DIKTAT

BAHASA INGGRIS

ENGLISH LEARNING ASSESSMENT

Disusun Oleh:

Diah Safithri Armin, M.Pd.

NIP. 199105282019032018

PROGRAM STUDI TADRIS BAHASA INGGRIS

FAKULTAS ILMU TARBIYAH DAN KEGURUAN

UNIVERSITAS ISLAM NEGERI

SUMATERA UTARA MEDAN

2021

SURAT REKOMENDASI

Saya yang bertanda tangan di bawah ini:

Nama : Rahmah Fithriani, Ph.D

NIP : 197908232008012009

Pangkat/Gol : Lektor/IIId

Unit Kerja : Prodi Tadris Bahasa Inggris Fakultas Ilmu Tarbiyah dan Keguruan

Menyatakan bahwa diktat saudara:

Nama : Diah Safithri Armin, M.Pd

NIP : 199105282019030218

Pangkat/Gol : Asisten Ahli/III b

Unit Kerja : Prodi Tadris Bahasa Inggris Fakultas Ilmu Tarbiyah dan Keguruan

Telah memenuhi syarat sebagai karya ilmiah (diktat) dalam mata kuliah English Learning Assessment pada Prodi Tadris Bahasa Inggris Fakultas Ilmu Tarbiyah dan Keguruan Universitas Islam Negeri Sumatera Utara Medan.

Demikian surat rekomendasi ini diberikan untuk dapat dipergunakan sebagaimana mestinya

Medan, 5 Mei 2021

Yang Menyatakan

Rahmah Fithriani, Ph.D.

NIP. 197908232008012009

ACKNOWLEDGMENT

Bismillahirahmanirrahim

First, all praise be to Allah SWT for all the opportunities and health that He bestows so that the writing of the English Learning Assessment handbook can be completed by the author even though it is still not perfect. This handbook is prepared as reading material for students of the English Education Department who take the English Learning Assessment course.

This handbook is prepared following the discussion presented in the lecture syllabus with additional discussions and studies. The teaching-learning activity is held for 16 meetings that discuss several topics using the lecture method, group discussions, independent assignments in compiling instruments for assessing students' language skills and critical journals, practicing using assessment instruments, and field observations.

The final product of the discussion of this handbook is an instrument for assessing students' language skills at both junior and senior high school levels and reports on the use of assessment instruments by English teachers in schools.

This book discusses several topics: testing and assessment in language teaching, assessing listening skills, assessing speaking skills, assessing reading skills, assessing writing skills, and testing for young learners.

The author realizes that this handbook is not perfect. Therefore, it is hoped that constructive suggestions will improve the contents of this book. Also, I would like to express my appreciation to my colleagues who helped and motivated me in the process of compiling this dictate.

Author,

Diah Safithri Armin, M.Pd

Table of Content

INTRODUCTION

In teaching English, assessing students' language skills is a crucial part of the learning process to know how far the students' skill have improved and to diagnose students' weakness, so the teacher can do better teaching to improve students' language proficiency. Assessment is always linked to test, and when people hear the word 'test' in classroom, they will think of something scary and stressful. However, what is exactly a test? Test is *a method of measuring a person's ability, performance, or knowledge in a specific domain*. First, a test is a method. It is an instrument—a series of methods, processes, or items—that allows the test-taker to execute. The process must be explicit and standardized to count as a test:

- multiple-choice questions with specified correct answers
- a writing prompt with a scoring rubric
- an oral interview based on a question script
- a checklist of planned responses to be filled out by the administrator

Second, a measurement must be calculable. Such tests measure general competence, while others focus on particular competencies or priorities. A multi-skill proficiency assessment assesses a broad level of ability, while a questionnaire on recognizing correct use of specific papers assesses individual abilities. The way the findings or measurements are communicated will vary. Some tests, such as a shot-answer essay exam given in a classroom, grant the test-taker a letter grade with negligible comments from the teacher. Others, such as large-scale quantitative tests, include a composite numerical ranking, a percentage grade, and perhaps several subscores. If an instrument does not specify a method of reporting measurement—a method of providing a result to the test-taker—then the procedure cannot be appropriately described as a test.

Also, a test assesses an individual's skill, expertise, or performance. The testers must identify the test-takers. What are their prior experience and educational backgrounds? Is the exam sufficient for their abilities? What do test takers do for their results?

