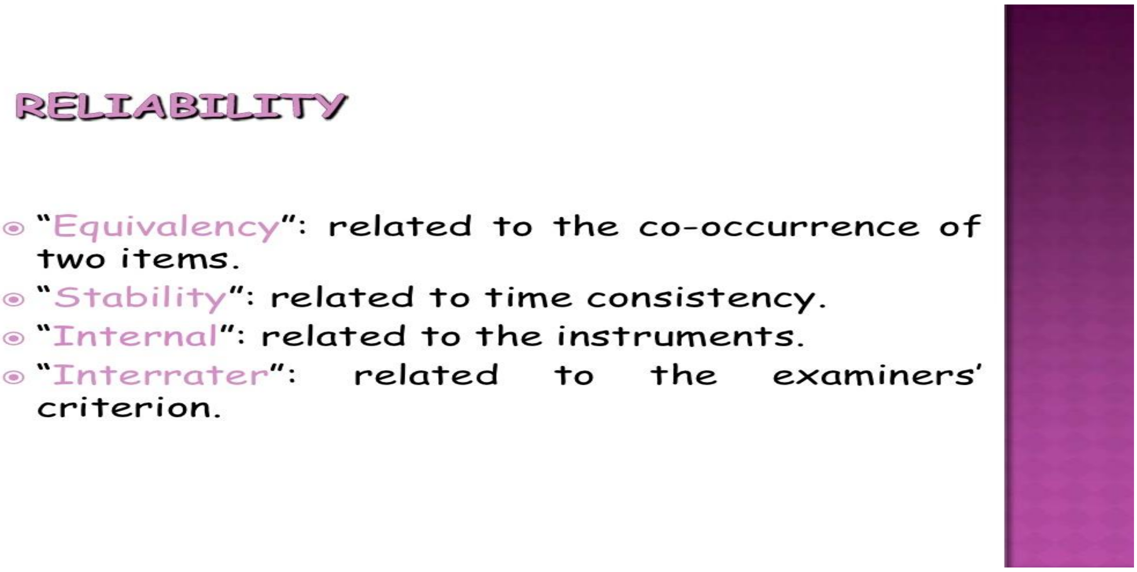
The bigger issue with practicality is that if you have a class of many students, it would be nearly impossible for all students to be able to complete a task like this in a reasonable amount of time. If you only needed one student to complete a task like shopping using a foreign language, it wouldn’t be much of a hassle; however, imagine having a class of ten students. You would need to take ten students to a place where you could administer the assessment, actually administer the assessment, and grade the assessment.



RELIABILITY

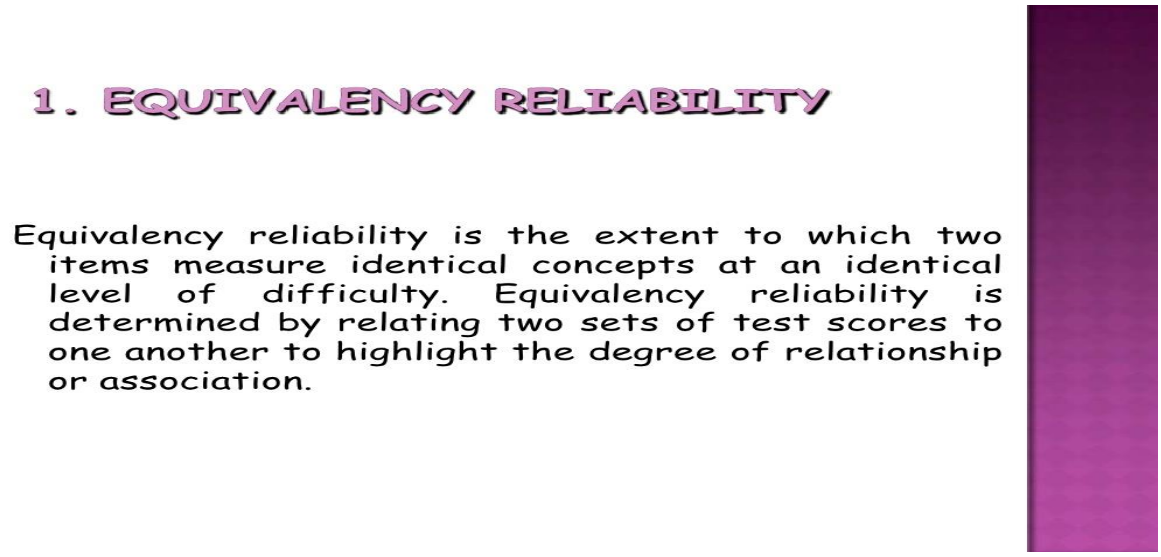
Reliability is the extent to which an experiment, test, or any measuring procedure shows the same result on repeated trials.

For researchers, four key types of reliability are:



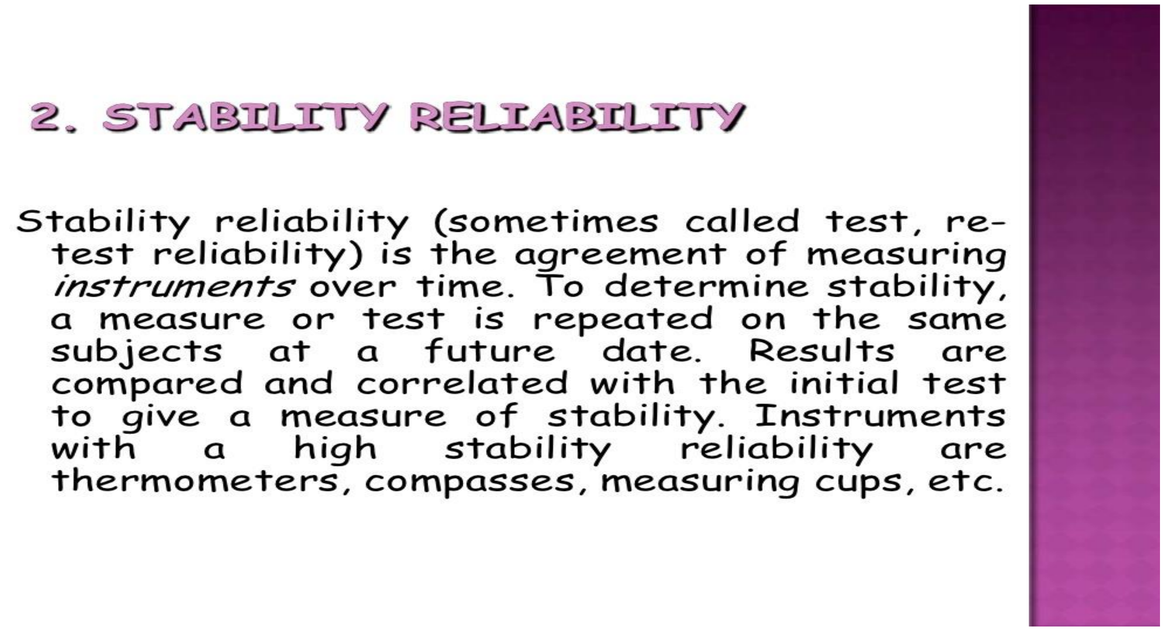
RELIABILITY

- “Equivalency”: related to the co-occurrence of two items.
- “Stability”: related to time consistency.
- “Internal”: related to the instruments.
- “Interrater”: related to the examiners’ criterion.



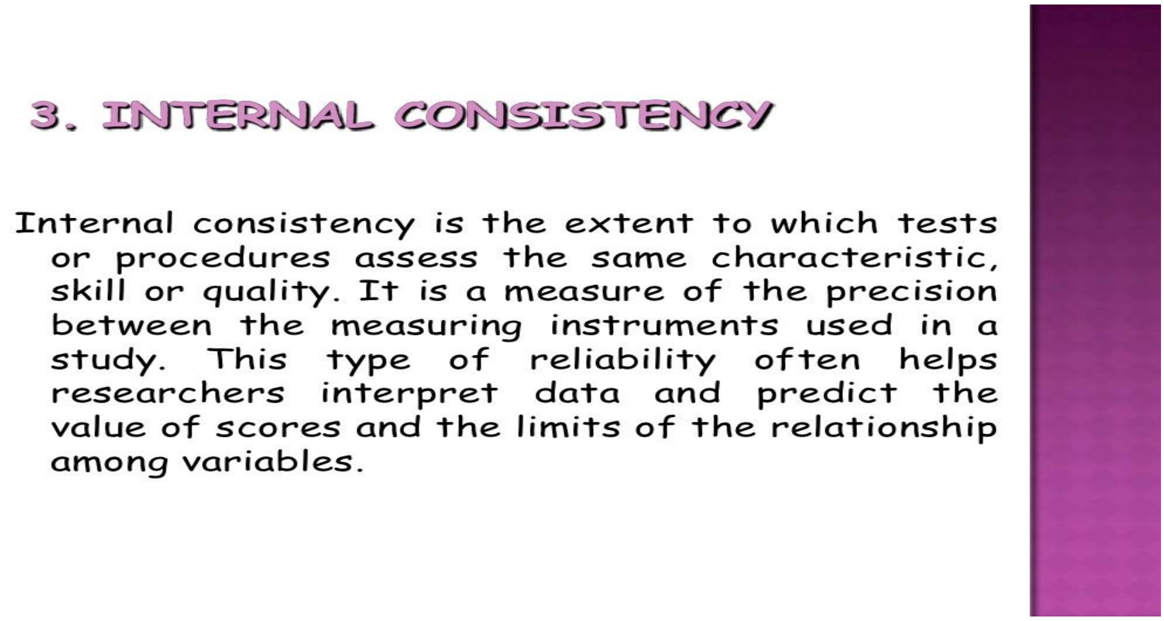
1. EQUIVALENCY RELIABILITY

Equivalency reliability is the extent to which two items measure identical concepts at an identical level of difficulty. Equivalency reliability is determined by relating two sets of test scores to one another to highlight the degree of relationship or association.



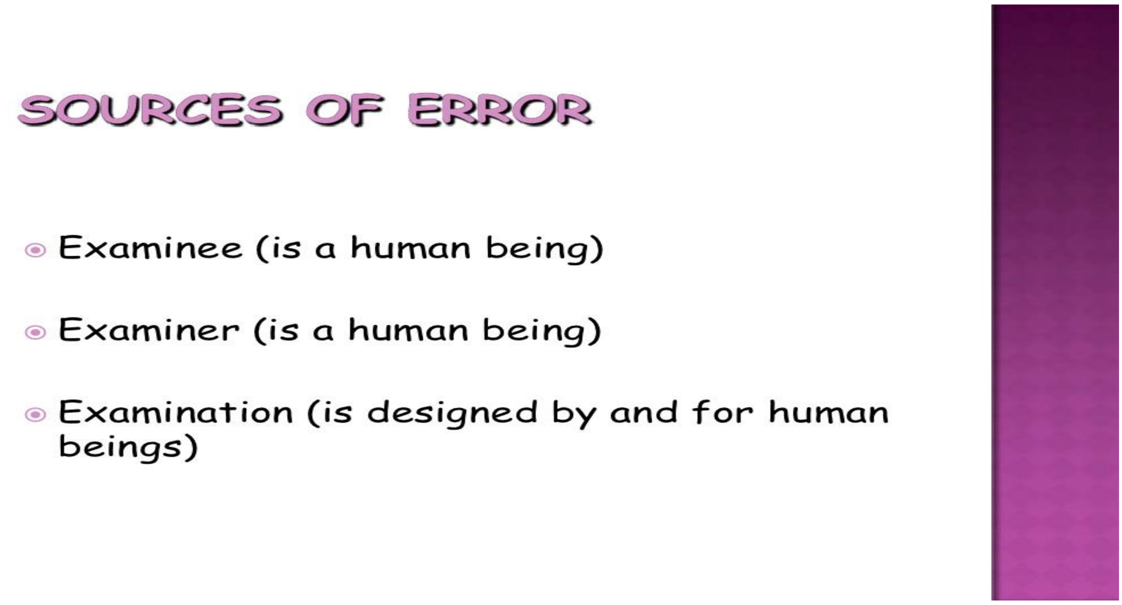
2. STABILITY RELIABILITY

Stability reliability (sometimes called test, re-test reliability) is the agreement of measuring *instruments* over time. To determine stability, a measure or test is repeated on the same subjects at a future date. Results are compared and correlated with the initial test to give a measure of stability. Instruments with a high stability reliability are thermometers, compasses, measuring cups, etc.



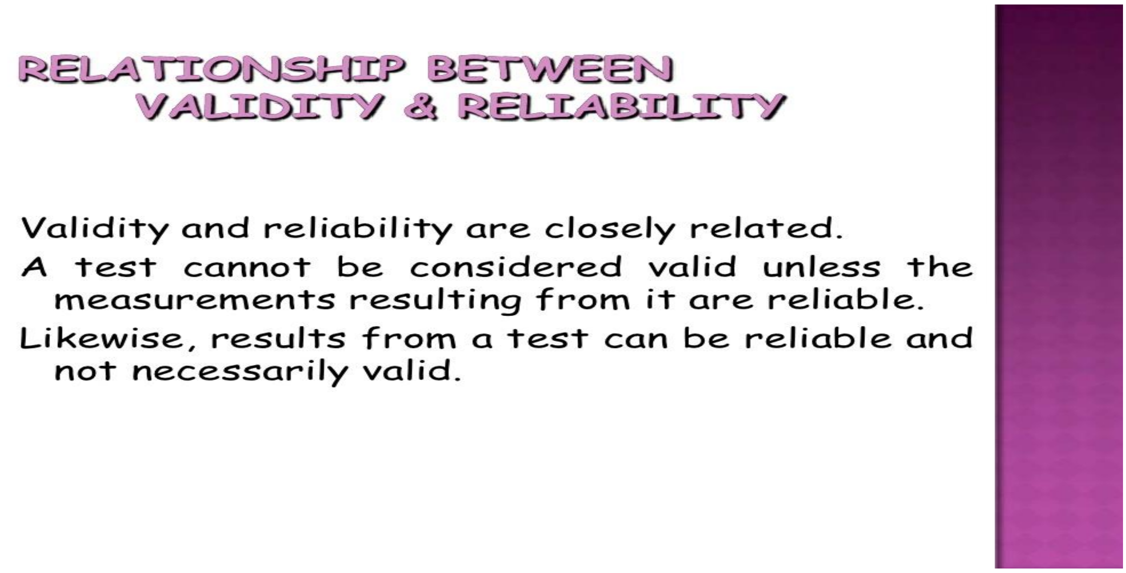
3. INTERNAL CONSISTENCY

Internal consistency is the extent to which tests or procedures assess the same characteristic, skill or quality. It is a measure of the precision between the measuring instruments used in a study. This type of reliability often helps researchers interpret data and predict the value of scores and the limits of the relationship among variables.



SOURCES OF ERROR

- Examinee (is a human being)
- Examiner (is a human being)
- Examination (is designed by and for human beings)

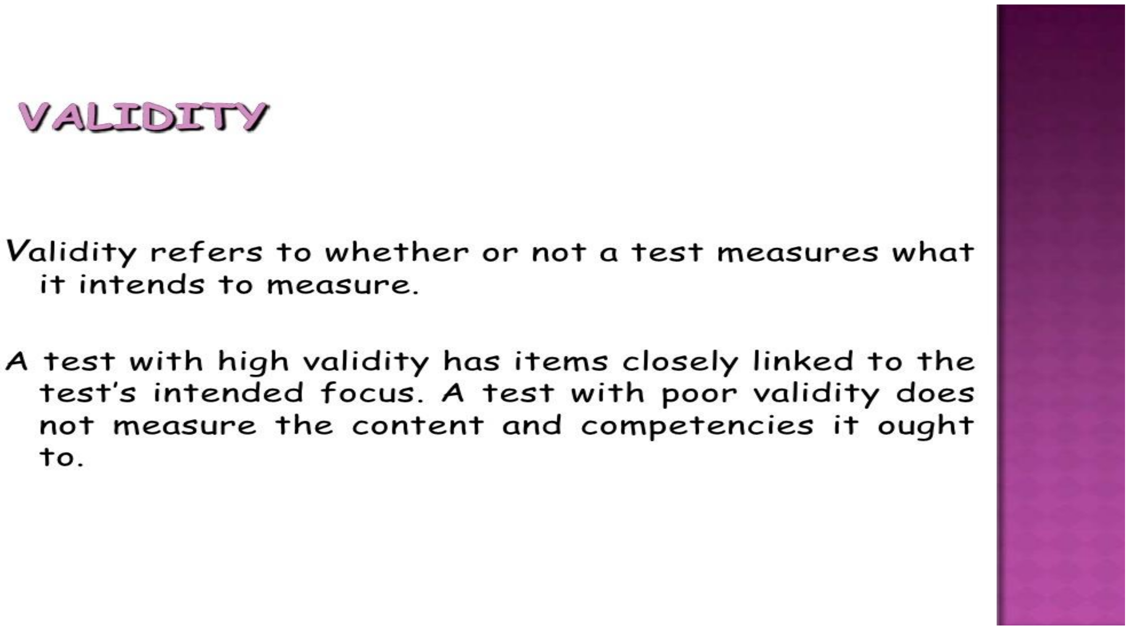


RELATIONSHIP BETWEEN VALIDITY & RELIABILITY

Validity and reliability are closely related.

A test cannot be considered valid unless the measurements resulting from it are reliable.

Likewise, results from a test can be reliable and not necessarily valid.



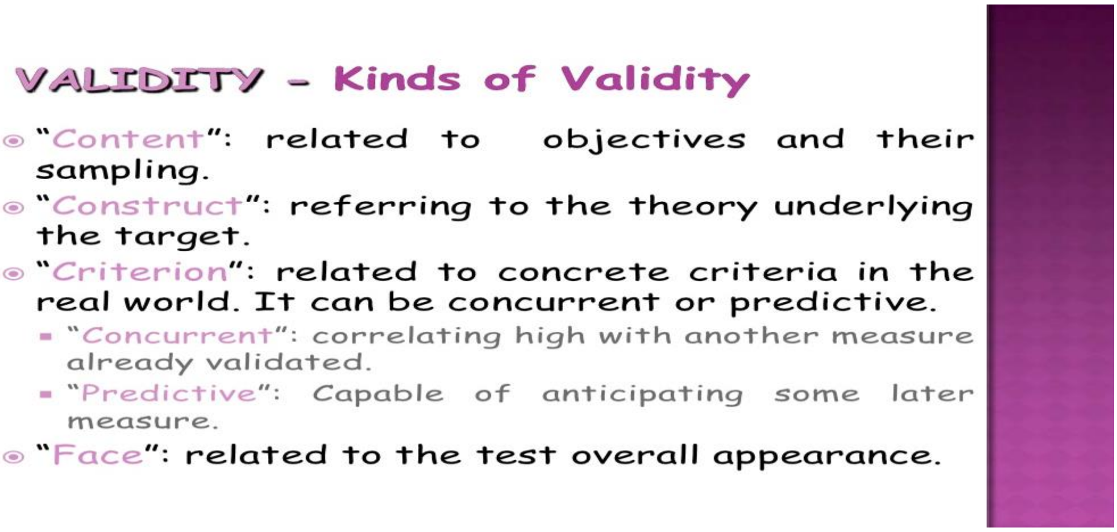
VALIDITY

Validity refers to whether or not a test measures what it intends to measure.

A test with high validity has items closely linked to the test's intended focus. A test with poor validity does not measure the content and competencies it ought to.

RELIABILITY

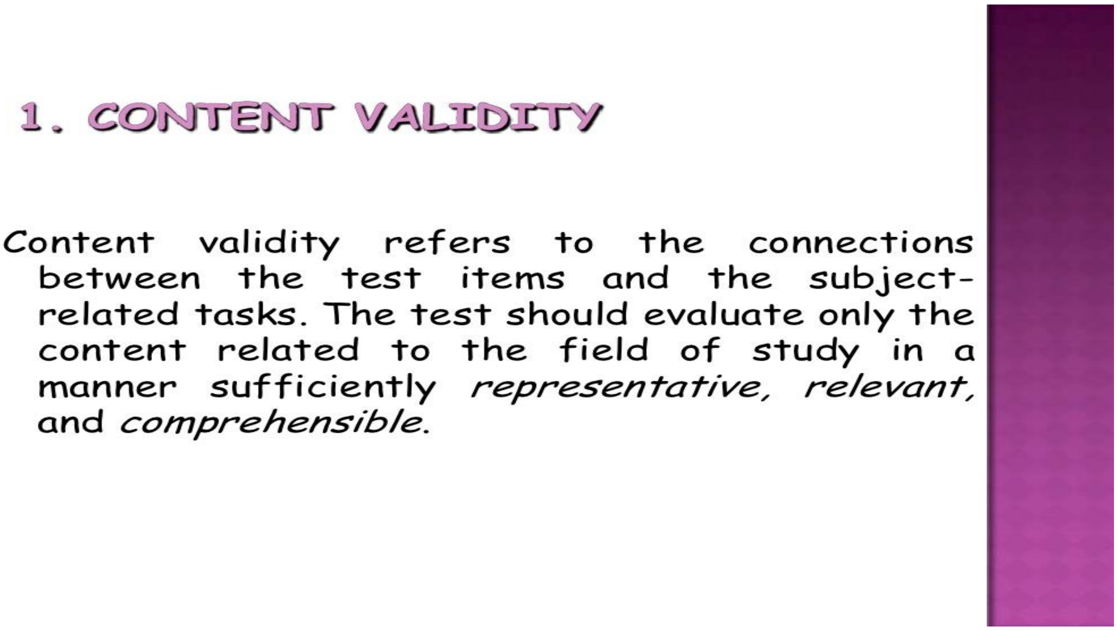
A test also has to be reliable. This means that the test results are consistent and dependable. If students of similar skill level take an assessment, they should receive a similar grade. Additionally, if the students were to retake the assessment, their scores should be similar to the previous score, assuming that the students didn't study more after taking the first assessment.



VALIDITY – Kinds of Validity

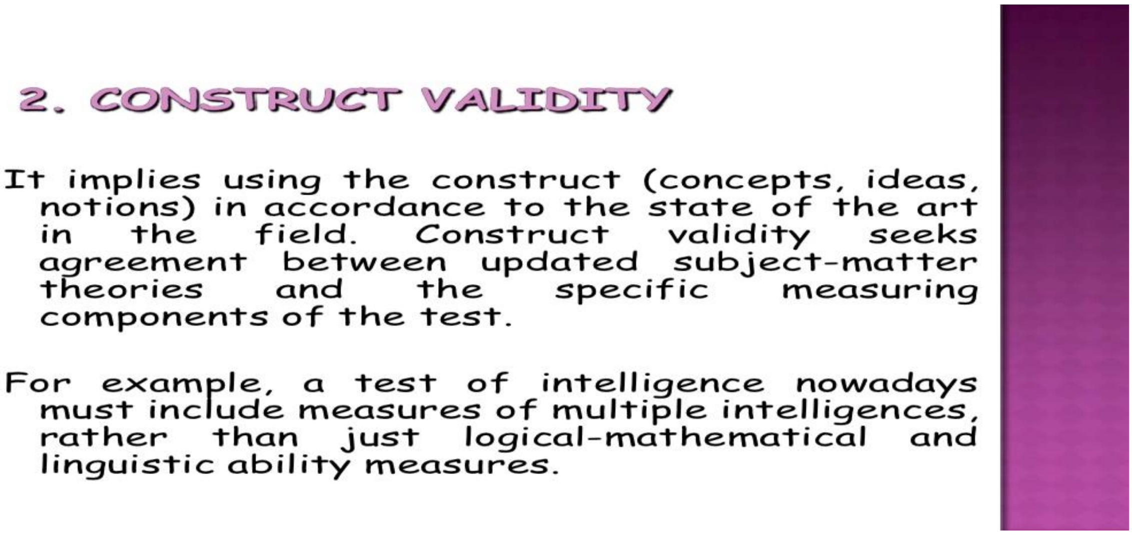
- "Content": related to objectives and their sampling.
- "Construct": referring to the theory underlying the target.
- "Criterion": related to concrete criteria in the real world. It can be concurrent or predictive.
  - "Concurrent": correlating high with another measure already validated.
  - "Predictive": Capable of anticipating some later measure.
- "Face": related to the test overall appearance.

The last thing a good test needs is validity. Validity answers the question "does the test actually measure what it is intended to measure?" There should be a strong relationship with what the assessment is measuring and how that reflects the student's ability to do the test in a real life situation.



1. CONTENT VALIDITY

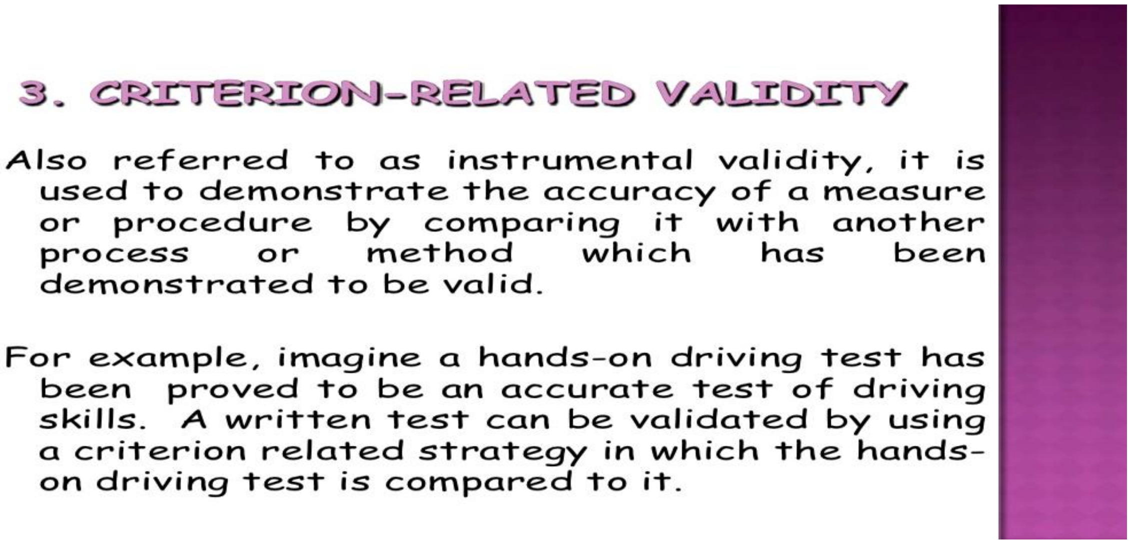
Content validity refers to the connections between the test items and the subject-related tasks. The test should evaluate only the content related to the field of study in a manner sufficiently *representative, relevant,* and *comprehensible.*



2. CONSTRUCT VALIDITY

It implies using the construct (concepts, ideas, notions) in accordance to the state of the art in the field. Construct validity seeks agreement between updated subject-matter theories and the specific measuring components of the test.

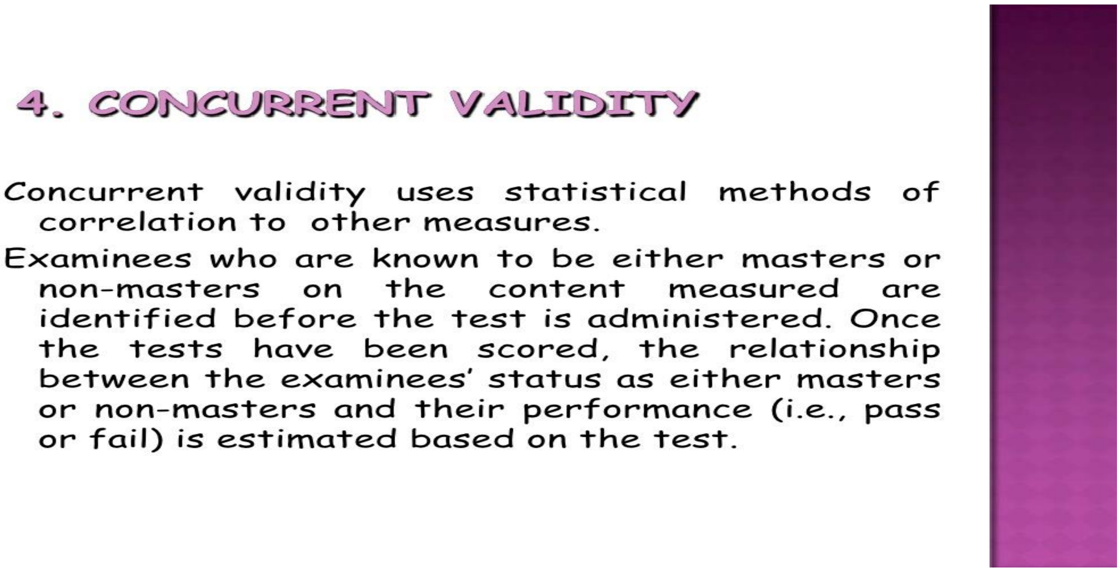
For example, a test of intelligence nowadays must include measures of multiple intelligences, rather than just logical-mathematical and linguistic ability measures.



3. CRITERION-RELATED VALIDITY

Also referred to as instrumental validity, it is used to demonstrate the accuracy of a measure or procedure by comparing it with another process or method which has been demonstrated to be valid.

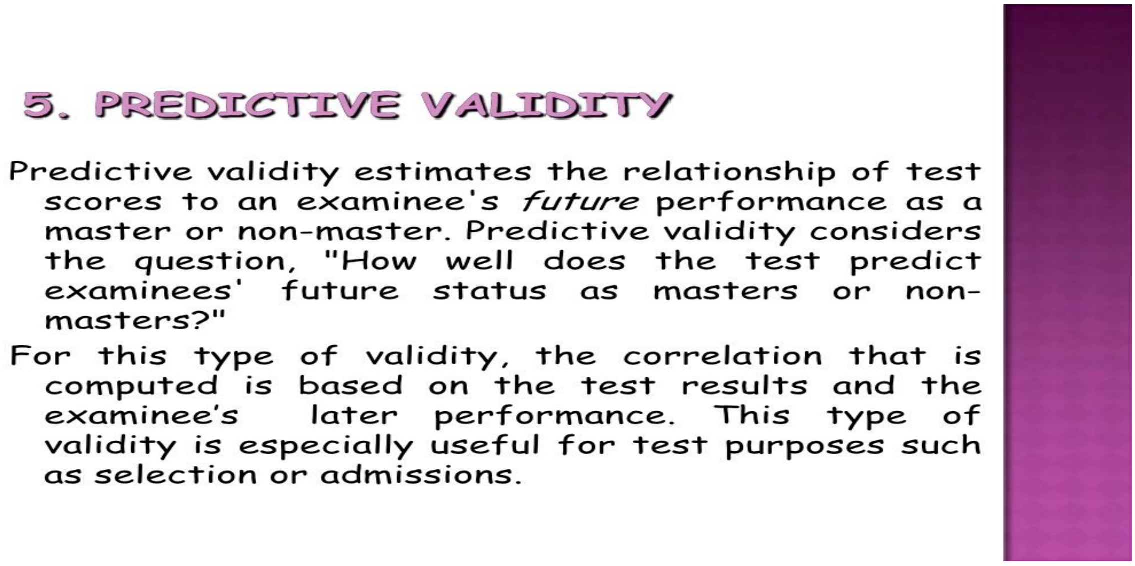
For example, imagine a hands-on driving test has been proved to be an accurate test of driving skills. A written test can be validated by using a criterion related strategy in which the hands-on driving test is compared to it.



4. CONCURRENT VALIDITY

Concurrent validity uses statistical methods of correlation to other measures.

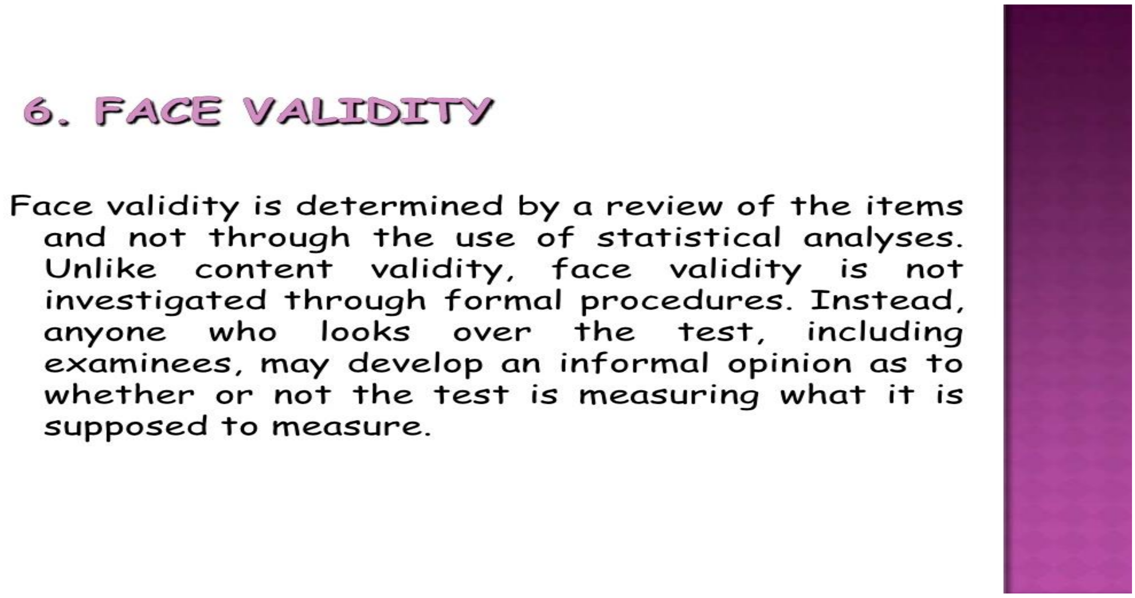
Examinees who are known to be either masters or non-masters on the content measured are identified before the test is administered. Once the tests have been scored, the relationship between the examinees' status as either masters or non-masters and their performance (i.e., pass or fail) is estimated based on the test.



5. PREDICTIVE VALIDITY

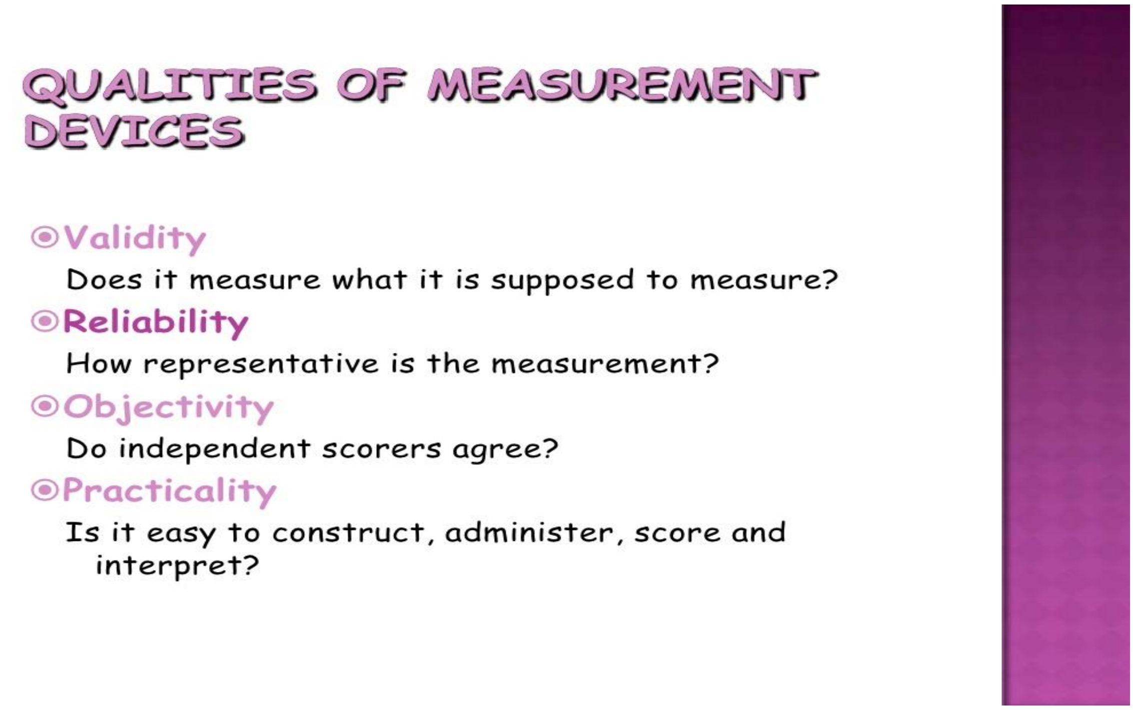
Predictive validity estimates the relationship of test scores to an examinee's *future* performance as a master or non-master. Predictive validity considers the question, "How well does the test predict examinees' future status as masters or non-masters?"

For this type of validity, the correlation that is computed is based on the test results and the examinee's later performance. This type of validity is especially useful for test purposes such as selection or admissions.



6. FACE VALIDITY

Face validity is determined by a review of the items and not through the use of statistical analyses. Unlike content validity, face validity is not investigated through formal procedures. Instead, anyone who looks over the test, including examinees, may develop an informal opinion as to whether or not the test is measuring what it is supposed to measure.



QUALITIES OF MEASUREMENT DEVICES

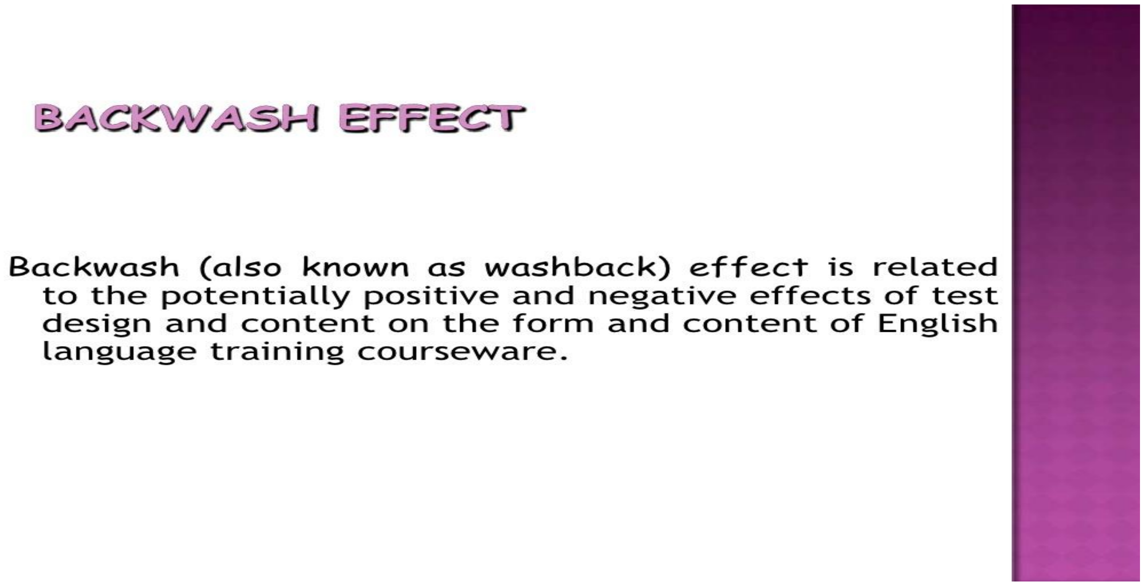
- **Validity**

  Does it measure what it is supposed to measure?
- **Reliability**

  How representative is the measurement?
- **Objectivity**

  Do independent scorers agree?
- **Practicality**

  Is it easy to construct, administer, score and interpret?



BACKWASH EFFECT

Backwash (also known as washback) effect is related to the potentially positive and negative effects of test design and content on the form and content of English language training courseware.

CONCLUSION

These are the three main things you should think about when designing any assessment. These factors should be considered if you are designing something like an exam in school or even if you are designing an assessment for professionals such as a licensing exam.

Comprehension questions:

1. How does validity affect reliability?

2. What is an example of reliability and validity?

3. What is practicality and efficiency in assessment?

4. Why the test must be reliability?

Test

1. … format means that the examinees are given brief notes of a public address and the task is to “unfold” these brief entries into full text.

a. Describing

b.Explaining

c.Contextual

d.Written response

2. …pertains to whether the text measures what it claims to measure

a. Reliability

b. Consistency

c. Construct validity

d. Concurrent validity

3. A ….

a. Breakthrough

b. Waystage

c. Threshold

d. Vantage

4. Can handle very short social exchanges even though they don't understand enough to keep the conversation going themselves…… What level is it?

a. A2

b. A1

c. B2

d. C1

5. CEFR is not….

a. Theoretical document

b. Descriptive document

c. A document tore flex

d. A starting point to develop new tools

6. Forms of formative assessment?

a. oral, written, individual, group

b. homework, projects

c. test, matching, multiplechoice

d. individual, pair and group work, homework tasks, indirect (implicit) form using different questions or plays

7. Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events, but lexical limitations cause repetition and even difficulty with formulation at times. What level is it?

a. A2

b. A1

c. B2

d. B1

8. High reliability, easier to write and limits guessing are advantages of which type of activity?

a. Multiple choice questions

b. Problem solution activity

c. Fill-in-the-Blank

d. Matching

9. How many challenges will be discussed along with solutions to help teachers using direct assessment in language classrooms?

a. 3

b. 4

c. 2

d. 5

10. I can write very simple personal letters expressing thanks and apology. What level is it?

a. A2

b. B1

c. B2

d. A1

THEME: 4. CRITERION AND NORM REFERENCING TEST

Getting started. Learn more.

Plan:

1. The main importance of criterion test.

2. The main features of norm-referenced test.

3. The differences between criterion and norm-referenced test.

What is the Criterion-referenced Test?

A criterion-referenced test is an assessment and test that measures student's performance. Also, these measures the performance of the students alongside fixed criteria. These criteria's include written and brief reports of what students are capable of doing at different stages. In other words, the Criterion reference test is a set of fixed criteria. That used to measure student's performance. Also, these assess the student's performance. Meaning of Criterion-referenced Test

Criterion reference test is a method which uses test score to judge students. Also, they help to generate statements about students' behavior. Also, they use test scores as their reference. Criterion reference mostly uses quizzes. The main objective of this is to check whether students have learned the topic or not. These generally have multiple-choice, true-false, and open-ended questions. They play an important role to take a decision about student's performance.

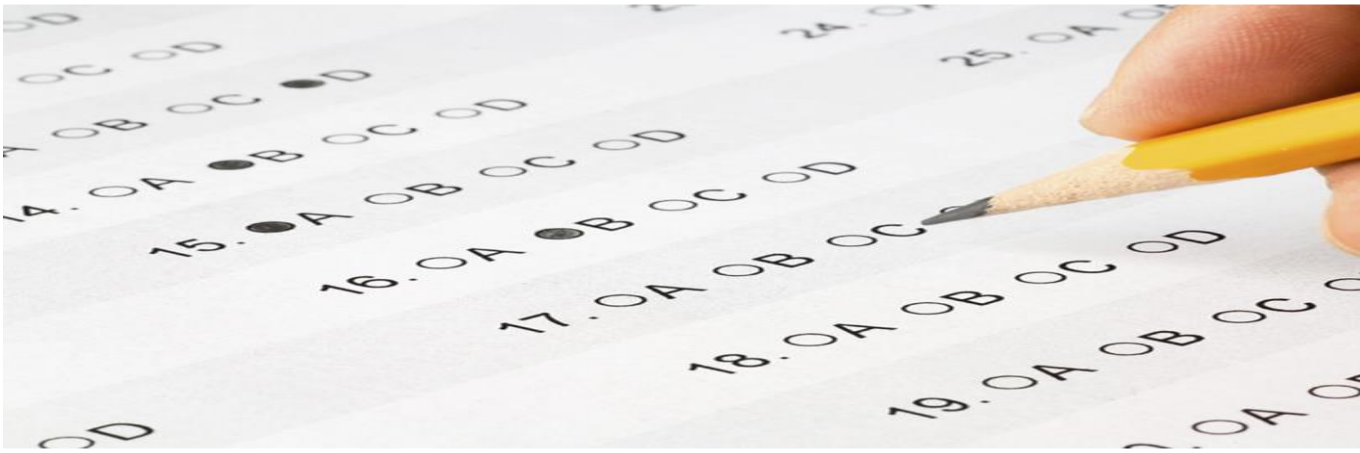
What are Norm-referenced tests?

These test measure student's performance in comparison to other students. Also, the age and question paper is same for both of them. They measure whether the students have performed better or worse than other test takers. It is the theoretical average determined by comparing scores.

Characteristics of the Criterion Reference Test:

- Authority
- Consistency

- Practically
- Subject Mastery
- Managed Locally



1. Authority

It actually assesses whether they measure what it claims or not. An individual item matches with its goal. Also, if the situations and performance specified in the aim signify in the item or not.

2. Consistency

It means that if it always measures what it states. Also, consistency means if they have a high degree of confidence in the scores or not. Any random error in the tool can make it unreliable.

3. Practicality

Not all assessment is reliable because of cost and time. It is not always possible to design reliable and accurate tests. Also, the decision should considerably relate to important factors.

4. Subject Mastery

This help in the pathway the performance of students within the course of study. Also, test items can be made to match precise purposes. Criterion reference test also judges how well the student knows and understand the topic.

5. Managed Locally

Generally, these developed at the classroom level. Also, the teacher can easily check if the standards are met or not. Besides, they also identify shortages. Results of tests are quickly obtained to give students helpful feedback on performance.

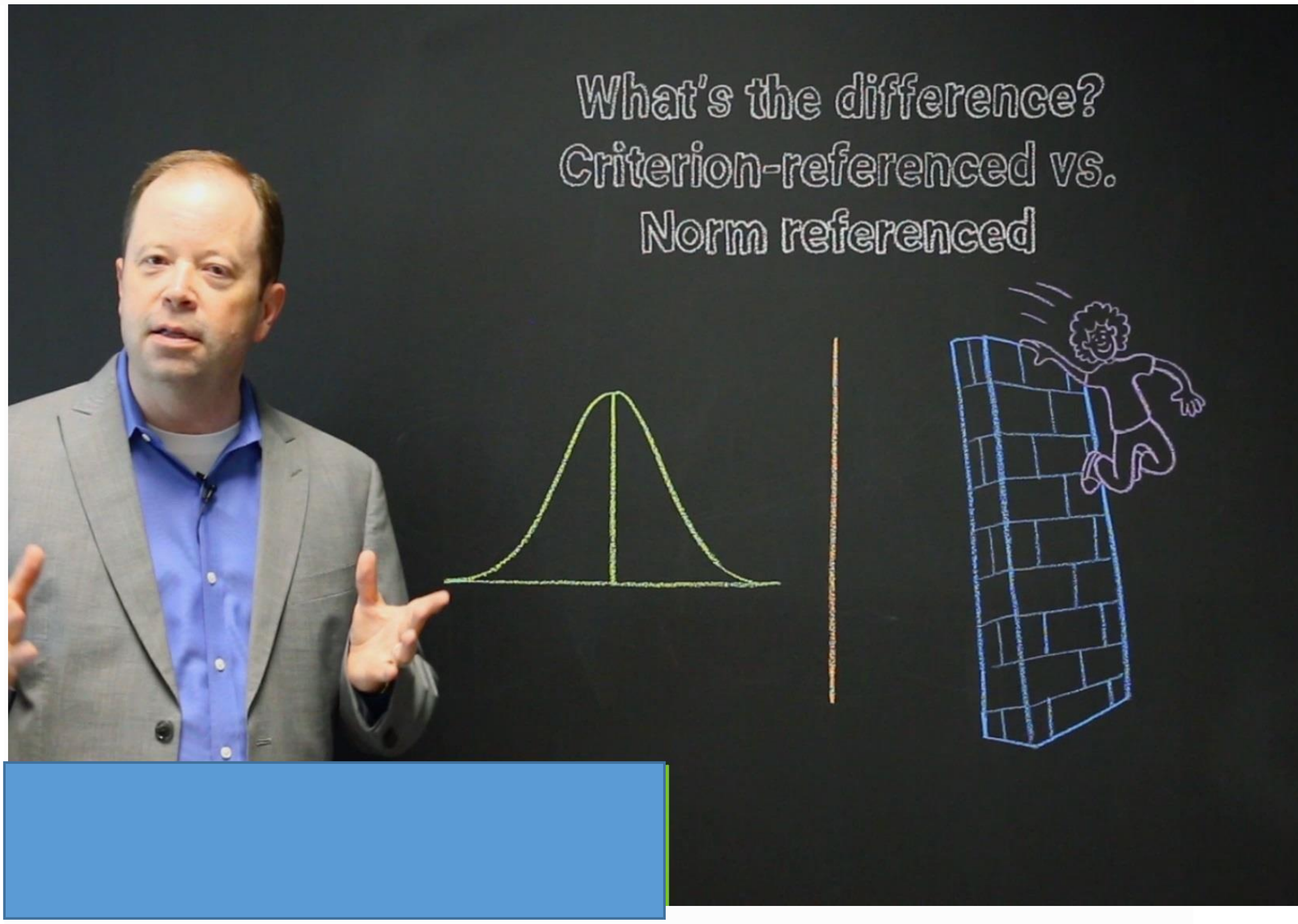
Difference between Norm and Criterion-referenced Test

Basis	**Criterion-Referenced Test**	**Norm-Referenced Test**
Performance	Each student is independently assessed.	Judged on the basis of other student's performance.
Comparison	It does not compare a student's performance with other students.	It compares a student's performance with other students.
Objective	Its main objective is to help students learn without getting questioned about grades.	Its main objective is to assess a student's performance with other students.
Criteria	They have fixed criteria for assessment.	Their criteria changes with outcomes.
Results	Results can be derived quickly.	Takes little time to derive results.
Examples	Clinical skill competency tools.	Class examination.

After going through the characteristics and difference between norm and criterion-referenced test we can conclude the following things. First of all, both are suitable for different tasks. Secondly, both have their own criteria of judgment. And lastly, they follow different norms and values.

Criterion-referenced vs. norm-referenced

To understand what happened, we need to understand the difference between criterion-referenced tests and norm-referenced tests. The first thing to understand is that even an assessment expert couldn't tell the difference between criterion-referenced test and a norm-referenced test just by looking at them. The difference is actually in the scores—and some tests can provide both criterion-referenced results and norm-referenced results!



How to interpret criterion-referenced test?

Criterion-referenced tests compare a person's knowledge or skills against a predetermined standard, learning goal, performance level, or other criterion. With criterion-referenced tests, each person's performance is compared directly to the standard, without considering how other students perform on the test. Criterion-referenced tests often use “cut scores” to place students into categories such as “basic,” “proficient,” and “advanced.”

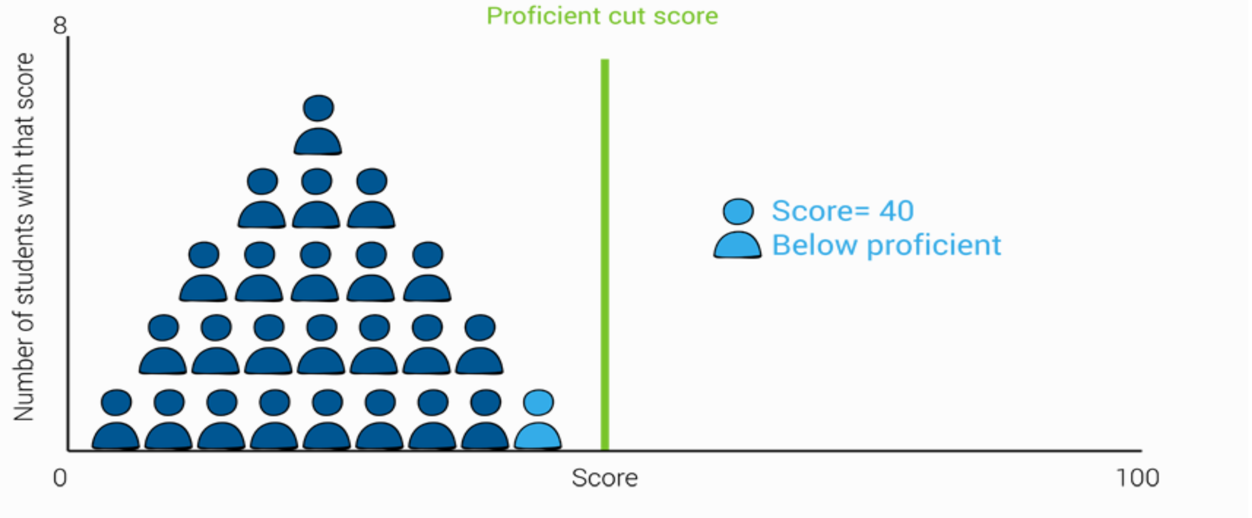
If you’ve ever been to a carnival or amusement park, think about the signs that read “You must be this tall to ride this ride!” with an arrow pointing to a specific line on a height chart. The line indicated by the arrow functions as the criterion; the ride operator compares each person’s height against it before allowing them to get on the ride.

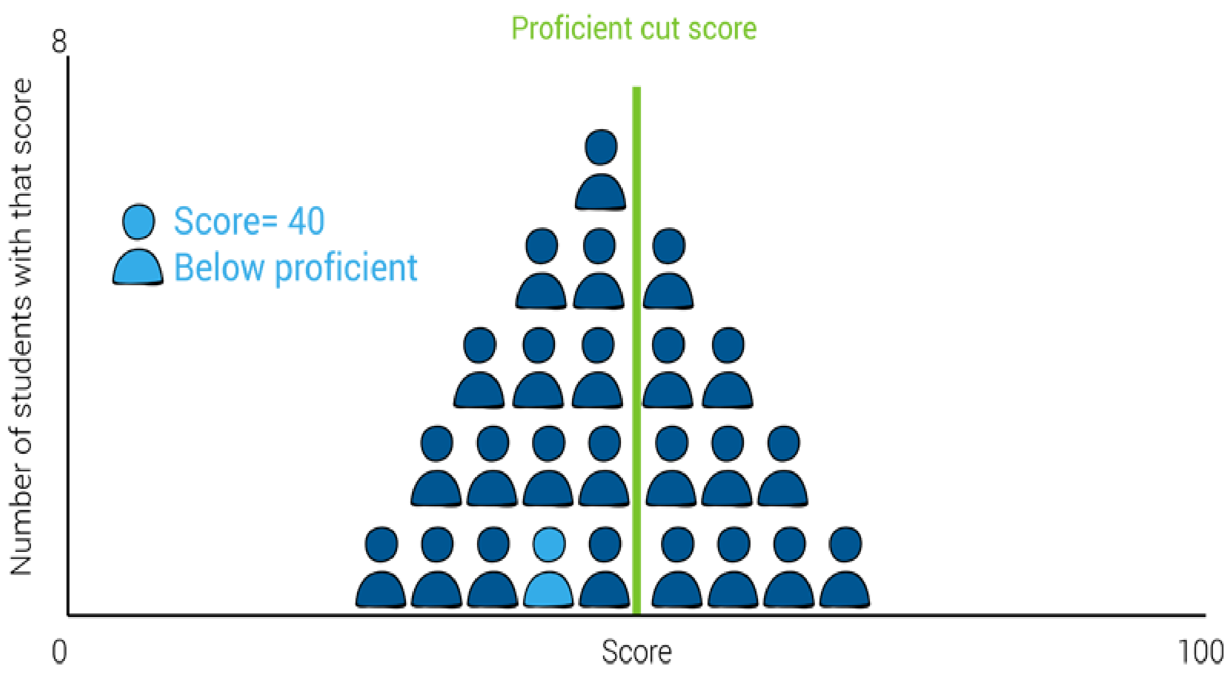
Criterion-referenced tests compare a student’s knowledge and skills against a predetermined standard, cut score, or other criterion.

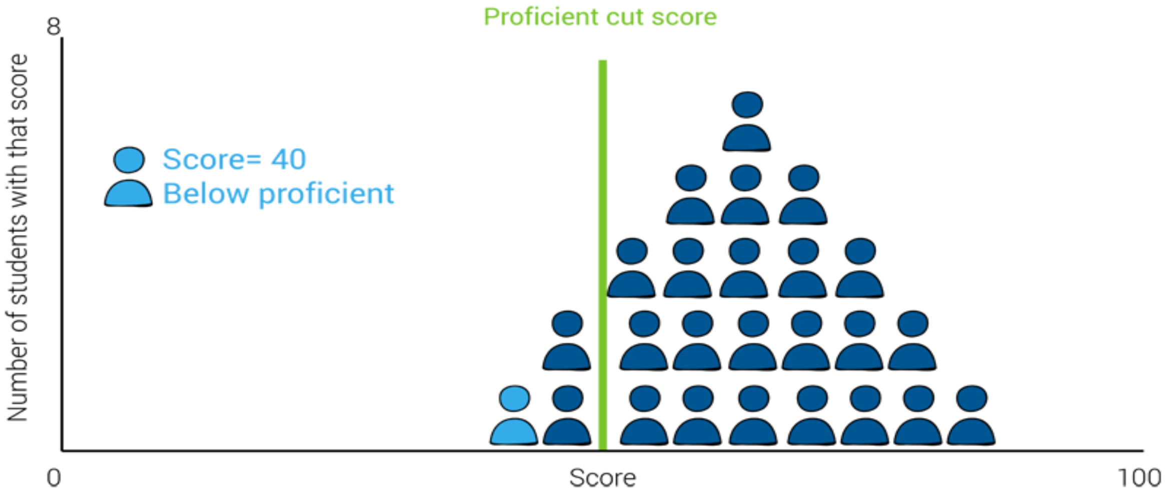
In criterion-referenced tests, the performance of other students does not affect a student’s score.

Note that it doesn’t matter how many other people are in line or how tall or short they are; whether or not you’re allowed to get on the ride is determined solely by your height. Even if you’re the tallest person in line, if the top of your head doesn’t reach the line on the height chart, you can’t ride. Criterion-referenced assessments work similarly: An individual’s score, and how that score is categorized, is not affected by the performance of other students. In the charts below, you can see the student’s score and performance category (“below proficient”) do not change, regardless of whether they are a top-performing student, in the middle, or a low-performing student.

On a criterion-referenced test, an individual student's score is not affected by the performance of their peers.





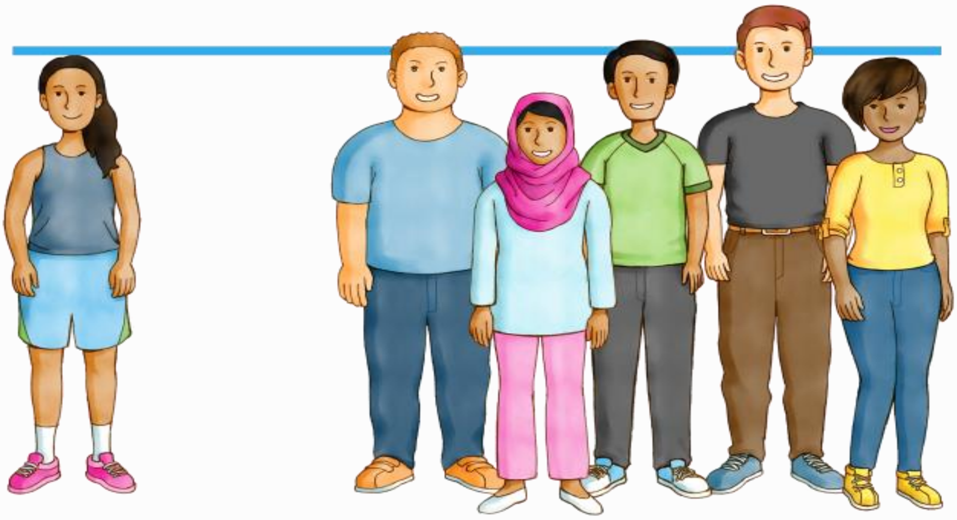


This means knowing a student’s score for a criterion-referenced test will only tell you how that specific student compared in relation to the criterion, but not whether they performed below-average, above-average, or average when compared to their peers.

How to interpret norm-referenced tests?

Norm-referenced measures compare a person’s knowledge or skills to the knowledge or skills of the norm group. The composition of the norm group depends on the assessment. For student assessments, the norm group is often a nationally representative sample of several thousand students in the same grade (and sometimes, at the same point in the school year). Norm groups may also be further narrowed by age, English Language Learner (ELL) status, socioeconomic level, race/ethnicity, or many other characteristics.

Norm-referenced tests compare a student’s performance against the performance of their peers.

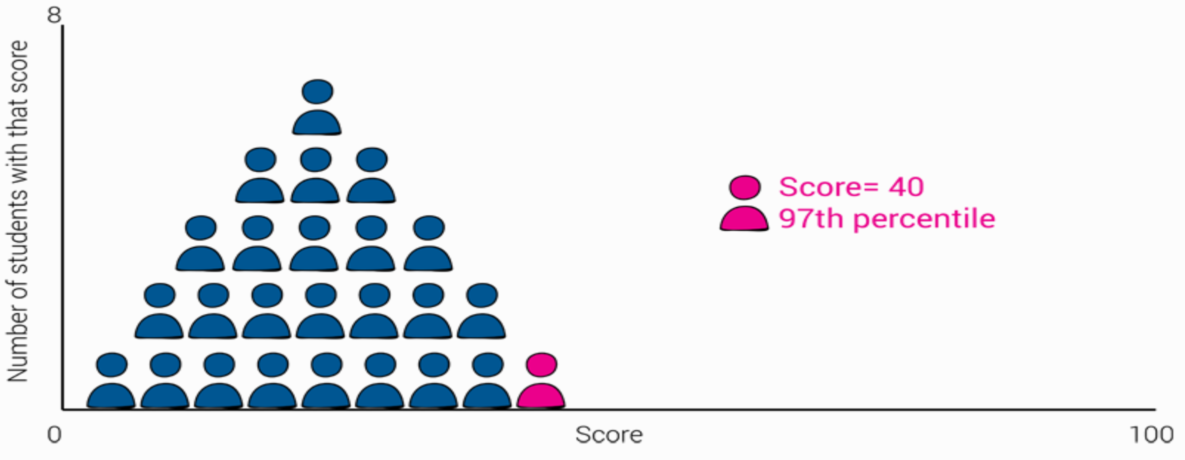


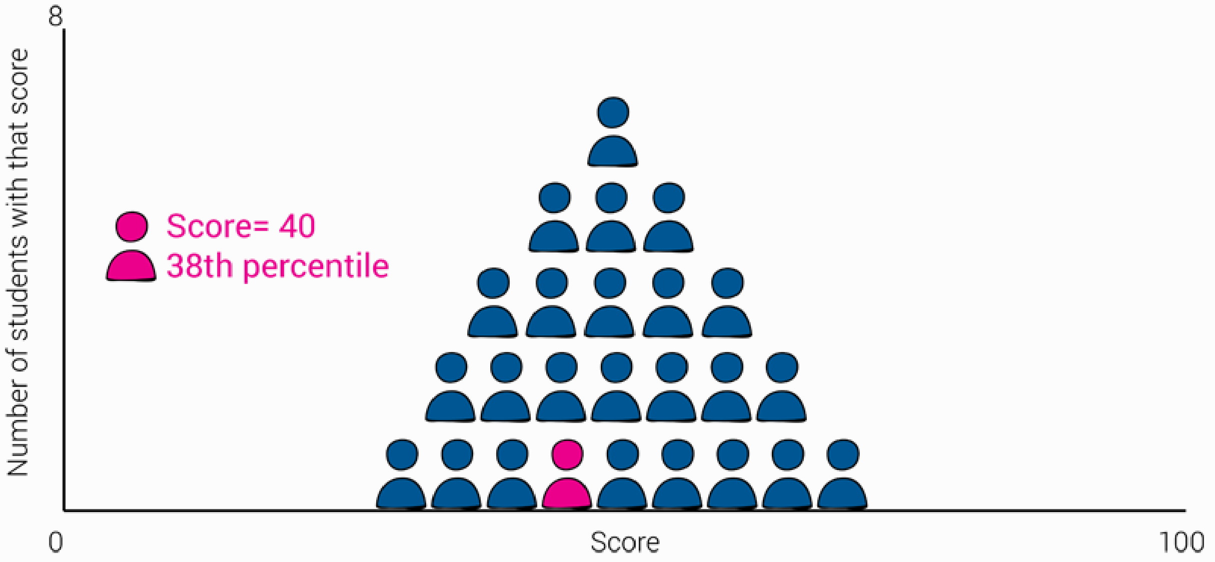
One norm-referenced measure that many families are familiar with is the baby weight growth charts in the pediatrician’s office, which show which percentile a child’s weight falls in. A child in the 50th percentile has an average weight; a child in the 75th percentile weighs *more* than 75% of the babies in the norm group and *the same as or less* than the heaviest 25% of babies in the norm group; and a child in the 25th percentile weighs *more* than 25% of the babies in the norm group and *the same as or less* than 75% of them. It’s important to note that these norm-referenced measures do not say whether a baby's birth weight is "healthy" or "unhealthy," only how it compares with the norm group.

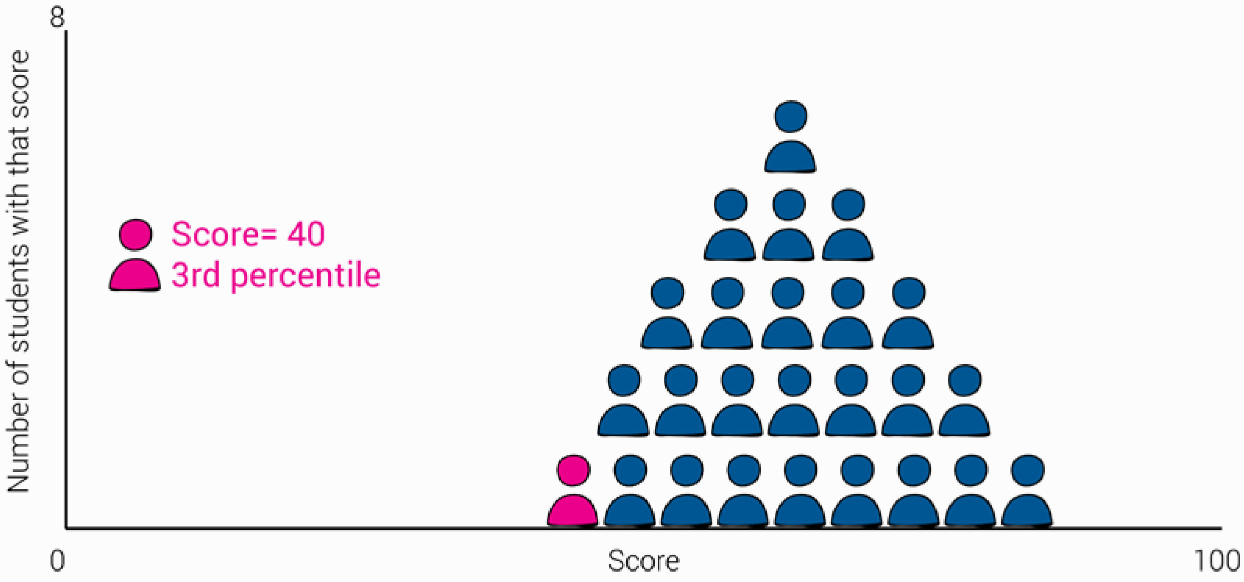
For example, a baby who weighed 2,600 grams at birth would be in the 7th percentile, weighing the same as or less than 93% of the babies in the norm group. However, despite the very low percentile, 2,600 grams is classified as a normal or healthy weight for babies born in the United States—a birth weight of 2,500 grams is the cut-off, or criterion, for a child to be considered low weight or at risk. (For the curious, 2,600 grams is about 5 pounds and 12 ounces.) Thus, knowing a baby's percentile rank for weight can tell you how they compare with their peers, but not if the baby's weight is "healthy" or "unhealthy." Norm-referenced assessments work similarly: An individual student's percentile rank describes their performance in comparison to the performance of students in the norm group, but does not indicate whether or not they met or exceed a specific standard or criterion.

In the charts below, you can see that, while the student's score doesn't change, their percentile rank *does* change depending on how well the students in the norm group performed. When the individual is a top-performing student, they have a high percentile rank; when they are a low-performing student, they have a low percentile rank. What we can't tell from these charts is whether or not the student should be categorized as proficient or below proficient.

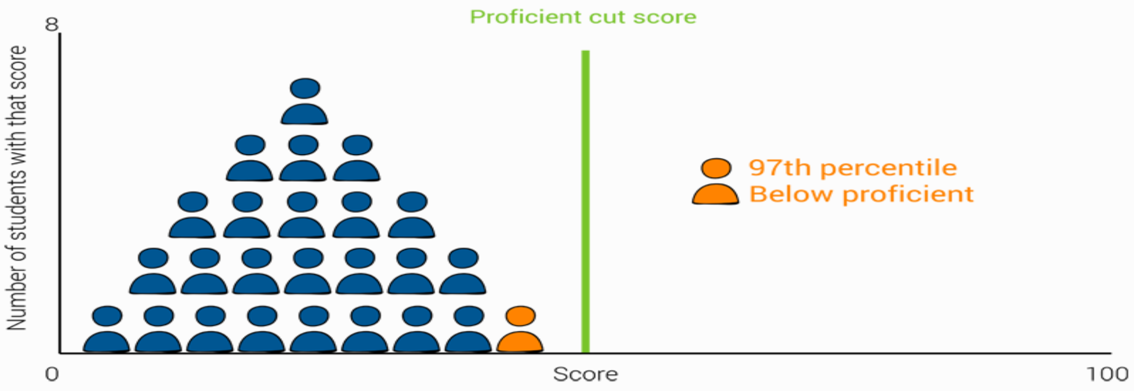
On a norm-referenced test, an individual student's percentile rank is calculated according to the performance of their peers.





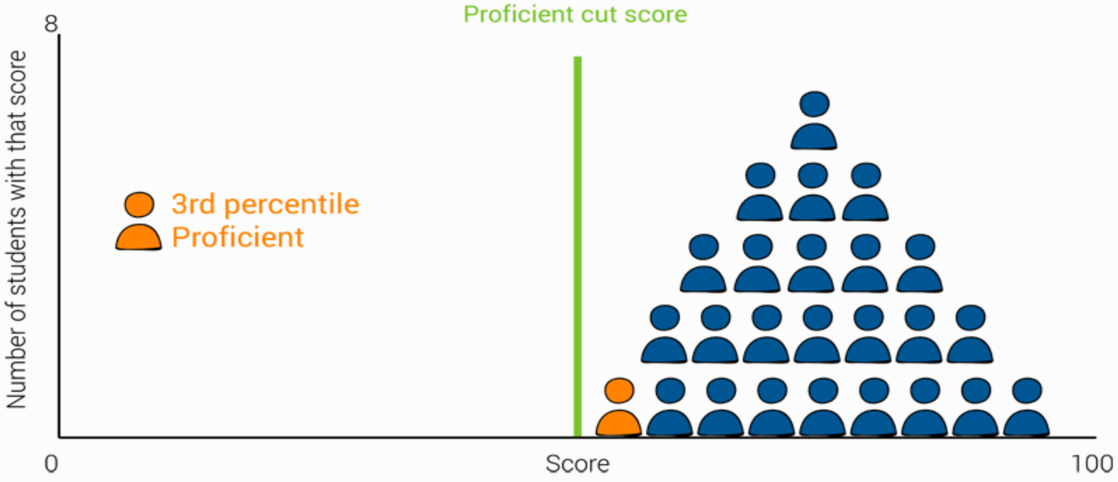


A student can have a high percentile rank but not achieve proficiency.



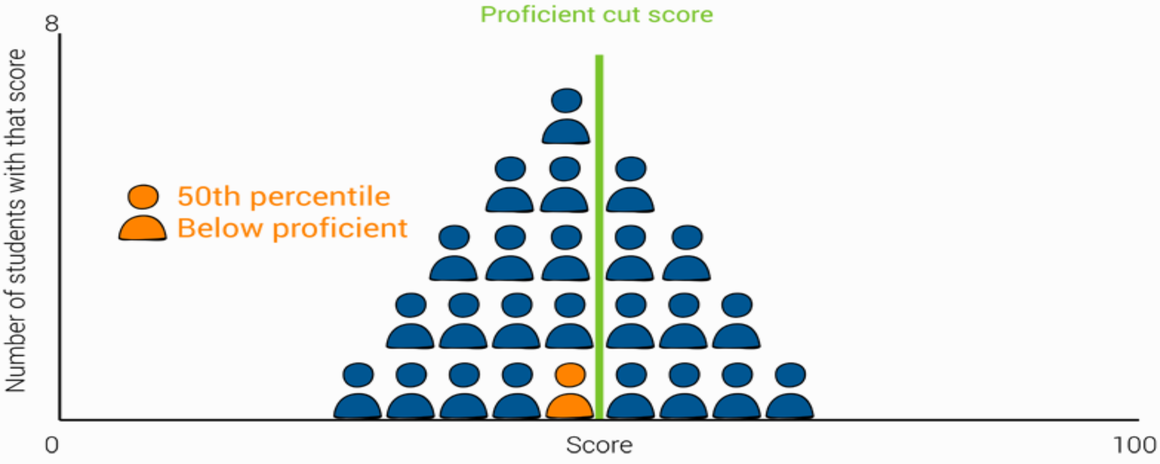
The opposite is also possible. A student could have a very low percentile rank, but still meet the criterion for proficiency. Is this student doing poorly, because they aren’t performing as well as their peers, or are they doing well, because they’ve achieved proficiency?

A student can have a low percentile rank but still achieve proficiency.



However, these are fairly extreme and rather unlikely cases. Perhaps more common is a “typical” or “average” student who does not achieve proficiency because the majority of students are not achieving proficiency. In fact, this is the pattern we see with National Assessment of Educational Progress (NAEP) scores, where the “typical” fourth-grade student (50th percentile) has a score of 226 and the “average” fourth-grade student (average of all student scores) has a score of 222, but proficiency requires a score of 238 or higher.

A student may be “typical” according to their norm-referenced measures but not achieve proficiency according to criterion-referenced measures.



In all of these cases, educators must use their professional judgement, knowledge of the student, familiarity with standards and expectations, understanding of available resources, and subject-area expertise to determine the best course of action for each individual student. The assessments — and the data they produce—merely provide information that the educator can use to help inform decisions.

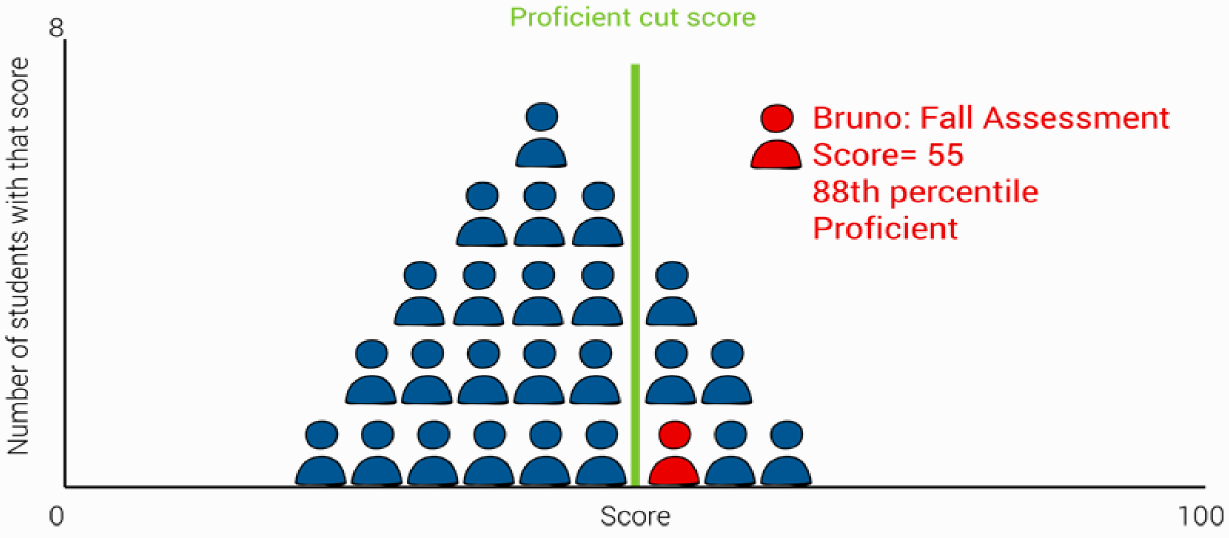
For example,

What happened to Bruno?

So what happened to Bruno in the scenario described at the beginning of this post?

In the fall, Bruno scored 55 out of 100 on his district's assessment. The district had set the cut-score for proficiency at 50, meaning that Bruno counts as "proficient." The district's assessment provider compared Bruno's score of 55 to the fall scores of their norm group, and found that Bruno scored higher than 88% percent of his peers in the norming group. This gives him a percentile rank of 88.

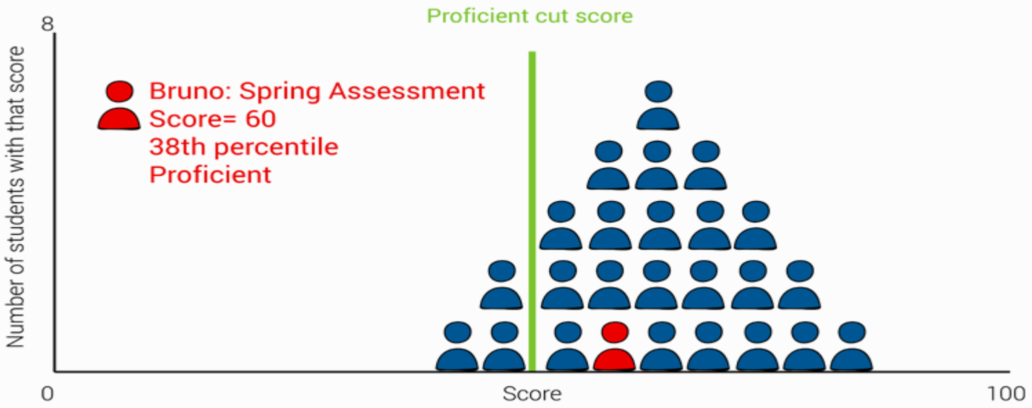
In the fall, Bruno scored 55 on the district assessment, putting him in the 88th percentile. He is categorized as proficient.



In the spring, Bruno takes the same test again. This time he scores 60, higher than this fall score. Since the district's criterion for proficiency hasn't changed, he is still categorized as proficient.

Just like Bruno, students in the norm group took the assessment twice—once in the fall and once in the spring. This time, the district's assessment provider compares Bruno's spring score to the spring scores of the norm group. In this case, the students in the norm group had notable gains and scored much higher in the spring than they did in the fall. Because students in the norm group generally had much larger gains from fall to spring than Bruno did, Bruno's spring score now puts him at the 38th percentile.

In the spring, Bruno scored 60 on the district assessment, putting him in the 38th percentile. He is categorized as proficient.



For Bruno's teacher, this is a sign of concern. Although Bruno is still categorized as proficient, he's not keeping up with his peers and may be at risk of falling behind in future years. In addition, if the district or state raises the criterion for proficiency—which can happen when standards or assessments change—he might fall short of that new criterion and struggle to make enough gains in one year to meet more rigorous expectations.

This is one reason why it's important for educators to monitor *growth* in addition to *gains*.

Comprehension question:

1. *What is the difference between criterion and norm referenced tests?*
2. *What are examples of criterion referenced tests?*
3. *What are norm and criterion referenced measures and when is each appropriate?*
4. *What is a criterion referenced score?*

THEME: 5. PROFICIENCY, ACHIEVEMENT AND PROGRESS TEST

Getting started. Learn more.

Plan:

1. The main importance of proficiency test.
2. The main features of achievement test.
3. The differences between achievement and progress test.

Test construction is a matter of problem solving, every teaching situation sets a different testing problem. In order to arrive at the best solution for any particular problem it is important to choose the most appropriate test or testing system. In our work we use four types of tests: proficiency tests, achievement tests, diagnostic tests, and placement tests. Proficiency tests Proficiency tests are designed to measure people's ability in a language, regardless of any training they may have had in that language. The content of a proficiency test, therefore, is not based on the content or objectives of language courses that people taking the test may have followed. It is based on a specification of what candidates have to be able to do in the language in order to be considered proficient, it means having sufficient command of the language for a particular purpose. The function of such tests is to show whether candidates have reached a certain standard with respect to a set of specified abilities. Though there is no particular purpose in mind for the language, these general proficiency tests should have detailed specifications saying just what it is that successful candidates have demonstrated that they can do. Each test should be seen to be based directly on these specifications. All users of a test (teachers, students, employers, etc.) can then judge whether the test is suitable for them, and can interpret test results. Despite differences between them of content and level of difficulty, all proficiency tests have in common the fact that they are not based on courses that candidates may have previously taken.

Achievement tests In contrast to proficiency tests, achievement tests are directly related to language courses, their purpose being to establish how successful individual students, groups of students, or the courses themselves have been in achieving objectives. They are of two kinds: final achievement tests and progress achievement tests. Final achievement tests are those administered at the end of a course of study. They may be written and administered by ministries of education, official examining boards, or by members of teaching institutions. Clearly the content of these tests must be related to the courses with which they are concerned. In the view of some testers, the content of a final achievement test should be based directly on a detailed course syllabus or on the books and other materials used. This has been referred to as the syllabuscontent approach. It has an obvious appeal, since the test only contains what it is thought that the students have actually encountered, and thus can be considered, in this respect at least, a fair test. The disadvantage is that if the syllabus is badly designed, or the books and other materials are badly chosen, the results of a test can be very misleading. Successful performance on the test may not truly indicate successful achievement of course objectives. For example, a course may have as an objective the development of conversational ability, but the course itself and the test may require students only to utter carefully prepared statements about their home town, the weather, or whatever. Another course may aim to develop a reading ability in English, but the test may limit itself to the vocabulary the students are known to have met. In each of these examples test results will fail to show what students have achieved in terms of course objectives. The alternative approach is to base the test content directly on the objectives of the course. This has a number of advantages. First, it compels course designers to be explicit about objectives. Secondly, it makes it possible for performance on the test to show just how far students have achieved those objectives. This in turn puts pressure on those responsible for the syllabus and for the selection of books and materials to ensure that these are consistent with the course objectives. Tests based on objectives work against the perpetuation of poor teaching practice, something which course-content-based tests, almost as if part of a conspiracy, fail to do. They will provide more accurate information about individual and group achievement, and it is likely to promote a more beneficial backwash effect on teaching. It might be argued that to base test content on objectives rather than on course content is unfair to students. If the course content does not fit well with objectives, they will be expected to do things for which they have not been prepared. In a sense this is true. But in another sense it is not. If a test is based on the content of a poor or inappropriate course, the students taking it will be misled as to the extent of their achievement and the quality of the course. Whereas if the test is based on objectives, not only will the information it gives be more useful, but there is less chance of the course surviving in its present unsatisfactory form. Initially some students may suffer, but future students will benefit from the pressure for change. The long term interests of students are best served by final achievement tests whose content is based on course objectives. Is there any real difference between final achievement tests and proficiency tests? If a test is based on the objectives of a course, and these are equivalent to language needs on which a proficiency test is based, there is no reason to expect a difference between the form and content of the two tests. Two things have to be remembered, however. First, objectives and needs will not typically coincide in this way. Secondly, many achievement tests are not in fact based on course objectives. Progress achievement tests, as their name suggests, are intended to measure the progress that students are making. They contribute to formative assessment. There should be established a series of well-defined short-term, objectives, progress tests based on short-term objectives will fit well with what has been taught. In addition to more formal achievement tests that require careful preparation, teachers should feel free to set their own ‘pop quizzes’. These serve both to make a rough check on students’ progress and to keep students on their toes. Diagnostic tests Diagnostic tests are used to identify learners’ strengths and weaknesses. They are intended primarily to ascertain what learning still needs to take place. These tests will help to see who is particularly weak in, say, speaking as opposed to reading in a language. We may be able to analyze samples of a person’s in writing or speaking in order to create profiles of the student’s ability with respect to such categories as ‘grammatical accuracy’ or ‘linguistic appropriacy’. Diagnostic tests are extremely useful for individualized instruction or self-instruction. Learners can be shown where gaps exist in their command of the language and can be directed to sources of information, exemplification practice. Placement tests Placement tests, as their name suggests, are intended to provide information that will help to place students at the stage (or in the part) or the teaching program most appropriate to their abilities.

Typically, they are used to assign students to classes at different levels. Placement tests that are most successful are those constructed for particular situations. Direct versus indirect testing is said to be direct when it requires the candidate to perform precisely the skills that we wish to measure. If we want to know how well the candidates can write compositions, we get them to write compositions. If we want to know how well the candidates pronounce a language, we get them to speak. The tasks and the texts that are used, should be as authentic as possible. Direct testing is easier to carry out when it is intended to measure the productive skills of speaking and writing. Direct testing has a number of attractions. First, provided that we are clear about what abilities we want to assess, it is relatively straightforward to create the conditions which will elicit the behavior on which to base our judgments. Secondly, at least in the case of the productive skills, the assessment and interpretation of students' performance is also quite straightforward. Thirdly, since practice for the test involves practice of the skills that we wish to foster, there is likely a helpful backwash effect. Indirect testing attempts to measure the abilities that underlie the skills in which we are interested. Perhaps the main appeal of indirect testing is that it seems to offer the possibility of testing a representative sample of a finite number of abilities which underlie a potentially indefinite large number of manifestations of them. The main problem with indirect tests is that the relationship between performance on them and performance of the skills in which we are usually more interested tends to be rather weak in strength and uncertain in nature. We do not yet know enough about the component of, say, composition writing to predict accurately composition writing ability from scores on tests that measure the abilities that we believe underlie it. We may construct tests of grammar, vocabulary, discourse markers, handwriting, punctuation, and what we will. But we will not be able to predict accurately scores on compositions (even if we make sure of the validity of the composition scores by having people write many compositions and by scoring these in a valid and reliable way). As far as proficiency and final achievement tests are concerned, it is preferable to rely on direct testing. Of course, to obtain diagnostic information on abilities, such as control of particular grammatical structures, indirect testing may be perfectly appropriate. Some tests are referred as semi-direct. The most obvious examples of these are speaking tests where candidates respond to tape recorded stimuli, with their own responses being recorded and later scored. These tests are semi-direct in the sense that, although not direct, they simulate direct testing. Discrete point versus integrative testing Discrete point testing refers to the testing of one element at a time, item by item. Integrative testing, by contrast, requires candidate to combine many language elements in the completion of the task. This might involve writing a composition, making notes while listening to a lecture, taking a dictation, or completing a cloze passage. Discrete point tests will almost always be indirect, while integrative tests will tend to be direct. Objective testing versus subjective testing The distinction here is between methods of scoring. If no judgment is required on the part of the scorer, then the scoring is objective. A multiple choice test, with the correct responses unambiguously identified, would be a case in point. If judgment is called for, the scoring is said to be subjective. There are different degrees of subjectivity in testing. The impressionistic scoring of a composition may be considered more subjective than the scoring of short answers in response to questions on a reading passage. Computer adaptive testing in most paper and pencil tests, the candidate is presented with all the items, usually in ascending order of difficulty, and is required to respond to as many of them as possible. This is not the most economical collecting information on someone’s ability. People of high ability (in relation to the test as a whole) will spend time responding to items that are very easy for them - all, or nearly all, of which they will get correct. We would have been able to predict their performance on this items from their correct response to more difficult items. Similarly, we could predict the performance of people of low ability on difficult items, simply by seeing their consistently incorrect response to easy items.

Computer adaptive testing offers a potentially more efficient way of collecting information on people's ability. All candidates are presented with an item of average difficulty. Those who respond correctly are presented with a more difficult item; those who respond incorrectly are presented with an easier item. The computer goes on in this way to present individual candidates with items that are appropriate for their apparent level of ability raising or lowering the level of difficulty until a dependable estimate of their ability is achieved. Oral interviews are typically a form of adaptive testing, with the interviewer's prompts and language being adapted to the apparent level of the candidate.

Comprehension question:

1. *What is the difference between criterion and norm referenced tests?*
2. *What are examples of criterion referenced tests?*
3. *What are norm and criterion referenced measures and when is each appropriate?*
4. *What is a criterion referenced score?*

Test

1. Subjective tests are…

a. *Tests in which the absence of predetermined or absolutely correct responses require the judgment of the teacher to determine correct and incorrect Answers*

b. *Assessments that involve learners in actually performing the behavior that one purports to measure*

c. *Tests that aim to measure, or summarize, what a student has grasped and typically occurs at the end of a course or unit of instruction*

d. *Test that are not limited to any one course, curriculum, or single skill in the language; rather, it tests overall global ability*

5. What is validity?

a. *The extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment*

b. *Statements that describe what a student can perform at a particular point on a rating scale; sometimes also called band descriptors*

c. *The effect of assessments on classroom teaching and learning*

d. *The extent to which resources and time available to design, develop, and administer a test are manageable and feasible*

3.What is integrative test?

a. *A test that treats language competence as a unified set of interacting abilities of grammar, vocabulary, reading, writing, speaking, and listening*

b. *The extent to which the linguistic criteria of the test (e.g., specified classroom objectives) are measured and implied predetermined levels of performance are actually reached*

c. *A test in which the absence of predetermined or absolutely correct responses require the judgment of the teacher to determine correct and incorrect answers*

d. *A test that is not limited to any one course, curriculum, or single skill in the language; rather, it tests overall global ability*

4. How many principles should be taken into consideration in all test specifications?

a. *There are 7 principles in test development: purpose of the test; learners' age and level; language skills; language elements; target language situation; text types; tasks.*

b. *There are 6 principles in test development: purpose of the test;; language skills; language elements; target language situation; text types; tasks.*

c. *There are 5 principles in test development: purpose of the test;;language elements; target language situation; text types; tasks.*

d. *There are 4 principles in test development: purpose of the test; target language situation; text types; tasks.*

5.How can newly create assessment tool checked?

a. The best way to find out whether the created assessment tool works well is piloting

b. It can be checked after the announcement of learners' results

c. The best way to find out whether the created assessment tool works well is using it many times

d. It can be checked by survey and questioner

6.How can newly be created assessment tool checked?

a. *The best way to find out whether the created assessment tool works well is piloting*

b. *It can be checked after the announcement of learners' results*

c. *The best way to find out whether the created assessment tool works well is using it many times*

d. *It can be checked by survey and questioner*

THEME: 6.THE COMMON EUROPEAN FRAMEWORK OF REFERENCE (CEFR)

Getting started. Learn more.

Plan:

1. The main importance of CEFR in testing students.
2. The main features of achievement test.
3. The differences between achievement and progress test.

4.How should the CEFR be used by recognizing institutions wishing to set language ability requirements?

Test users may find the Council of Europe's Common European Framework of Reference for Languages (CEFR) helpful. The framework is a series of descriptions of abilities at different learning levels that can be applied to any language. It can provide a starting point for interpreting and comparing different language qualifications and is increasingly used as a way of benchmarking language ability around the world.

IELTS and the CEFR

To help test users understand the relationship between IELTS band scores and the six CEFR levels, Cambridge Assessment English has conducted several studies to map the IELTS 9-band scale to the CEFR, drawing on the interrelationship between IELTS and other Cambridge Assessment English qualifications and the known relationship of these latter qualifications to the CEFR.

In fulfilling its purpose as a common reference tool, the CEFR was not designed to provide the basis for precise equating, nor was it intended to be a prescriptive tool to impose standardized solutions. Rather it was designed as a common framework of reference, primarily intended as ‘a tool for reflection, communications and empowerment’, as described by John Trim, its coordinating author (Saville, N 2005).

Therefore, we would recommend that all recognizing institutions should look at the IELTS band score descriptors and use the IELTS Scores Guide DVD to ascertain the appropriate level of language ability required for their institution or course.

General information

Making comparisons between scores on different tests is challenging because many of the current range of test products differ in their design, purpose, and format (Taylor, 2004a). Test takers' aptitude and preparation for a particular type of test may also vary and individual test takers or groups of test takers may perform better in certain tests than in others.

Specifying the relationship between a test product and the CEFR is challenging because, in order to function as a framework, the CEFR is deliberately underspecified (Davidson & Fulcher, 2007; Milanovic, 2009; Weir, 2005). Establishing the relationship is also not a one-off activity, but rather involves the accumulation of evidence over time (e.g. it needs to be shown that test quality and standards are maintained).

Cambridge Assessment English has been working since the 1990s to refine its understanding of the relationship between its different assessment products, including IELTS, and the CEFR. The relationship of IELTS with the CEFR is complex as IELTS is not a level-based test, but rather designed to span a much broader proficiency continuum. It also utilises a different 9-point band scoring system; thus, there will not be a one-to-one correspondence between IELTS scores and CEFR levels. It is important to bear in mind the differences in test purpose, test format, test populations, and measurement scales when seeking to make comparisons.

With the above in mind, Cambridge Assessment English has conducted a number of research projects since the late 1990s to explore how IELTS band scores align with the CEFR levels. A number of these were summarised in Taylor (2004b), while cautioning that, “As we grow in our understanding of the relationship between IELTS and the CEFR levels, so the frame of reference may need to be revised accordingly.

Note that the IELTS band scores referred to in the figure are the overall band scores, not the individual module band scores for Listening, Reading, Writing and Speaking. It is important to recognise that the purpose of this figure is to communicate the relationship between IELTS and the CEFR. They should not be interpreted as reflecting strong claims about exact equivalence between assessment products or the scores they generate, for the reasons given in Taylor (2004a).

The current alignment is based upon a growing body of internal and external research, some of which has also appeared in peer-reviewed academic journals, attesting to their quality (e.g. Hawkey & Barker, 2004; Lim, Geranpayeh, Khalifa & Buckendahl, 2013). This research has been further combined with long established experience of test use within education and society, as well as feedback from a range of stakeholders regarding the uses of test results for particular purposes.

As further work, such as that being undertaken in the English Profile project, enriches our understanding of the CEFR levels, further refinements may be possible.

Questions about IELTS and the CEFR

1.Has the IELTS test changed?

2.Why is IELTS changing the way the band scores relate to the CEFR?

We have always been committed to providing ongoing revision as we grow in our understanding of the relationship between IELTS, other examinations and the CEFR levels.

The CEFR is becoming more prominent in how institutions consider language ability requirements. It is important therefore that we provide updated advice as to how to interpret IELTS scores in CEFR terms. The table previously on the website did not show half-band scores, and predated the introduction of half-band reporting for Writing and Speaking in July 2007.

3.Has IELTS been made more difficult?

The way it is examined and the way band scores are awarded remain the same.

4. Should institutions and organizations which use IELTS scores change the band scores they expect students to achieve as a result of the revised CEFR mapping?

The test has not changed and the performance represented by each band score remains the same. The IELTS Scores Explained provides samples of those performances so that institutions can judge what level is appropriate to their needs. There is no need for institutions to make changes where they have previously been satisfied with their particular score requirements.

5. How should institutions and organizations interpret this?

As IELTS preceded the CEFR, IELTS band scores have never aligned exactly with the CEFR transition points. The new table makes this clearer. Previously we provided advice as to the score on IELTS that a test taker who was at a given CEFR level might achieve. However, our research shows that a C1 minimum threshold would fall between the 6.5 and 7 bands on the IELTS scale. Therefore, whilst many 6.5 test takers would be at C1, a number will be marginally below. So if an institution requires a high degree of confidence that an applicant is at C1, they may wish to set a requirement of 7, rather than 6.5.

6. Does IELTS differentiate at C2 level?

Band scores of 8.5 and higher are recognised as C2. Band 8 is borderline.

7. If a student already has an IELTS score of 6.5, shown as C1 in the previous mapping, should this now be treated as a B2 equivalent score?

The score 6.5 is borderline B2/C1. The real-world level of performance represented by the result has not changed. It is for institutions to decide whether they wish to change their requirements, if alignment to a particular level of the CEFR is critical (see response to q5 above). The advice in the IELTS Guide for Educational Institutions as to probable levels required for different types of course still holds.

8. Should institutions and organizations that offer English courses to prepare students for university study, or to facilitate university study, change the format, content or level of their courses?

Nothing within the test content has changed.

9. What is the research behind these new mappings?

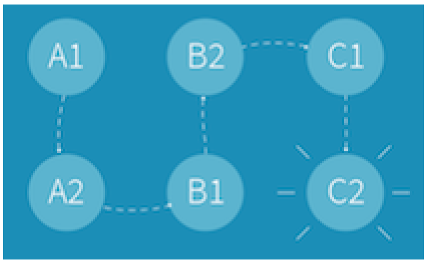
This is a response to the increased prominence of the CEFR in how institutions consider language ability requirements, rather than the findings of a particular research project. The new presentation draws on the previous evidence, on benchmarking exercises conducted in 2009, and on studies of the performance of test takers for other Cambridge English exams at B2 and C1 level on IELTS-type materials in 2009 and 2010.

10. How does this compare to the mappings that other language testers have published?

We do not comment on the benchmarking exercises that other language testers have provided.

English language levels (CEFR)

CEFR standard (Common European Framework of Reference for Languages). The six reference English levels are widely accepted as the global standard for grading an individual's language proficiency.



CEFR English levels are used by all modern English language books and English language schools. It is recommended to use CEFR levels in job resumes (curriculum vitae, CV, Europass CV) and other English levels references. We list here the CEFR descriptors for language proficiency level with the approximate equivalent to other global English evaluation schemes- Cambridge ESOL, Canadian Language Benchmarks / Canadian English Language Proficiency Index Program (CLB/CELPIP), Canadian Academic English Language Assessment (CAEL), BULATS, IELTS and TOEFL.

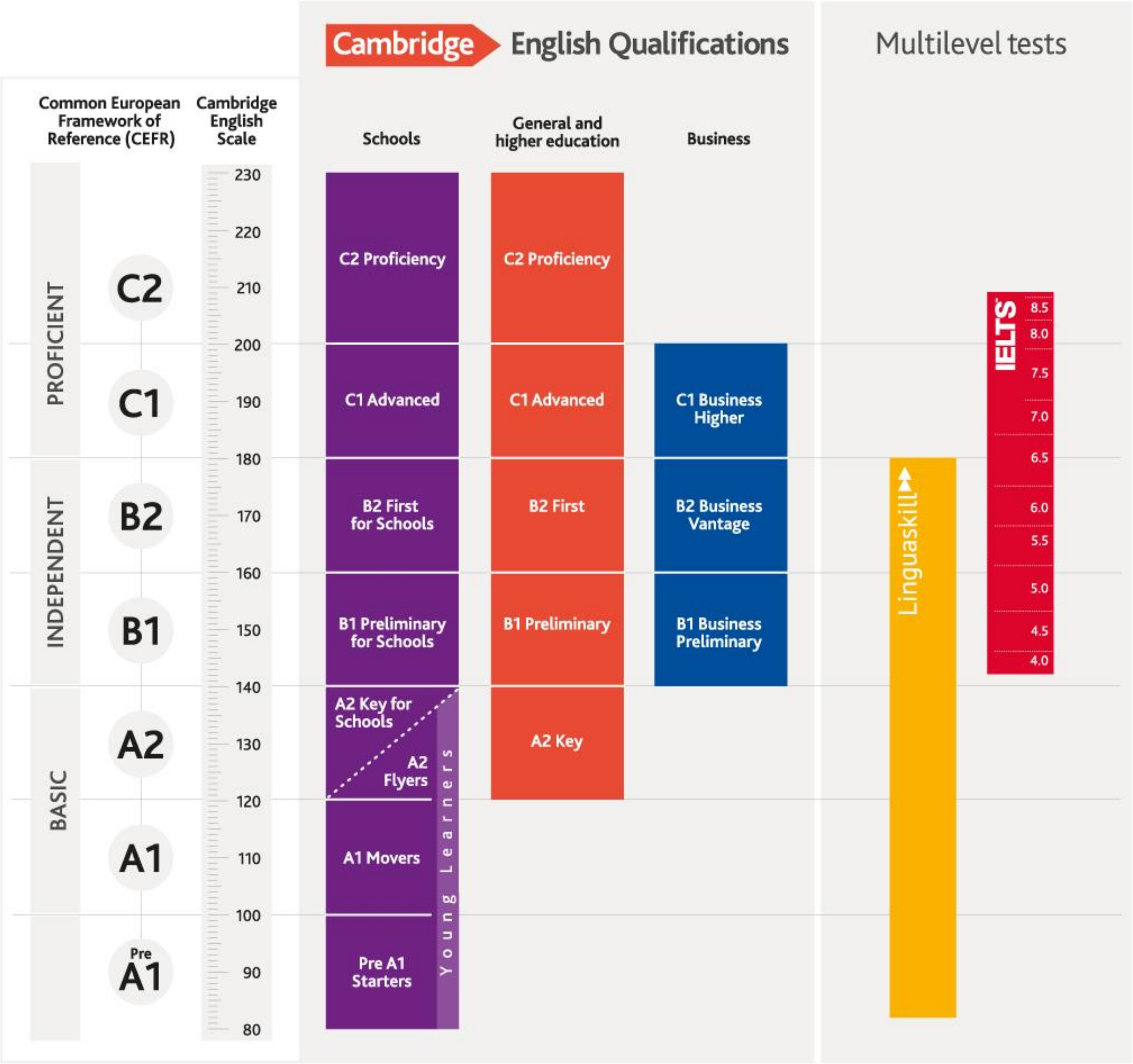
About the Common European Framework of Reference for Languages (CEFR)

The Common European Framework of Reference for Languages (CEFR) is an international standard for describing language ability. It describes language ability on a six-point scale, from A1 for beginners, up to C2 for those who have mastered a language. This makes it easy for anyone involved in language teaching and testing, such as teachers or learners, to see the level of different qualifications. It also means that employers and educational institutions can easily compare our qualifications to other exams in their country.

The CEFR Levels

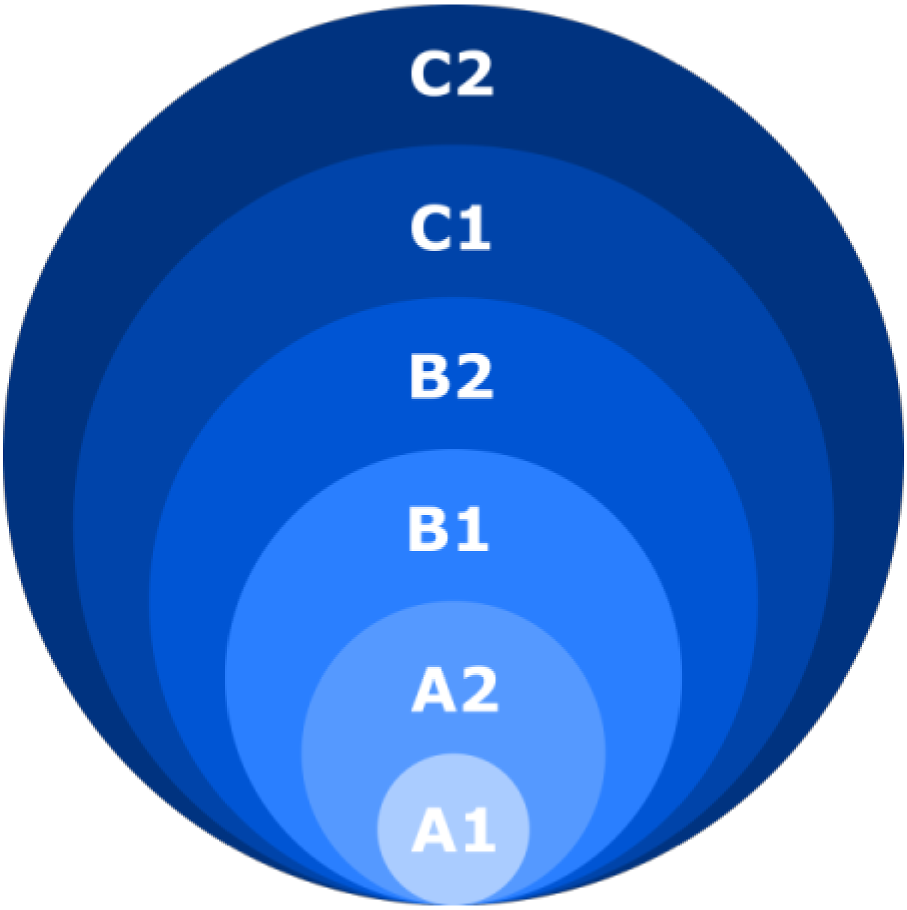
Some of the instruments produced within the Council of Europe have played a decisive role in the teaching of so-called “foreign” languages by promoting

methodological innovations and new approaches to designing teaching programmes, notably the development of a communicative approach.
The diagram below shows all of our English exams on the CEFR.



They have facilitated a fresh approach to communicating these teaching methods in a manner potentially more conducive to operational appropriation of unknown languages. By thus identifying language needs, they were able to pinpoint the knowledge and know-how required for attaining this communication “threshold”. The CEFR organises language proficiency in six levels, A1 to C2, which can be regrouped into three broad levels: Basic User, Independent User and

Proficient User, and that can be further subdivided according to the needs of the local context.



The levels are defined through 'can-do' descriptors. The levels did not suddenly appear from nowhere in 2001, but were a development over a period of time, as described below.

The CEFR: a turning point

The first specification of this "threshold level" was formulated for the English language (*Threshold level*, 1975), quickly followed by French (*Un Niveau Seuil*, 1976). These two instruments have been used de facto as models for the same type of reference instruments that were produced subsequently for other languages, but they were adapted to suit the peculiar features of each language. In order to meet the teaching and certification requirements, the level concept as defined was extended to cover specification of levels lying immediately below and above the threshold level. In the light of the developments in this field, particularly as regards the CEFR, other levels were developed for a number of languages. These proficiency levels constitute one of the origins of the six-level scale of the CEFR.

Launched in 2001, the CEFR marked a major turning point as it can be adapted and used for multiple contexts and applied for all languages.

The CEFR is based on all these achievements and has developed a description of the process of mastering an unknown language by type of competence and sub-competence, using descriptors for each competence or sub-competence, on which we shall not go into further detail here. These descriptors were created without reference to any specific language, which guarantees their relevance and across-the-board applicability. The descriptors specify progressive mastery of each skill, which is graded on a six-level scale (A1, A2, B1, B2, C1, C2).

However, for textbook authors, teachers and other professionals, the specification set out in the CEFR may appear excessively broad, particularly since individual languages are not addressed. The **Reference Level Descriptions** (RLD) for national and regional languages, which provide detailed content specifications for different CEFR levels, have been developed to address this issue.

CEFR: three tables used to introduce the Common Reference Levels

The following three tables, which are used to introduce the Common Reference Levels, are summarised from the original bank of "illustrative descriptors" developed and validated for the CEFR in the Swiss National Research project described in Appendix B of the volume. These formulations have been mathematically scaled to these levels by analysing the way in which they have been interpreted in the assessment of large numbers of learners.

Global scale

It is desirable that the common reference points are presented in different ways for different purposes. For some purposes it will however be appropriate to summarise the set of proposed Common Reference Levels in a holistic summarized table. Such a simple 'global' representation will make it easier to communicate the system to non-specialist users and will provide teachers and curriculum planners with orientation points.

Official translations of the CEFR Global scale

***Self-assessment grid*.**

In order to orient learners, teachers and other users within the educational system for some practical purpose, a more detailed overview is necessary. Table 2 is a draft for a self-assessment orientation tool intended to help learners to profile their main language skills, and decide at which level they might look at a checklist of more detailed descriptors in order to self-assess their level of proficiency.

Table 3 (CECR 3.3): Common Reference levels

Qualitative aspects of spoken language use

The chart in this table was designed to assess spoken performances. It focuses on different qualitative aspects of language use.

Tests

1. Fill the blank with appropriate definition. The … is an internationally recognised framework that describes 6 levels of language ability from A1 for beginners up to C2 for those who have mastered a language.

a. CEFR

b. TOEFL

c. IELTS

2. This test named ……… is the easiest of the Cambridge exams. Difficulty level: A2 elementary. Which test is being described?

a. PET (Preliminary English Test)

b. KET (Key English Test)

c. FCE (First Certificate in English)

3. This test named …. is the most important of the Cambridge exams. Difficulty level: B2 /Upper Intermediate. Which test is being described?

a. KET (Key English Test)

b. PET (Preliminary English Test)

*c. FCE (First Certificate in English) **

4. This test named ... is one of the Cambridge ESOL exams. Difficulty level:

B1 / low intermediate. Which test is being described?

a. KET (Key English Test)

b. PET (Preliminary English Test)

c. FCE (First Certificate in English)

5.What is being described? The study and practice of teaching methods appropriate to working with adults is called…

a. pedagogy   b. pedagogical competence   c. andragogy d. no answer

6. How is it called "In the opposite direction to the movement of the hands of a clock".

a. Anticlockwise   b. Clockwise   c. Backwards   No correct answer

THEME 7: TEST METHODS

Getting started. Learn more.

Plan:

1. The main importance of test methods in organizing tests.
2. The main features of test methods.
3. The types of test methods.

Types of Language Tests The needs of assessing the outcome of learning have led to the development and elaboration of different test formats. Testing language has traditionally taken the form of testing knowledge about language, usually the testing of knowledge of vocabulary and grammar. Stern (1983, p. 340) notes that „if the ultimate objective of language teaching is effective language learning, then our main concern must be the learning outcome‟. In the same line of thought, Wigglesworth (2008, p. 111) further adds that "In the assessment of languages, tasks are designed to measure learners‟ productive language skills through performances which allow candidates to demonstrate the kinds of language skills that may be required in a real world context." This is because a "specific purpose language test is one in which test content and methods are derived from an analysis of a specific purposes target language use situation, so that test tasks and content are authentically representative of tasks in the target situation" (Douglas, 2000, p. 19). Thus, the issue of authenticity is central to the assessment of language for specific functions. This is another way of saying that testing is a socially situated activity although the social aspects have been relatively under-explored (Wigglesworth, 2008). Yet, language tests differ with respect to how they are designed, and what they are for, in other words, in respect to test method and test purpose. In terms of method, we can broadly distinguish traditional paper-and-pencil language tests from performance tests. Paper-and-pencil language tests are typically used for the assessment either of separate components of language knowledge (grammar, vocabulary etc.), or of a receptive understanding (listening and reading comprehension). In performance-based tests, the language skills are assessed in an act of communication. Performance tests1 are most commonly tests of speaking and writing, for instance, to ask a language learner to introduce himself or herself formally or informally and to write a composition, a paragraph or an essay, 1 A performance test is "a test in which the ability of candidates to perform particular tasks, usually associated with job or study requirements, is assessed" (Davies et al., 1999, p. 144). 2 on the way he or she spent her summer holidays. These examples are elicited in the context of simulations of real-world tasks in realistic contexts. In terms of purpose, several types of language tests have devised to measure the learning outcomes accordingly. However, each test has its specific purpose, properties and criterion to be measured 2. The test types that will be dealt with in this part have been laid-out not in terms of importance, they are all of equal importance, but on the basis of alphabetical order. Yet, dictation, the traditional testing device which focuses much more on discrete language items, will have its fair of attention in terms of its pro‟s and con‟s. 1. Achievement Test an achievement test, also referred to as attainment or summative test, are devised to measure how much of a language someone has learned with reference to a particular course of study or programme of instruction, e.g. end-of-year tests designed to show mastery of a language. An achievement test might be a listening comprehension test based on a particular set of situational dialogues in a textbook. The test has a two-fold objective: 1) To help the teachers judge the success of their teaching. 2) To identify the weaknesses of their learners. In more practical and pedagogical terms, Brown (1994, p. 259) defines an achievement test as „tests that are limited to particular material covered in a curriculum within a particular time frame". In other words, they are designed primarily to measure individual progress rather than as a means of motivating or reinforcing language. Ideally, achievement tests are rarely constructed by classroom teacher for a particular class. 2 Richards et al. (1985) define a criterion-referenced test (CRT) as: a test which measures a student's performance according to a particular standard or criterion which has been agreed upon. The student must reach this level of performance to pass the test, and a student's score is therefore interpreted with reference to the criterion score, rather to the scores of the students. That definition is very different from their definition for a norm-referenced test (NRT) which they say is: a test which is designed to measure how the performance of a particular student or group of students compares with the performance of another student or group of students whose scores are given as the norm. a student's score is therefore interpreted with reference to the scores of other students or group of students, rather than to an agreed criterion score. 3 2. Cloze Test A cloze test, also alternately referred to as cloze procedure, consists of a set of techniques for measuring, for example, reading comprehension. In a cloze test words are removed from a reading passage at regular intervals, leaving blanks. For example every fifth word may be removed. The reader must then read the passage and try to guess the missing words.

For example, a cloze passage looks like this: A passage used in ……….. cloze test is a ………….. of written material in ……….. words have been regularly………... The learners must then ……….. to reconstruct the passage ……….. filling the missing …………. (Adapted from Richards et al., 1989, p. 4. Here, the test-taker or the reader has to guess the following missing words: a, passage, which, removed, try, by and words. The cloze test can also be used to judge the difficulty of reading materials. If the cloze procedure is being used for language testing, the test-taker is given a score according to how well the words guessed match the original words, or whether or not they make sense.

Two types of scoring procedure are used:

1) The reader must guess the exact word which was used in the original (as in the example) above. This is called exact word method.

2) The reader can guess any word that is appropriate or acceptable in the context.

This is called the acceptable word method. Another illustrative example of close test looks something like the following: 'A week has seven ....'. The only word which will fit in this blank is „days‟. But sometimes one can choose between two or more words, as in: 'We write with a.....'. In this blank one can write „pen‟ or „pencil‟ or even „chalk‟, „computer‟ or „typewriter‟. However, two substantial criticisms have been made to the cloze-test types (Broughton et al., 1980). The first of these criticisms is that such tests rarely afford the person being tested any opportunity to produce language spontaneously. The second is that they are fundamentally trying to test that knowledge of the language system that underlies any actual instance of its use –linguistic competence in the Chomsky an sense- they are not concerned with the ability to 4 master the language system for particular purposes with particular people in particular situations. 3.Diagnostic Test As its name denotes, a diagnostic test is primarily designed to diagnose some particular linguistic aspects. Diagnostic tests in pronunciation, for example, might have the purpose of determining which particular phonological features of the English language are more likely to pose problems and difficulties for a group of learners. One of the well-known diagnostic tests in English is Prator's (1972) Diagnostic Passage. It consists of a short written passage that the learner reads orally; the teacher then examines a tape recording of that reading against a very detailed checklist of pronunciation errors. Basically, diagnostic language tests have a threefold objective:

1.To provide learners with a way to start learning with their own personal learning programme or what would be called in the literature of testing learning paths.

2. To provide learners with a way to test their knowledge of a language.

3. To provide learners with better information about their strengths and weaknesses. Ideally, diagnostic tests are designed to assess students‟ linguistic knowledge (knowledge of and about the language) and language skills (listening, speaking, reading and writing) before a course is begun. However, the term formative is sometimes used to designate a diagnostic test. One of the main advantages of a diagnostic test is that it offers useful pedagogical solutions for mixed-ability classes. In this very specific context, Broughton et al. (1980) contend that: There will certainly be a large block in the middle of the ability range who can be separated off as a group for some parts of the lesson, or for some lessons, and will form a more homogenous teaching group. If this strategy is adopted, the poor ones and the better ones must receive their due time and attention. (Broughton et al. 1980, p. 189)

4.Discrete-Point Test The discrete-point test, also called discrete-item test, is a language test which measures knowledge of individual language items, such as a grammar test which has different sections on tenses, adverbs and prepositions. Discrete-point tests are based on the theory that language consists of different parts such as speech sounds, grammar and vocabulary, and different skills such as listening, speaking, reading and writing, and these are made up of elements that can be 5 tested separately. Test consisting of multiple-choice questions are usually regarded as discrete point tests. Discrete-point tests are all too often contrasted with what are called integrative tests. An integrative test is one which requires a learner to use several skills at the same time. An essay-writing is an integrative test because it leans heavily on the knowledge of grammar, vocabulary, and rules of discourse; a dictation is also an integrative test as it requires knowledge of grammar, vocabulary and listening comprehension skills. In this vein, Harmer notes the following distinction between discrete-point testing and integrative testing, “Whereas discrete point-testing only tests on thing at a time such as asking students to choose the correct tense of a verb, integrative test items expect students to use a variety of language at any one given time – as they will have to do when writing a composition or doing a conversational oral test" (Harmer, 2001, p. 323). In the same line of thought and Broughton et al. ,more than some thirty years ago, noted that "Since language is seen as a number of systems, there will be items to test knowledge of both the production and reception of the sound segment system, of the stress system, the intonation system, and morphemic system, the grammatical system, the lexical system and so on" (Broughton et al., 1980, pp. 149-150).

5.Language Aptitude Test Before one ventures into defining what a language aptitude test is, it would be wiser to start first by defining what a language aptitude is. Language aptitude, as a hybrid concept part linguistic and part psychological, refers to the genuine ability one is endowed with to learn a language. It is thought to be a combination of several abilities: • Phonological ability, i.e. the ability to detect phonetic differences (e.g. of stress, intonation, vowel quality) in a new language. • Syntactic ability, i.e., the ability to recognize the different grammatical functions of words in sentences. • Psychological ability, i.e. rote-learning abilities and the ability to make inferences and inductive learning. Additionally, Crystal (1989, p. 371) suggests other variables conducive to successful language learning such as „empathy and adaptability, assertiveness and independence with good drive and powers of application‟. A high language-aptitude person can learn more quickly and 6 easily than a low language-aptitude individual. The evidence in such assertion is axiomatic in a language aptitude test. A language aptitude test tends to measure a learner aptitude for language learning, be it second or foreign, i.e. students performance in a language. Thus, it is used to identify those learners who are most likely to succeed. Language aptitude tests usually consist of several different test items which measures such abilities as: • Sound-coding ability, i.e. the ability to identify and remember new sounds in a new language. • Grammar-coding ability, i.e. the ability to identify the grammatical functions of different parts of sentences. • Inductive-learning ability, i.e. the ability to work out meanings without explanation in the new language. • Memorization, i.e. the ability to remember and to recall words, patterns, rules in the new language. Two well-known standardized language aptitude tests have been used in the United States, the Modern Language Aptitude Test (Carroll and Sapon, 1958) and the Primsleur Language Aptitude Battery (Primsleur, 1966). Both of these are English tests and require students to perform such tasks as learning numbers, listening, detecting spelling clues and grammatical patterns and memorizing (Brown, 1994).

6.Placement Test A placement test, as its name implies, is originally designed to place learners at an appropriate level in a programme or course. The term "placement test" as Richards et al. (1989) note does not refer to what a test contains or how it is constructed, but to the purpose for which it used. Various types or testing procedures such as dictation, interview or a grammar test (discrete or integrative) can be used for placement purposes. The English Placement test (EPT), which is a well-known test in America, is an illustrative example of this test-type. The EPT is designed to assess the level of reading and writing skills of entering undergraduate students so that they can be placed in appropriate courses. Those undergraduate students who do not demonstrate college or university-level skills will be directed to remedial courses or programmes to help them attain these skills.

7.Proficiency Test A proficiency test is devised to measure how much of a language someone has learned. It is not linked to any particular course of instruction, but measures the learner‟s general level of language mastery. Most English language proficiency tests base their testing items on high frequency-count vocabulary and general basic grammar. Some proficiency tests have been standardized for worldwide use, such as the well-known American tests, the TOEFL, and the English Language Proficiency Test (ELPT)3 which are used to measure the English language proficiency of foreign students intending further study at English-speaking institutions, namely the USA. However, the Cambridge Certificate of Proficiency in English or CPE, as it is generally referred to, is the most advanced remains the only British top-value and highprestige standardized4 language test. It is the most advanced general English exam provided by the University of Cambridge. The Certificate is recognized by universities and employees throughout the world. The English level of those who pass the CPE is supposed to similar to that of a fairly educated native speaker of English. Clearly, as Valette posits, „the aim of a proficiency test is to determine whether this language ability corresponds to specific language requirements" (Valette, 1977, p. 6) Actually, there are four other types of Cambridge proficiency tests, the Cambridge Key English Test (KET), the Cambridge Preliminary English Test (PET), The Cambridge First Certificate of English (FCE) and the Cambridge Certificate in Advanced English (CAE). The material contained in proficiency tests can be used for teaching as well as for testing. In essence, a proficiency test measures what the student has learned in relation to a specific purpose, e.g. does the student know enough English to follow a course offered in English? 3 The English Language Proficiency Test (ELPT) was the name of a test last administered in January 2005. It was a one-hour multiple choice question given on English language proficiency. A student whose native language was not English could have chosen to take this test instead of or in addition to the TOEFL for college or university entrance depending upon the requirements of the schools in which the student was planning to apply. Until 1994, the tests were known as Achievement Tests. The ELPT assessed both the understanding of spoken and written standard American English and the ability to function in a classroom where English is spoken. 4 A standardized test is an exam which has been developed from tryouts and experimentation to ensure that it is reliable and valid. It is also a test for which norms have been established and it provides uniform procedures for administering (time limits, response format, and number of questions) and for scoring the test. “Standardized tests are often used by school systems for high-stakes decision making” (Menken, 2008, p. 402).

8.Progress Test A progress test is an achievement-like test. It is closely related to a particular set of teaching materials or a particular course of instruction. Progress tests are usually administered at the end of a unit, a course, or term. A progress test may be viewed as similar to an achievement test but much narrower and much more specific in scope (Richards et al., 1989). They help examiners in general and language teachers in particular to assess the degree of success of their programmes and teaching and therefore to identify their shortcomings and weaknesses respectively. Progress tests can also be diagnostic to some degree, in the sense that they help identify areas of difficulties encountered by learners in general. 9. TOEFL The Test of English as a Foreign Language, or TOEFL for short, is a large-scale language assessment. It is, "arguably the most well-known and widely used large-scale language assessment in the world" . It was first developed in 1963 in the United States to help in the assessment of the language competence of non-native speakers. As a test type, it is a standardized test of English proficiency administered by the Educational Testing Service, Princeton. It is widely used to measure the English-language proficiency of foreign students wishing to enter American colleges and universities. According to Taylor and Angelis (cited in Kunnan, 2008) the first TOEFL was administered in 1964 at 57 test centres to 920 test candidates. Recently, the TOEFL has widely been recognized as a model test and have-take-test for our students, graduate and postgraduate, as well as our teachers and researchers in universities and higher education institutions wishing to read for higher degrees and develop further their research potential in North American universities5. Kunnan (2008, p. 141) notes that, "Over the years, the TOEFL became mandatory for non-American and non-Canadian native speakers of English applicants to undergraduate and graduate programs in U.S. and Canadian English-medium universities". One of the most important realizations in the TOEFL enterprise was the launching of a more innovative test, the iBTOEFL, internet-based TOEFL, in 2005. This iB TOEFL is 5 The International English Language Testing System, IELTS, is designed to assess the language ability of candidates who wish to study or work in countries where English is the language of communication. IELTS is required for admission to British universities and colleges. It is also recognized by universities and employers in Australia, Canada, and the USA. IELTS is jointly managed by the

University of Cambridge, British Council and IDP Education.9 regarded as a significant development over the previous TOEFL forms and the TOEFL CBT, Computer-Based Test, launched in 1996. The novel features of the TOEFL are a speaking section consisting of independent and integrated skills tasks, a listening section with longer lectures and conversations with note-taking, a reading section made up of questions that ask test-takers to categorize information and fill in a chart or complete a summary and a writing section that has both an independent and integrated task.

THEME 8. HOW TO TEST AND ASSESS INTEGRATED SKILLS

Getting started. Learn more.

Plan:

1. Why should we integrate the four skills?
2. How can we integrate the four skills?
3. How can we test speaking skills?
4. How language skills are taught in an integrated way?

INTEGRATING THE LANGUAGE SKILLS

One image for teaching English as a second or foreign language (ESL/EFL) is that of a tapestry. The tapestry is woven from many strands, such as the characteristics of the teacher, the learner, the setting, and the relevant languages (i.e., English and the native languages of the learners and the teacher). For the instructional loom to produce a large, strong, beautiful, colorful tapestry, all of these strands must be interwoven in positive ways. For example, the instructor's teaching style must address the learning style of the learner, the learner must be motivated, and the setting must provide resources and values that strongly support the teaching of the language. However, if the strands are not woven together effectively, the instructional loom is likely to produce something small, weak, ragged, and pale--not recognizable as a tapestry at all.

In addition to the four strands mentioned above-teacher, learner, setting, and relevant languages-other important strands exist in the tapestry. In a practical sense, one of the most crucial of these strands consists of the four primary skills of listening, reading, speaking, and writing. This strand also includes associated or related skills such as knowledge of vocabulary, spelling, pronunciation, syntax, meaning, and usage. The skill strand of the tapestry leads to optimal ESL/EFL communication when the skills are interwoven during instruction. This is known as the integrated-skill approach. If this weaving together does not occur, the strand consists merely of discrete, segregated skills-parallel threads that do not touch, support, or interact with each other. This is sometimes known as the segregated-skill approach. Another title for this mode of instruction is the language-based approach, because the language itself is the focus of instruction (language for language's sake). In this approach, the emphasis is not on learning for authentic communication. By examining segregated-skill instruction, we can see the advantages of integrating the skills and move toward improving teaching for English language learners.

SEGREGATED-SKILL INSTRUCTION

In the segregated-skill approach, the mastery of discrete language skills such as reading and speaking is seen as the key to successful learning, and language learning is typically separate from content learning (Mohan, 1986). This is contrary to the integrated way that people use language skills in normal communication, and it clashes with the direction in which language teaching experts have been moving in recent years.

Skill segregation is reflected in traditional ESL/EFL programs that offer classes focusing on segregated language skills. Why do they offer such classes? Perhaps teachers and administrators think it is logistically easier to present courses on writing divorced from speaking, or on listening isolated from reading. They may believe that it is instructionally impossible to concentrate on more than one skill at a time.

Even if it were possible to fully develop one or two skills in the absence of all the others, such an approach would not ensure adequate preparation for later success in academic communication, career-related language use, or everyday interaction in the language. An extreme example is the grammar-translation method, which teaches students to analyze grammar and to translate (usually in writing) from one language to another. This method restricts language learning to a very narrow, noncommunicative range that does not prepare students to use the language in everyday life.

Frequently, segregated-skill ESL/EFL classes present instruction in terms of skill-linked learning strategies: reading strategies, listening strategies, speaking strategies, and writing strategies (see Peregoy & Boyle, 2001). Learning strategies are strategies that students employ, most often consciously, to improve their learning. Examples are guessing meaning based on context, breaking a sentence or word down into parts to understand the meaning, and practicing the language with someone else.

Very frequently, experts demonstrate strategies as though they were linked to only one particular skill, such as reading or writing (e.g., Peregoy & Boyle, 2001). However, it can be confusing or misleading to believe that a given strategy is associated with only one specific language skill. Many strategies, such as paying selective attention, self-evaluating, asking questions, analyzing, synthesizing, planning, and predicting, are applicable across skill areas (see Oxford, 1990). Common strategies help weave the skills together. Teaching students to improve their learning strategies in one skill area can often enhance performance in all language skills (Oxford, 1996).

Fortunately, in many instances where an ESL or EFL course is labeled by a single skill, the segregation of language skills might be only partial or even illusory. If the teacher is creative, a course bearing a discrete-skill title might actually involve multiple, integrated skills. For example, in a course on intermediate reading, the teacher probably gives all of the directions orally in English, thus causing students to use their listening ability to understand the assignment. In this course, students might discuss their readings, thus employing speaking and listening skills and certain associated skills, such as pronunciation, syntax, and social usage. Students might be asked to summarize or analyze readings in written form, thus activating their writing skills. In a real sense, then, some courses that are labeled according to one specific skill might actually reflect an integrated-skill approach after all. The same can be said for ESL/EFL textbooks. A particular series might highlight certain skills in one book or another, but all the language skills might nevertheless be present in the tasks in each book. In this way, students have the benefit of practicing all the language skills in an integrated, natural, communicative way, even if one skill is the main focus of a given volume.

In contrast to segregated-skill instruction, both actual and apparent, there are at least two forms of instruction that are clearly oriented toward integrating the skills.

TWO FORMS OF INTEGRATED-SKILL INSTRUCTION

Two types of integrated-skill instruction are content-based language instruction and task-based instruction. The first of these emphasizes learning content through language, while the second stresses doing tasks that require communicative language use. Both of these benefit from a diverse range of materials, textbooks, and technologies for the ESL or EFL classroom.

"Content-Based Instruction". In content-based instruction, students practice all the language skills in a highly integrated, communicative fashion while learning content such as science, mathematics, and social studies. Content-based language instruction is valuable at all levels of proficiency, but the nature of the content might differ by proficiency level. For beginners, the content often involves basic social and interpersonal communication skills, but past the beginning level, the content can become increasingly academic and complex. The Cognitive Academic Language Learning Approach (CALLA), created by Chamot and O'Malley (1994) shows how language learning strategies can be integrated into the simultaneous learning of content and language.

At least three general models of content-based language instruction exist: theme-based, adjunct, and sheltered (Scarcella & Oxford, 1992). The theme-based model integrates the language skills into the study of a theme (e.g., urban violence, cross cultural differences in marriage practices, natural wonders of the world, or a broad topic such as change). The theme must be very interesting to students and must allow a wide variety of language skills to be practiced, always in the service of communicating about the theme. This is the most useful and widespread form of content-based instruction today, and it is found in many innovative ESL and EFL textbooks. In the adjunct model, language and content courses are taught separately but are carefully coordinated. In the sheltered model, the subject matter is taught in simplified English tailored to students' English proficiency level.

"Task-Based Instruction". In task-based instruction, students participate in communicative tasks in English. Tasks are defined as activities that can stand alone as fundamental units and that require comprehending, producing, manipulating, or interacting in authentic language while attention is principally paid to meaning rather than form (Nunan, 1989).

The task-based model is beginning to influence the measurement of learning strategies, not just the teaching of ESL and EFL. In task-based instruction, basic pair work and group work are often used to increase student interaction and collaboration. For instance, students work together to write and edit a class newspaper, develop a television commercial, enact scenes from a play, or take part in other joint tasks. More structured cooperative learning formats can also be used in task-based instruction. Task-based instruction is relevant to all levels of language proficiency, but the nature of the task varies from one level to the other. Tasks become increasingly complex at higher proficiency levels. For instance, beginners might be asked to introduce each other and share one item of information about each other. More advanced students might do more intricate and demanding tasks, such as taking a public opinion poll at school, the university, or a shopping mall.

ADVANTAGES OF THE INTEGRATED-SKILL APPROACH

The integrated-skill approach, as contrasted with the purely segregated approach, exposes English language learners to authentic language and challenges them to interact naturally in the language. Learners rapidly gain a true picture of the richness and complexity of the English language as employed for communication. Moreover, this approach stresses that English is not just an object of academic interest nor merely a key to passing an examination; instead, English becomes a real means of interaction and sharing among people. This approach allows teachers to track students' progress in multiple skills at the same time. Integrating the language skills also promotes the learning of real content, not just the dissection of language forms. Finally, the integrated-skill approach, whether found in content-based or task-based language instruction or some hybrid form, can be highly motivating to students of all ages and backgrounds.

INTEGRATING THE LANGUAGE SKILLS

In order to integrate the language skills in ESL/EFL instruction, teachers should consider taking these steps:

* *Learn more about the various ways to integrate language skills in the classroom (e.g., content-based, task-based, or a combination).*

* *Reflect on their current approach and evaluate the extent to which the skills are integrated.*

* *Choose instructional materials, textbooks, and technologies that promote the integration of listening, reading, speaking, and writing, as well as the associated skills of syntax, vocabulary, and so on.*

* *Even if a given course is labeled according to just one skill, remember that it is possible to integrate the other language skills through appropriate tasks.*

* *Teach language learning strategies and emphasize that a given strategy can often enhance performance in multiple skills.*

With careful reflection and planning, any teacher can integrate the language skills and strengthen the tapestry of language teaching and learning. When the tapestry is woven well, learners can use English effectively for communication.

What is Integrated Skills Assessment

Assessment that incorporates several skills within one test to determine whether a student can tackle the complexity of real-world tasks in academia that require multiple skill sets. Within an EAP program, integrated skills tests require students to produce written or oral work that incorporates meaningful uses of source evidence, both conceptually - to comprehend, synthesize, and present ideas from sources – and through writing – to conform with stylistic convention for presenting ideas from sources, and acknowledging those sources.

In any educational endeavor, the tripartite relationship of teaching, learning, and assessment are of critical importance. This chapter concentrates on the assessment of the English for academic purposes literacies (EAPL), which is defined as the assessment of second language and literacies of multilingual international students (henceforth considered as L2 learners) at an entry level English for Academic Purposes program in higher educational context. The chapter starts by distinguishing second language tests and academic literacies assessment. It then presents a historical overview of language assessment and assessment of academic literacies, specifically in the context of English for Academic Purposes (EAP). Of particular importance are the definition and construct of academic literacies assessment. Based on this review, the chapter then presents practical applications of the discussed EAPL construct.

EAPL refers to teaching, learning, and assessment of English that covers a wide range of general as well as discipline-specific topics which aim to help students, "to develop their academic literacy skills to facilitate their effective participation in academic communities" (Hamp-Lyons, 2011, p. 100). Two key considerations need to be taken into account: 1. students need an understanding of how knowledge is created, presented, and debated in any discipline, in addition to linguistic support; and 2. academic literacies (reading, writing, and reasoning in any discipline) are difficult for native and non-native speakers alike (Wingate & Tribble, 2012).

Depending on their uses, roles, purposes, and contexts, English language tests and assessments can generally be divided into two types: internally mandated and externally mandated (Davidson & Lynch, 2002). Externally mandated summative assessment occurs at the end of a particular unit of instruction (e.g., chapter, unit or semester), and its purpose is to primarily categorize students' performances. Internally mandated assessment, on the other hand, can be administered at the beginning or during the course of a unit of instruction.

Internally mandated tests, also referred to as formative assessment (i.e., assessments for learning or classroom-based assessments), are related to "the needs of the teachers and learners working within a particular context and … are generally ecologically sensitive" (Fulcher, 2010, pp. 1-2). Such forms of assessment are an essential component of classroom work and are used to inform teaching and learning and ultimately to raise the standards of achievement (Black & Wiliam, 1998). In addition to diagnosing difficulties in individual learners, internally mandated tests can also be used for placement or achievement purposes.

On the other hand, *summative assessment* (or assessment *of* learning), is externally mandated by a group of people who often "do not know a great deal about the local learning ecology [context], and probably don't even know the teachers and learners who will have to cope with the required testing regime" (Fulcher, 2010, p. 2). Such tests (e.g., General Certificate of Secondary Education (GCSE) examinations in England and the College English Test (CET) in China) are high-stakes in nature and are used by policymakers or other stakeholders to make judgments about proficiency and achievement of learners at the end of a study period where learners are expected to have reached a particular standard.

Other than stakes, test purposes are also one of the important considerations in using any language test. *Proficiency* and *achievement* are two important test purposes of an English for Academic Purposes assessment. A proficiency test measures general ability in language and not specific content, course, curriculum area, or skills in the language. It is based on what learners can do with the language (Fulcher, 2010). An achievement test, on the other hand, is directly related to a language course and aims to measure what has been taught. These tests are based on a detailed course syllabus and objectives and are usually administered at the end of a course or a unit of study.

Key Terms. <u>Formative Assessment</u>: Assessment that takes place during an instructional cycle. It can be used for enhancing and informing teaching and learning in classroom.

High-Stakes Tests: Tests that are viewed as powerful measures to change the course of student's progress in an EAP program. One example of high-stakes tests is proficiency tests for universities in Anglophone universities where failure can mean expulsion from the program.

Multimodality: Combining different individual modes such as discipline-specific texts, audios, images, and videos to create meaningful communication that encourages interaction and learning in an EAP context.

Performance-Based Assessment: A more valid construct of EAPL, performance assessment tasks are authentic tasks that use real-world contexts. They require learners to work independently and use 21 st century skills such as higher-level thinking and problem solving. These tasks also help teachers in providing constructive feedback to students about their strength and weaknesses. Examples of performance based EAPL tasks include searching and selecting relevant sources, taking notes, writing essays, and making presentations.

Washback: All the intended (or positive) and unintended (or negative) effects of assessment on teaching and learning in the classroom.

Needs Assessment: Assessment that helps teachers elicit information about students' needs and design effective course materials.

Summative Assessment: Assessment that takes place at the end of a teaching cycle to measure students' learning e.g., achievement testing.

Integrated Skills Assessment: Assessment that incorporates several skills within one test to determine whether a student can tackle the complexity of real-world tasks in academia that require multiple skill sets. Within an EAP program, integrated skills tests require students to produce written or oral work that incorporates meaningful uses of source evidence, both conceptually - to comprehend, synthesize, and present ideas from sources – and through writing – to conform with stylistic convention for presenting ideas from sources, and acknowledging those sources.

ISE levels and the CEFR

ISE Foundation to ISE align with the levels of the Common European Framework of Reference (CEFR) for Languages (Council of Europe,2001) as follows:

ISE level	CEFR level
ISE Foundation	A2
ISE I	B1
ISE II	B2
ISE III	C1

Integrated skills assessment – structure of the qualification.

ISE is taken in two modules – Reading and Writing and Listening. Once the two modules have been passed at the same level a certificate for the full qualification is awarded. The four skills are assessed both independently and in an integrated way:

Module	Component	Method
Reading and Writing	Long reading Multi-text reading Reading into writing Extended writing	Reading a single text and short questions Reading three or four shorter texts and short questions Reading texts and producing a short piece of writing using the texts as source material
Speaking and Listening	Independent listening Independent listening into speaking Integrated speaking and	Listening to a recording and reporting information either on paper or verbally

	listening	A short piece of writing similar to the kind of writing done in school or college Listening to a recording and verbally reporting and discussing the content A phased speaking exam including discussion of topic, a conversation and a collaborative task (depending on the level)

Reading assessment

Reading is dichotomously scored. The reading exam consists of 30 item over two tasks. The table below shows how reading is assessed, for assessing reading skill we'll implement reading strategies for assessing reading skill on the work of "I, Robot" by Isaac Asimov. In this case it is used the reading sub skills: inferring the work the multiple choice, matching, True/false test methods

We know that while assessing reading the effective readers use strategies to understand what they read before, during, and after reading. To improve student's reading comprehension, teachers should introduce the seven cognitive strategies of effective readers: Activating, Inferring, Monitoring-clarifying, Questioning, Searching-selecting, Summarizing and Visualizing-organizing.

Ex: for assessing reading skill of students on the work of "I, Robot" by Isaac Asimov we use reading sub-skills – skimming, scanning and in-depth reading. Using intensive reading activities, it's included skimming a reading material ("I, Robot" by I. Asimov) for specific information to answer true or false statements of filling gaps in a summary, scanning a text to match headings to paragraphs, scanning jumbled paragraphs and reading them carefully to put them into the correct order.

THEME 9: HOW TO ASSESS WRITING SKILL

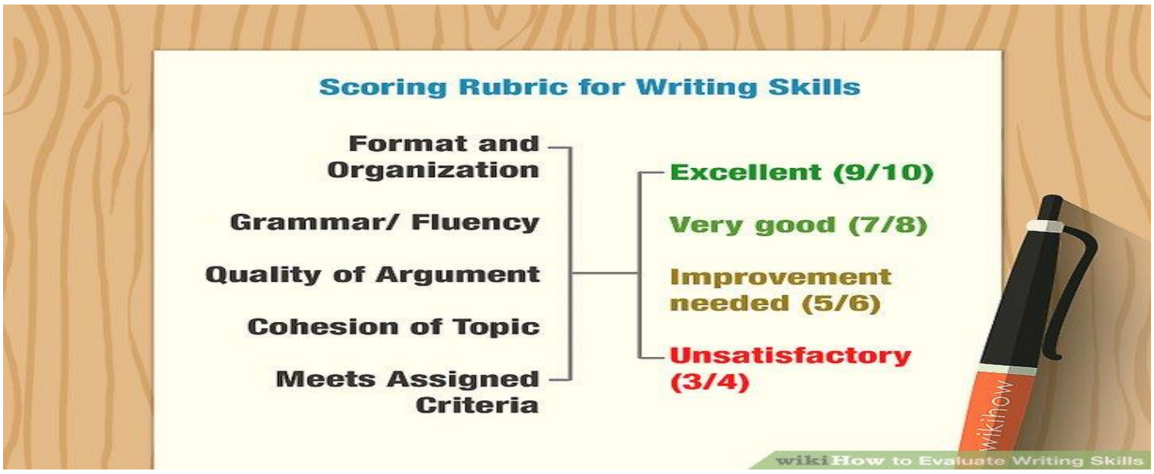
Getting started. Learn more.

Plan:

1. **What is a writing assessment test?**
2. **How do you demonstrate good writing skills?**
3. **What is testing writing?**

Determination your assessment criteria.

1. The use of proper **writing** conventions, such as good spelling, grammar, syntax, capitalization, and punctuation.
2. The writer's mastery of **written** vocabulary.
3. The clarity and fluency with which the writer presents their arguments.
4. The use of clear and logical structure within the text.



Simple Ways to Assess the Writing Skills of Students with Learning Disabilities

Student writing can be evaluated on five product factors: fluency, content, conventions, syntax, and vocabulary. Writing samples also should be assessed across a variety of purposes for writing to give a complete picture of a student's writing performance across different text structures and genres.

These simple classroom help in identifying strengths and weaknesses, planning instruction, evaluating instructional activities, giving feedback, monitoring performance, and reporting progress. (*Stephen L. Isaacson Portland State University This article is adapted for LD On Line from a similar article by Isaacson published in The Volta Review, 1996, Vol. 98, No. 1, pp. 183-199)*

Simple ways to assess the writing skills of students with learning disabilities

A teacher's first responsibility is to provide opportunities for writing and encouragement for students who attempt to write. A teacher's second responsibility is to promote students' success in writing. The teacher does this by carefully monitoring students' writing to assess strengths and weaknesses, teaching specific skills and strategies in response to student needs, and giving careful feedback that will reinforce newly learned skills and correct recurring problems. These responsibilities reveal, upon inspection, that assessment is clearly an integral part of good instruction. In their review of the existing research on effective instruction Christenson, Ysseldyke, and Thurlow (1989) found that, in addition to other factors, the following conditions were positively correlated to pupil achievement:

a. The degree to which there is an appropriate instructional match between student characteristics and task characteristics (in other words, teachers must assess the student's prior knowledge and current level of skills in order to match them to a task that is relevant and appropriate to their aptitudes);

b. The degree to which the teacher actively monitors students' understanding and progress; and

c. The degree to which student performance is evaluated frequently and appropriately (congruent with what is taught).

Assessment, therefore, is an essential component of effective instruction. Airasian (1996) identified three types of classroom assessments. The first he called "sizing-up" assessments, usually done during the first week of school to provide the teacher with quick information about the students when beginning their instruction. The second type, instructional assessments, are used for the daily tasks of planning instruction, giving feedback, and monitoring student progress. The third type he referred to as official assessments, which are the periodic formal functions of assessment for grouping, grading, and reporting. In other words, teachers use assessment for identifying strengths and weaknesses, planning instruction to fit diagnosed needs, evaluating instructional activities, giving feedback, monitoring performance, and reporting progress. Simple curriculum-based methods for assessing written expression can meet all these purposes.

Process, product, and purpose

Curriculum-based assessment must start with an inspection of the curriculum. Many writing curricula are based on a conceptual model that takes into account process, product, and purpose. This conceptual model, therefore, forms the framework for the simple assessment techniques that follow.

Simple ways to assess the process

The diagnostic uses of assessment (determining the reasons for writing problems and the student's instructional needs) are best met by looking at the process of writing, i.e., the steps students go through and strategies they use as they work at writing. How much planning does the student do before he or she writes? Does she have a strategy for organizing ideas? What seem to be the obstacles to getting thoughts down on paper? How does the student attempt to spell words she does not know? Does the student reread what she has written? Does the student talk about or share her work with others as she is writing it? What kind of changes does the student make to her first draft?

In order to make instructionally relevant observations, the observer must work from a conceptual model of what the writing process should be. Educators have reached little consensus regarding the number of steps in the writing process. Writing experts have proposed as few as two (Elbow, 1981) and as many as nine (Frank, 1979). Englert, Raphael, Anderson, Anthony, and Stevens (1991) provided a model of a five-step writing process using the acronym POWER: Plan, Organize, Write, Edit, and Revise. Each step has its own sub-steps and strategies that become more sophisticated as the students become more mature as writers, accommodating their style to specific text structures and purposes of writing. Assessment of the writing process can be done through observation of students as they go through the steps of writing.

Having students assess their own writing process is also important for two reasons. First, self-assessment allows students an opportunity to observe and reflect on their own approach, drawing attention to important steps that may be overlooked. Second, self-assessment following a conceptual model like POWER is a means of internalizing an explicit strategy, allowing opportunities for the student to mentally rehearse the strategy steps. Figure 1 is a format for both self-observation and teacher observation of the writing process following the POWER strategy. Similar self-assessments or observation checklists could be constructed for other conceptual models of the writing process.

Figure 1. Using a five-step conceptual model for student and teacher observation of the writing process

POWER Looking at How I Write			
My Comments			Teacher Comments
Plan			
I chose a good topic	Yes	No	
I read about my topic	Yes	No	
I thought about what the readers will want to know	Yes	No	
I wrote down all my ideas on a "think sheet"	Yes	No	
Organize			
I put similar ideas together	Yes	No	
I chose the best ideas for my composition	Yes	No	

I numbered my ideas in logical order	Yes	No	
Write			
I wrote down my ideas in sentences	Yes	No	
When I needed help I… ____did the best I could ____looked in a book ____asked my partner ____asked the teacher			
Edit			
I read my first draft to myself	Yes	No	
I marked the parts I like	Yes	No	
I marked the parts I might want to change	Yes	No	
I read my first draft to my partner	Yes	No	
I listened to my partner's suggestions	Yes	No	
Rewrite			
I made changes to my composition	Yes	No	
I edited for correctness	Yes	No	
I wrote the final draft in my best writing	Yes	No	

Simple ways to assess the product

An effective writing process should lead to a successful product. A writing product fulfills its communicative intent if it is of appropriate length, is logical and coherent, and has a readable format. It is a pleasure to read if it is composed of well-constructed sentences and a rich variety of words that clearly convey the author's meaning. When various conceptual models of writing are compared side by side (Isaacson, 1984) five product variables seem to emerge: fluency, content, conventions, syntax, and vocabulary. Too often teachers focus their attention primarily on surface features of a student's composition related to the mechanical aspects of writing, or conventions. A balanced assessment should look at all five aspects of a student's writing. The following are simple methods for assessing each product variable. In some instances, quantifiable measures are used; in others, qualitative assessments seem more appropriate.

Fluency

The first writing skill a teacher might assess with a beginning writer is fluency: being able to translate one's thoughts into written words. As concepts of print and fine motor skills develop, the student should become more proficient at writing down words and sentences into compositions of gradually increasing length. The developmental route of very young writers involves trying to understand what written language is about as they look at books, become aware of environmental print, and put pencil to paper (Clay, 1982). Then children try to relate their experiences in writing using invented spelling. As they begin to construct little stories they explore spelling patterns and develop new language patterns. Clay (1979, 1993) recommends a simple rating scale for emerging writing skills that focuses on language level (from only letters to sentences and paragraphs), message quality, and directional principles

Figure 2. Rating a child's early attempts at writing (Clay, 1993)

Language Level

Record the highest level of linguistic organization used by the child:

1. Alphabetical
2. Word (any recognizable word)
3. Word group (any two-word phrase)
4. Sentence (any simple sentence)
5. Punctuated story (of two or more sentences)
6. Paragraphed story (two themes)

Message Quality

Record the number for the best description on the child's sample:

1. *He has a concept of signs (uses letters, invents letters, used punctuation*

2. He has a concept that a message is conveyed

3. A message is copied

4. Repetitive use of sentence patterns such as "Here is a…"

5. Attempts to record own ideas

6. Successful composition

Directional Principles

Record the number of the highest rating for which there is no error in the sample of the child's writing:

1. *No evidence of directional knowledge*

2. Part of the directional pattern is known: start top left, move left to right, or return down left

3. Reversal of the directional pattern (right to left and return down right)

4. Correct directional pattern

5. Correct directional pattern and spaces between words

6. Extensive text without any difficulties of arrangement and spacing of text

A simple curriculum-based measure of fluency is total number of words written during a short writing assignment. When fluency is the focus, misspellings, poor word choice, and faulty punctuation are not considered. Attention is only directed to the student's facility in translating thoughts into words. A baseline of at least three writing samples should be collected and the total number of words counted for each. For the purpose of evaluation, this total can be compared with those of proficient writers of the same age or grade level. However, total words may be used best in monitoring the student's progress, comparing performance with his or her own previous fluency.

A resulting IEP objective might be written like this: After a group prewriting discussion with the teacher, Daniel will write original narrative compositions of [40] words or more. A rough guideline for setting the criterion can be established from research reported by Deno, Mirkin, and Wesson (1984) and Parker and Tindal (1989):

a. If the total number of words is less than 20, aim for doubling it by the end of the school year.

b. If the number of words is between 25 and 30, aim for a 50% increase.

c. If the number of words is between 35 and 45, aim for a 25% increase.

d. If the number of words is greater than 50, choose another objective.

Content

Content is the second factor to consider in the writing product. Content features include the composition's organization, cohesion, accuracy (in expository writing), and originality (in creative writing).

General questions the classroom teacher can ask regarding a composition's organization include:

a. Is there a good beginning sentence?

b. Is there a clear ending?

c. Is there a logical sequence of subtopics or events?

Cohesion questions include:

a. Does the writer stick to the topic?

b. Is it clear what words like it, that, and they refer to?

c. Does the writer use key words that cue the reader to the direction of the discourse (First… , Then… , Therefore… , On the other hand…)?

Originality is assessed through questions like:

a. Did the writer attempt humor?

b. Did the writer present a unique point of view?

Analytical scales are the best way to lend some objectivity to evaluation of content. One can choose from a general rating scale, appropriate to almost any writing assignment, or one tailored to a specific genre or text structure. Spandel and Culham (1993) developed an analytical trait scoring guide for six aspects of writing, three of which address content: Ideas and content, organization, and voice. (Voice refers to the author's own unique personality, style, and honesty reflected in the writing.) Each of these traits is scored on a five-point scale. For example, organization is scored using the following guidelines:

A composition that is somewhat better organized than described by the guidelines for 3 but does not quite fit the descriptors for 5 would receive a rating of 4. Similarly, a rating of 2 falls between the descriptors for 1 and 3.

Analytical scoring guidelines such as these are used in many state writing assessments. There are two limitations to scales such as these. First, teachers must spend many hours learning the rubrics and discussing student compositions in order to establish any degree of integrater reliability. Second, these scales may not be sensitive enough to measure growth in students with emerging literacy skills who are unable to achieve a rating above 1 or-at the most-2.

For many students, writing instruction begins with smaller units of discourse, such as a paragraph. Welch and Link (1992) recommended an informal paragraph assessment that focuses on each of a paragraph's three parts: topic sentence, supporting sentences, and clincher sentence (Figure 3). Each part can receive a point for its existence, its form (grammatical correctness), and its function (relevance to the topic). Both topic sentence and clincher sentence can earn only one point for each of the three criteria, but up to three supporting sentences can be scored for existence, form, and function. This scale could be used to evaluate almost any kind of paragraph.

Figure 3. Informal assessment of a paragraph composition

Source:Welch, M. & Link, D.P. (1992) Informal assessment of paragraph composition. Intervention in School and Clinic, 27(3), 145-149.

Saguaro Cactus. The large cactus you see in pictures the desert is saguaro cactus. The Squaro cactus is very painfull if you toutch it. But it isn't as painful as being stabbed with a knife. It is against the law kill saguaros in the desert. I have seen som with about therty arms.

TOPIC SENTENCE:

Existence	1			(A topic sentence was written, but it was not grammatically correct.)
Form	0			
Function	1			
SUPPORTING SENTENCES:				
Existence	1	1	1	(Scored on the 2nd, 3rd, and 4th sentences.)
Form	1	1	0	
Function	1	0	1	(The 3rd sentence does not support the topic. The 4th is not grammatical.)
CLINCHER SENTENCE:				
Existence	0			No clincher sentence was written.
Form	0			
Function	0			
Total points earned	=	9		
Total points possible	=	15		
Total points earned Total points possible	x	100	=	60%

Writing instruction for students with special needs also may focus on specific text structures. An example of a structure-specific scale is one that Isaacson (1995) devised for evaluating factual paragraphs written by middle school students (Figure 4). Isaac son's scale reflects the conceptual definition of fact paragraphs taught to the students: (a) A fact paragraph has more than one sentence; (b) The first sentence tells the topic; (c) All other sentences are about the topic; (d) Sentences tell facts, not opinions; and (e) The most important information is given first. Judgments of factual accuracy and fact vs. opinion make the scale specific to factual paragraphs.

Figure 4. Analytical scale for factual paragraphs

Content			
Does the first sentence tell the topic?		0	1
Are all the other sentences about the topic?		0	1
Do the sentences tell about facts, not opinions?		0	1
Are the facts accurate?	0	1	2
0 = Some facts are clearly inconsistent with source material 1 = Some facts are questionable (content not covered in source material 2 = All facts seem accurate			
Is amount of information sufficient?		0	1
0 = Very little information given to reader or information is of trivial nature 1 = Sufficient information is provided			
Is information presented in logical order?	0	1	2
0=Random or stream-of-consciousness order 1=Some improvement possible 2 = Clear, logical order			

Is the most important information or main idea first?		0	1
TOTAL SCORE	____/ 9		

Harris and Graham (1992) provided another example of a structure-explicit measure for assessing the inclusion and quality of eight story elements in stories written by students with learning disabilities: introduction of the main character, description of the locale, the time in which the story takes place, a precipitating event (or starter event), the goal formulated by the character in response to the starter event, action(s) carried out in an attempt to achieve the goal, the ending result, and the final reaction of the main character to the outcome. Each story element receives a numerical score for its inclusion and quality of development. The validity of the scale was demonstrated by its correlation with Thematic Maturity scores on the Test of Written Language and holistic ratings of story quality (Graham & Harris, 1986). A resulting IEP objective for content might read: Using a story map, John will plan, write, and revise a story which includes a description of the character, setting, problem or goal, two or more events, and conclusion. (A story map is a planning sheet that prompts students to think about and write down their ideas concerning the character, setting, and other components of a good story before they write.)

Conventions

In order to fulfill the communicative function of writing, the product must be readable. Writers are expected to follow the standard conventions of written English: correct spelling, punctuation, capitalization, and grammar and legible handwriting. Consequently, even if the message is communicated, readers tend to be negatively predisposed to compositions that are not presentable in their form or appearance. Teachers traditionally have been more strongly influenced by length of paper, spelling, word usage, and appearance than by appropriateness of content or organization (Charney, 1984; Moran, 1982).

Counting correct word sequences is one quantitative method of measuring and monitoring students' use of conventions. Correct word sequences (CWS) are two adjacent, correctly spelled words that are grammatically acceptable within the context of the phrase (Videen, Deno, & Marston, 1982). Capitalization and punctuation also can be considered within the sequence. To calculate the proportion of CWS:

a. *Place a caret (^) over every correct sequence between the two words that form the sequence.*

b. *Place a large dot between every incorrect sequence. Place dots before and after misspelled words.*
Example: o my ^ dog o chasd o the ^ ball^.

c. *The first sequence is not comprised of two words but marks how the sentence was begun. (Sentence beginning to first word my is marked as an incorrect sequence because the M is not capitalized.) The last sequence is the last word to period, question mark, or other appropriate ending punctuation.*

d. *To control for length of composition either (a) time the writing sample for 3 minutes (the student may continue writing after a mark is made indicating the last word written in the 3-minute period) and/or (b) divide the number of CWS by the total number of sequences (correct and incorrect), which gives the proportion of CWS.*

e. *Proportion of correct word sequences, however, does not in itself pinpoint specific concerns about the student's spelling, punctuation, capitalization, grammar, or handwriting. The diagnostic function of assessment will only be met if the teacher also notes the student's strengths and weaknesses as in Figure.*

Figure 5. Diagnostic analysis of conventions

About Sell My Cow. I go to the Ranch at 5:30 in morning. I Ride my Horse with My Dad. get my Cow in the Barn. I Leave My cow and Calf. My DaD gave Shot to Calf. We took My Calf to Downtown. My fReind ride my horse. My horse is Black. My freind have red horse. But I need my cow to Born in feB 1st 1992. I am sell my Cow to calf for town But I have fun in Ranch in town. But I Like my money Back to for sell my Calf. But I need money Back to me. My Dad Siad no money back now Wait little to me.

Convention	Strengths	Errors
Spelling	Almost all words spelled correctly	Reversals in vowel combinations: ie/ei (friend), ai/ia (said)
Capitalization	Begins all sentences but one with uppercase letters.	Irregular use of uppercase where not required and even in middle of words. Month ("feB") not capitalized.
Punctuation	Correct ending punctuation in every sentence but one. Use of colon for time (5:30).	No comma in date (feB 1st 1992) or before the word but in compound sentence.
Grammar	Simple sentences are grammatically correct.	Inconsistent use of past tense. Missing articles ("My DaD gave Shot to Calf.") Problems with gerunds ("am sell"/am selling).
Handwriting	Legible. Good spacing and alignment.	

Like the other assessments discussed in this article, these methods can be useful for instructional planning. A resulting IEP objective addressing conventions, for example, might read: Using a 4-step editing strategy, Kevin will reread his composition checking for correct capitals, punctuation, spelling, and overall appearance, writing a final draft with 2 or less mechanical errors.

Syntax

As discussed previously, a child's early attempts at writing move from writing single words to writing word groups and sentences (Clay, 1993). Beginning writers often produce sentences that follow a repeated subject-verb (S-V) or subject-verb-object (S-V-O) pattern. The composition in Figure 5 was written by a ten-year-old female deaf student. The beginning of the composition reveals this typical repetitious pattern to a certain degree in its first few sentences: "I go… I Ride my Horse… [I] get my Cow… I Leave My cow…" A more mature writer will vary the sentence pattern and combine short S-V and S-V-O sentences into longer, more complex sentences.

Powers and Wilgus (1983) examined three parameters of syntactic maturity: (a) variations in the use of sentence patterns, (b) first expansions (six basic sentence patterns formed by the addition of adverbial phrases, infinitives, and object complements, and the formation of simple compound sentences), and (c) transformations that result in relative and subordinate clauses. Adapting Power and Wilgus's analysis of patterns suggests a simple schema for evaluating the syntactic maturity of a student's writing:

Fragment: A group of words that does not make a complete sentence

Examples: His old shirt. Nina and Fred too.

Level 1 Repetitious use of a single pattern (simple sentences). Example: I like my horse. I like my dog. I like my kitty. I like to feed my kitty.

Level 2 Use of a variety of simple sentence patterns.

Examples: I have a new toy. (S-V-O) It is big. (S-Vbe -Adj) It came in the mail. (S-V-PP)

Level 3 First expansions: (a) addition of an adverbial or gerund phrase, or (b) the making of a compound sentence by combining two simple sentences with the word and.

Examples: Our baby sitter sleeps all the time. To go faster, we push it. I ate the cookie and my brother ate the candy bar.

Level 4 Complex sentences (transformations in which one sentence is embedded within another as a subordinate clause)

Examples: The man wants to live where there is no pollution. Since John was late, we had to start without him.

Seldom does a student write sentences at only one level of syntactic maturity. One determines a syntactic level by analyzing all the sentences in the sample and summarizing them according to the type most often used. Occasionally one might characterize a student's syntactic level as being a transitional Level 2/Level 3 or Level 3/Level 4.

A resulting IEP objective for syntax might read: Daniel will plan, write, and revise a descriptive paragraph using mature sentences, at least half containing embedded clauses or adverbial phrases.

Vocabulary

The words used in a student's composition can be evaluated according to the uniqueness or maturity of the words used in the composition. Both quantitative and qualitative methods can be used to evaluate vocabulary. Quantitative methods include calculating the use of unrepeated words in relation to the total number of words, such as Morris and Crump's (1982) corrected type-token ratio. A simpler classroom-based method of looking at vocabulary is to simply make note of words used repetitiously (over-used words) as well as new and mature words the student uses.

Example: Over-Used Words: New Mature Words:

1. awesome
2. inspiring

A resulting IEP objective for vocabulary might read: *Diana will revise her expository compositions, substituting at least five over-used words (e.g., is) for more interesting action words.*

Taking into account the purpose

Being skilled is not just knowing how to perform some action but also knowing when to perform it and adapt it to varied circumstances (Resnick &

Klopfer, 1989, p. 4). Being a skilled writer requires knowing how to employ the writing process across a range of writing tasks and adapt the process to the specific purpose for writing.

Instruction often begins with story structures because they represent the genre most familiar to children. Children also use and depend upon narrative as their principal mode of thinking (Moffett, 1983). However, several educators (Hennings, 1982; Sinatra, 1991; Stotsky, 1984) have called for more emphasis on descriptive and expository text structures which relate more closely to real life writing tasks. Different purposes for writing call for different text structures. Writing a story calls for a narrative text structure that includes a character, setting, problem, etc. Writing about one's beliefs calls for a persuasive text structure that includes discussion of the problem, statement of belief, two or three reasons for the belief, facts and examples that support the reasons, etc.

Assessment of writing skills, therefore, should take into account a variety of purposes and text structures. Purposes and genres to consider include: personal narrative (my trip to the state fair), story narrative, descriptive, explanation of a process (how to give your dog a bath), factual report, letter, compare-contrast (compare the Allegheny Mountains with the Rocky Mountains), and persuasive.

Summary

Simple curriculum-based assessments can be used to assess the writing process and products of students with learning disabilities, as well as take into account purpose. The assessments recommended in this article also adequately fulfill the purposes of assessment as discussed at the beginning of the article: identifying strengths and weaknesses, planning instruction to fit diagnosed needs, evaluating instructional activities, giving feedback, monitoring performance, and reporting progress. A teacher might use these methods at the beginning of the year to do a quick sizing-up of student instructional needs. The process checklist in Figure 1 gives the teacher important diagnostic information about the strategies a student does or does not use when writing.

A quick assessment of product variables from the first two or three writing assignments also gives the teacher important diagnostic information about skill strengths and weaknesses. The teacher then should use the initial assessment to identify instructional targets. Some students, for example, may do pretty well at planning their composition, but do little in the way of effective editing. Other students may have creative ideas, but need considerable work on conventions. Some students may do pretty well with writing stories, but need to learn how to write factual paragraphs.

All classroom-based assessment should involve the student. Self-assessment helps students take ownership for their own writing and helps them internalize the strategies they are learning. The teacher's feedback should be given judiciously: generous in the encouragement of ideas and improved skills, but cautious in correction. Corrective feedback should only focus on those few skill targets that have been addressed in instruction.

Simple classroom-based methods also can be used to monitor student performance and report progress. Figure 6 is an assessment summary sheet that could be used to give a profile of a student's skills across a variety of writing purposes and genres. In an assessment portfolio the summary sheet would be accompanied by representative samples of a student's writing with both the student's and teacher's evaluations. After an initial assessment of student strengths and weakness across fluency, content, conventions, syntax, and vocabulary, the teacher would not necessarily need to monitor all the product factors, just those that focus on the student's greatest challenges and priority instructional objectives.

Figure 6. Assessment summary sheet

Writing Portfolio Summary	
Student:	Teacher:
Date:	Genre:
Fluency	

Number of Words	
Approximate Time	
Content	
Structure (Beginning, middle, end; story schema or other text structure)	
Cohesion (Adherence to topic; use of key words)	
Originality (Unique point of view; attempts at humor)	
Conventions	
% Correct Word Sentences	
Spelling Problems, punctuation or capitalization errors, grammar, other	
Syntax	
% Fragments	
Level 1 (simple repeated)	
Level 2 (simple varied)	
Level 3 (expansions)	
Level 4 (complex)	
Vocabulary	
Unique/Mature Words	

In conclusion, on-going assessment of writing is integral to effective teaching of writing. A teacher cannot make an appropriate instructional match between a student's skills and appropriate tasks without assessment. A teacher cannot ensure a student's success and make necessary adjustments in instruction without engaging in frequent assessment. Careful, thorough assessment of a student's writing requires that the teacher have a sound conceptual model of written expression taking into account process, product, and purpose.

GLOSSARY OF ASSESSMENT TERMINOLOGY

Accountability

This term has dominated educational reform for at least the past decade. In its best sense, it means shared responsibility for constantly improving educational practices and short- and long-term educational consequences such as student learning and the qualities of the society the students develop. Policymakers, researchers, administrators, families, community members, teachers, and students all share this responsibility. Often, however, accountability focuses on the shortterm responsibilities of teachers and students, such that primarily teachers and students experience the consequences when there are changes in achievement as measured by high-stakes tests. When teachers and students are held account able only for short-term consequences, such as what can be measured on a test, longer term goals, particularly those not easily measured on a test, tend to be neglected. When only a subset of the community feels responsibility for educational improvement, education will not be well served and burn-out is likely to occur. An analogous situation would be holding a doctor accountable for a child's physical and mental health when the child has no health insurance (and therefore does not seek regular medical care) and his family's eating, exercising, and interaction patterns are not under the doctor's control.

Aggregation

In assessment, aggregation is the process of collecting data for the purpose of making a more general statement. For example, it is common practice for school districts to add together all students' test scores to find the average performance of students in the district. This process strips away all of the differences among the various cultural groups, schools, and students within the district in order to make the general statement. Even an individual student's test score is a result of aggregating all the items to which the student responded on the test to make a general statement about a student's "ability." It is also common to "disaggregate" scores to see how subgroups performed within the larger group or to investigate the students' performance in various subareas of reading (e.g., word identification, vocabulary, comprehension).

There are powerful tensions around aggregation reflecting, on the one hand, the need to make general statements about students, teachers, and schools and, on the other, the problem of stripping away the particulars of individual performances and situations in the process. Not everyone agrees that it is reasonable to reduce students or schools to numbers—let alone the purposes for or the grounds on which that might be done. It is often argued that administrators need highly aggregated data to make programmatic and budgetary decisions.

However, both in education and in industry, administrators make different decisions when facing aggregated data than they do when presented with data about individual people and situations. Decision making needs to consider a balance of both kinds of data.

Authentic Assessment

For assessment to be considered authentic, it must include tasks that are a good reflection of the real-world activities of interest. This term arose from the realization that widely employed assessment tools generally have been poor reflections of what literate people actually do when they read, write, and speak. The logic of authentic assessment suggests, for example, that merely identifying grammatical elements or proofreading for potential flaws does not yield an acceptable measure of writing ability. Writing assessment tasks should reflect the audiences and purposes expected in life outside of school, with the real challenges those conditions impose. Similarly, reading very short passages and answering a limited number of multiple-choice questions is not a good measure of what literate people normally do when they read. Authentic assessments of reading employ tasks that reflect real-world reading practices and challenges. The authenticity of an assessment is very much a matter of the extent to which the assessment task measures what it purports to measure – a matter of construct validity.

Criterion-Referenced Assessment

We assess for particular purposes. When we want to know what children know and can do in a given domain, particularly whether they perform at a defined level on a specific task, we choose criterion-referenced assessment. Items in a criterion-referenced assessment are chosen because they discriminate what a person (or group) knows and can do and who has and has not reached a criterion level of performance. They are not chosen because they discriminate among individuals in determining who is better than whom. An item that genuinely measures a particular skill would not be eliminated from an assessment because everyone got it right. For example, a driver's test intends to determine whether a person is knowledgeable and capable enough to be allowed on the road, not whether one driver is more accomplished than another.

To be criterion referenced, a test must clearly define the characteristics that go into acceptable performance. In literacy, criterion-referenced assessments commonly compare students' performance on a specific task against established benchmarks. These benchmarks or criteria can be expressed as numerical ranges that define levels of achievement. For example, an 80–85 score may mean strong performance among levels of achievement ranging from unsatisfactory to outstanding. Criterion-based assessment can also involve holistic scoring of writing, for example, where a score is based on a set of pre-established criteria.

Curriculum

We can think of curriculum as having three components: (1) the envisioned curriculum, (2) the enacted curriculum, and (3) the experienced curriculum. The envisioned curriculum is the intended proficiency of students as a consequence of instruction and participation in classroom events. The enacted curriculum is the daily attempt in classrooms to put the envisioned curriculum into practice.

The experienced curriculum is the sense the learner makes of the enacted curriculum in the classroom and, thus, is constructed within the language of that classroom. For example, it is possible to intend to teach a particular lesson (e.g., authors' perspective) but that students not learn the lesson—either because it is not taught well (e.g., insufficient modeling, practice, support) or because the experiences of the students don't support the learning (e.g., they aren't provided with materials and experiences that invite perspective taking). As another example, if most of the reading material in one class includes racial or gender stereotypes, then that is likely to be reflected in students' learning. By contrast, students are likely to construct different knowledge about human relationships from a more balanced selection of reading material. However, the knowledge and attitudes students construct from those works are strongly influenced by the way teachers talk about them, the way teachers and other students respond to one another, and the nature of group discussions. Ultimately, it is the experienced curriculum that is our concern, and that is why students must be our primary curricular informants. However, the discrepancies among envisioned, enacted, and experienced curricula are what drive curriculum inquiry and the process of assessment.

Curriculum-Based Measurement (CBM)

This form of measurement was developed to help teachers evaluate a student's rate of growth in learning to read. The original idea was to have assessments that were embedded in the curriculum so they not only took no time away from teaching and learning but also did not distract teachers from the larger instructional picture. Originating in special education, a CBM of oral reading measures the number of words a child can read accurately in a minute from a standardized text (though there are comparable measures in spelling and writing). CBM assumes that a proxy variable, reading speed and accuracy (often mistakenly referred to as oral reading fluency), is an effective estimate of the larger construct of reading achievement and that the use of such estimates positively directs instruction.

Because these assessments now use texts and word lists that are standardized and that are not part of the curriculum, the term curriculum based is no longer particularly applicable. Other assessments not normally subsumed under the category of curriculum based, such as running records of children's reading and evidence of student work collected for a portfolio, are more clearly curriculum based since they are taken while the children are working within the actual classroom curriculum.

Equity

Issues of fairness surround literacy assessment. Testing originated as a means to control nepotism in job selection, providing an independent perspective on selection to uphold fairness. But equity cannot be assured through testing alone.

Those who control the assessment process control what counts, what is valued.

As we point out in this book’s Introduction, language and literacy assessment is laden with cultural issues and biases. Although equity cannot be assured through assessment, it must be pursued relentlessly in assessment and in schooling. It is more likely to be achieved through the involvement of multiple, independent perspectives than through the use of a single perspective.

Tests have traditionally been administered, their results published, and their impact on instruction instigated with little regard to issues such as cultural, economic, or gender equity. But many equity issues affect assessment, rendering comparisons difficult and often invalid. Because traditional tests frequently reflect narrow cultural values, students and schools with different backgrounds and concerns often have not been fairly assessed.

Being equitable requires ensuring comparable educational experiences for those facing similar assessments, particularly in certification or gate-keeping situations. Questions of access to sound instruction, appropriate materials, and enriching learning opportunities are critical. Educators have become increasingly aware of the connections between assessment results and levels of safety, health, and welfare support in addition to physical accessibility.

Formative Assessment

Formative assessment, often referred to as assessment for learning, is the assessment that is done before and during teaching to inform instruction. It is assessment that informs instruction. Formative assessment includes things like teacher–student conferences, listening in on student book discussions, taking records of children’s oral reading, examining students’ writing pieces, and so forth. Though these assessments might be standardized, they often are not. To be formative, an assessment must affect instruction.

Compare to summative assessment.

High-Stakes Testing

These tests have significant consequences for those viewed as responsible for performance on the tests, and also for the student. For example, tests that determine whether one is accepted or rejected into the military, a university, or an educational program have significant consequences for the individual test takers. Consequences can be felt among a broader range of people, however.

In the United States today, student test scores are not only used to determine whether children move on to the next grade level, but they also influence where educational resources are allocated and whether a school may continue to operate. Often, local news media publish school test scores, and property values are affected when families make decisions about where to purchase a home based on the local school’s performance. When major consequences—such as the adjustment of teachers’ salaries –are attached to their students’ test scores, teachers will emphasize in their instruction what the test measures and reduce their emphasis on areas not covered by the test. This has consequences for the breadth of the curriculum and, thus, for the students’ lives.

Both the National Council of Teachers of English and the International Reading Association have position statements regarding high-stakes testing.

Both organizations recommend minimizing the stakes where possible and not relying on single measures, particularly when the stakes are high.

Inquiry

The process of inquiry begins with a genuine question, that is, a question that motivates the questioner to persist in seeking the answers. Authentic questions are rarely well formulated or structured at the outset. Rather, structure emerges through the process of inquiry. Inquiry is not merely a matter of asking and answering questions. It is a way of engaging the world and other people.

Communication and social relationships play an important role in inquiry as questioners seek the advice and expertise of peers and more knowledgeable others, share their findings, reflect upon the results of the inquiry, and take up new questions that arise.

In a traditional view of classroom learning, teachers deliver information. They ask the children questions to which they already know the answers, and the students are to show they know the correct answers as well. This approach has not been very successful at helping all students become the critical, creative, and socially responsible citizens our society needs. In an inquiry classroom, on the other hand, students and teachers have a different relationship. Teacher and peers are resources for helping students answer their own questions. The community relationships are different. Instruction is based on engaging in sustained examination of personally significant topics.

Assessment as inquiry involves the same principles. It requires teachers to pose questions about the teaching and learning in their classrooms and to seek answers to those questions using assessment data and the resources of their learning community.

Multimodal Literacy

For centuries, the book has been the central medium of communication, expressed on paper largely through the mode of writing. Today, the screen is becoming the dominant medium of communication, with increasing reliance on the mode of image. A mode is a resource for communication and representation. Examples include speech, dance, gesture, music, sculpture, photography, and writing. Humans may express themselves through a single mode, such as writing, but with growing frequency we combine modes to communicate. This results in multimodal texts such as a PowerPoint presentation or YouTube video that combines words, images, music, and movement, or an advertisement in which print and image are merged. Today's and tomorrow's learners need to acquire competence in this multimodal literacy.

Norm-Referenced Assessment

When we want to know how a child performs relative to other children in a particular domain, we use norm-referenced assessment. Items in a norm-referenced assessment are chosen because they discriminate between individuals rather than assessing what a person (or group) knows and can do. To make norm referenced assessments, assessment practices need to be standardized and test item selection must focus on maximizing the differences among individuals on a scale. An item that genuinely measured a particular skill but which all students got correct would not be used because it would not discriminate who was better than whom.

Norm-referenced interpretations are based on comparisons with others, usually resulting in a ranking. For example, a norm-referenced interpretation of a student's writing might assert that the sample is "as good as that of 20% of the students in that grade nationally".

Norm-referenced testing is the most prevalent form of large-scale testing, in which large groups of students take a test and the scores are grouped and interpreted in relation to other scores. In other words, the score of any student or group (school, district, state, or nation) has meaning only in relation to all the other scores of like entities (e.g., school to school, district to district, state to state). In order to make such comparisons, we have to make the assumption of "all else being equal," which is rarely justifiable. National norm-referenced tests assume that all students in our society have had similar cultural and curricular experiences. Uses of these tests also commonly ignore differences in curriculum, culture, gender, ethnicity, economic circumstance, per-pupil funding, and so forth.

The main advantage of such assessments is the simplicity of the linear scale. The seductiveness of this scale is also the main disadvantage, because the scores appear readily interpretable and objective. However, the score oversimplifies the complexities of literacy and assessment. Unfortunately, norm-referenced test scores often become the most important criterion for decisions about placement and promotion, which have a powerful impact on students' and teachers' lives.

Compare to criterion-referenced assessment.

Performance-Based Assessment

Performance-based assessment refers to assessment that involves the demonstration of a particular skill and often the process of accomplishing a performance specific to that skill. Performance assessments can include, for example, such complex activities as group collaboration to write and produce a play. The concept of performance-based assessment is related to the concept of authentic assessment in that it arose from a realization of the limitations of multiple-choice tests, and other assessments of complex skills, and the difficulty in making inferences about complex skills from such assessments.

Portfolio Assessment

A portfolio approach to assessment uses a systematic and multifaceted collection of work that represents a student's development. For example, a portfolio might include a range of writing pieces, a book log, self-reflections, group projects, and multimedia work. Because of the nature of the contents, portfolios are both curriculum based and performance based. A primary emphasis in most portfolio assessment is on student involvement and the development of self-assessment or reflectiveness. However, in some applications, portfolios can also include teacher and parent observations.

Reliability

Broadly speaking, reliability is an index of the extent to which a set of results or interpretations can be generalized over time, across tasks, and among interpreters. In other words, it is a particular kind of generalizability. For example, a common concern raised by newer forms of literacy assessment is whether different examiners, evaluating a complex response and using complex scoring criteria, will draw similar conclusions about a student's performance (whether an assessment will generalize across different examiners). Experience from scoring complex student writing samples suggests that high rates of agreement can be achieved when people are well trained in the application of specific criteria.

Another example of reliability is whether a score obtained by a student on a test would remain the same if the student took the test the following day, assuming no new learning has taken place—in other words, whether the performance generalizes over time. In general, the more samples of student work we collect, the more reliable and consistent an assessment will be.

Reliability is only important within the context of validity—the extent to which the assessment measures what it is supposed to measure and leads to useful, meaningful conclusions and consequences. Reliability does not guarantee a high quality assessment. It is possible that consistent scoring can be achieved on poorly designed tests or tests of trivial skills. Indeed, reliability is easiest to obtain on low-level skills.

Summative Assessment

Summative assessment, often referred to as assessment of learning, is the after the-fact assessment in which we look back at what students have learned, such as end-of-course or end-of-year examinations. The most familiar forms are the end-of-year standardized tests, though in classrooms we also assess students' learning at the end of a unit. These assessments are likely to be uniform or standardized.

Compare to formative assessment.

Validity

Historically, a common definition of a valid measure is that it measures the construct it purports to measure. This is called construct validity. For example, if we

RECOMMENDED LITERATURES:

1. Brown D.P., Nacino-Brown (1990) Effective Teaching Practice. Cheltenham:
2. Nelson Thornes Ltd.,
3. Johnson K. (2001). An Introduction to Foreign Language Learning and Teaching.Harlow: Pearson Education.
4. Khoshimov U., I. Yokubov, (2003) Ingliz tili o’qitish metodikasi Tashkent, Sharq.
5. Larsen-Freeman, D. (1986). Techniques and principles in language teaching. New York: Oxford University Press.
6. Fulcher, G. An English language placement test: issues in reliability and validity/ G.Fulcher. - Cambridge University Press, 2007.- 114 c.
7. Hughes, A. Testing for Language Teachers / A.Hughes. –Cambridge University Press, 2011.- 110 c.

ADDITIONAL SOURCES:

1. Black, Paul, & William, Dylan (October 1998). "Inside the Black Box: Raising Standards Through Classroom Assessment."Phi Beta Kappan. Available at PDKintl.org. Retrieved January 28, 2009.
2. Cottrell, S. (1999) The Study Skills Handbook. Hampshire: Palgrave
3. Committee on Standards for Educational Evaluation. (2003). The Student Evaluation Standards: How to Improve Evaluations of Students. Newbury Park, CA: Corwin Press.
4. Earl, Lorna (2003). Assessment as Learning: Using Classroom Assessment to Maximise Student Learning. Thousand Oaks, CA, Corwin Press. ISBN 0-7619-4626-8. Available at, Accessed January 23, 2009.
5. Joint Information Systems Committee (JISC). "What Do We Mean by e-Assessment?". Retrieved January 29, 2009.
6. Materials Evaluation and Design for Language Teaching (Edinburgh Textbooks in Applied Linguistics)

7. Mctighe, Jay; O'Connor, Ken (November 2005). "Seven practices for effective learning". Educational Leadership. 63 (3): 10–17.
8. Moskal Barbara M., & Leydens, Jon A (2000). "Scoring Rubric Development: Validity and Reliability." Practical Assessment, Research & Evaluation, 7(10). Retrieved January 30, 2009.
9. Nelson, Robert; Dawson, Phillip (2014). "A contribution to the history of assessment: how a conversation simulator redeems Socratic method". Assessment & Evaluation in Higher Education. 39 (2): 195–204.
10. Reed, Daniel. "Diagnostic Assessment in Language Teaching and Learning." Center for Language Education and Research, available at Google.com. Retrieved January 28, 2009.
11. Scriven, M. (1991). Evaluation thesaurus. 4th ed. Newbury Park, CA:Sage Publications.
12. Tanner, R &C. Green (1998). Tasks for Teacher Education: a Reflective
13. Approach. Longman
14. Thornbury, S. (1999) How to teach grammar. Longman Pearson
15. Thornbury, S. (2002) How to teach vocabulary. Longman Pearson.
16. Ur,P. (1996) A Course in Language Teaching: Practice and Theory. Cambridge: CUP
17. Woodward, T. (2001). Planning Lessons and Courses. Cambridge, CUP

Websites:

1. www.teachingenglish.org.uk
2. www.onestopenglish.com
3. www.businessenglishonline.net
4. www.elgazette.com
5. www.tesol.org
6. www.tefl.com
7. www.teachertrainingvideos.com
8. www.learnenglish.org.uk
9. www.educationuk.org

10. www.bbc.co.uk\worldservice\learningenglish/

11. www.channel4.com\learning\

12. www.better-english.com \exerciselist.html

13. www.bbc.co.uk\worldservice\learningenglish\business\index.shtml

14. www.englishclub.com\index.htm