A test tests accuracy, but the findings mean the test-taker skill or expertise, to use a linguistics term. The majority of language tests assess an individual's ability

to practice language, that is, to talk, write, interpret, or listen to a subset of language. On the other hand, it is not unusual to come across a test designed to assess a test-knowledge taker's of language: describing a vocabulary object, reciting a grammatical law, or recognizing a rhetorical characteristic of written discourse. Performance-based evaluations collect data on the test-actual taker's language use, but the test administrator infers general expertise from those data. A reading comprehension test, for example, could consist of many brief reading passages accompanied by a limited number of comprehension questions—a small sampling of a second language learner's overall reading activity. However, based on the results of that examination, the examiner can assume a degree of general reading skill.

A well-designed test is an instrument that gives a precise measure of the test-takers ability in a specific domain. The concept seems straightforward, but creating a successful test is a complex challenge that requires both science and art.

In today's educational practice, assessment is a common and often confusing word. You may be tempted to consider assessing and testing to be synonyms, but they are not. Tests are planned administrative procedures that arise at specific points in a program where students must summon all of their faculties to work at their best, recognizing that their reactions are being assessed and tested. On the other hand, assessment is a continuous phase that covers a much broader range of topics. When a student answers a challenge, makes a statement, or tries out a new word or structure, the instructor evaluates the student's success subconsciously. From a scribbled sentence to a structured essay, written work is a performance that is eventually evaluated by the author, the instructor, and potentially other students. Reading and listening exercises usually necessitate constructive output, which the teacher indirectly evaluates, but peripherally. A good teacher never stops assessing pupils, whether such tests are unintentional or intentional.

Tests are, therefore, a category of assessment; they are by no means the only type of assessment that an instructor should conduct. Tests can be helpful tools, but they are just one of the processes and assignments that teachers can use to evaluate students in the long run.

However, you might be wondering, if tests are made any time you teach something in the classroom, does all teaching require assessment? Are teachers actively judging pupils with no assessment-free interaction?

The response is dependent on your point of view. For optimum learning to occur, students in the classroom must be allowed to experiment, to test their ideas about language without feeling as though their general ability is being measured based on such trials and errors. In the same way, that tournament tennis players must have the right to exercise their skills before a tournament with no consequences for their final placement on the day of days, and learners must have chances to "play" with language in a classroom without being officially graded. Teaching establishes the practice games of language learning: opportunities for learners to listen, reflect, take chances, set goals, and process input from the "coach—and then recycle into the skills that they are attempting to master.

Chapter I
Testing and Assessment in Language Teaching

Competence

The students comprehend what testing and assessment is in language teaching and how to arrange valid and reliable English skill assessment instrument.

Definition and Dimension of Assessment

In learning English, one of the essential tasks that the teacher must carry out is an assessment to ensure the quality of the learning process that has been carried out. Assessment refers to all activities carried out by teachers and students as their own self-evaluation to obtain modified feedback on their learning activities (Black and William, 1998, p. 2). In this sense, there are two important points conveyed by Black and William; the first assessment can be carried out by teachers and students, or students with students. Second, the assessment includes daily assessment activities and more extensive assessments, such as semester exams or language proficiency tests (TOEFL, IELTS, TOEIC).

According to Taylor and Nolen (2008), assessment has four basic aspects: assessment activities, assessment tools, assessment processes, and assessment decisions. Activity assessment, for example, when the teacher holds listening activities. Listening activities can help students improve their listening skills if they are carried out with the right frequency. Thus the teacher can find out whether the instruction used is successful or still requires more instruction. Assessment tools could support the learning process if the tools used help students understand essential parts of the lesson and good work criteria. Also, an assessment tool is vital in gathering evidence of student learning. Therefore, it is imperative to determine the appropriate assessment tool by the skill to be assessed.

The assessment process is how teachers carry out assessment activities. In the assessment process, feedback is expected to help students be more focused and

better understand what is asked for the given assignment. Therefore, feedback is central to the assessment process.

Then, the assessment decision is a decision made by the teacher following the assessment reflection results. Assessment decisions will help students in the learning process if the value obtained from the assessment is valid or describes the students' abilities. An example of an assessment decision is what will be done in the following learning process, is there a part of the material that has been taught that must be deepened or can continue with the following material.

Assessment has two dimensions:

1. Assessment for learning. Assessment for learning is the process of finding and interpreting the results of the assessment, which are used to determine where students are "where" in the learning process, "where" they have to go, and "how" students can reach their intended places.
2. Assessment of learning. This dimension refers to the assessment carried out after the learning process to determine whether learning has taken place successfully or not.

In the immediate learning process in the field, teachers should combine the two dimensions above.

Assessment can also be defined in two forms, namely formative assessment, and summative assessment. Black and William (2009) define formative assessment as:

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction. (p. 9)

Meanwhile, according to Cizek (2010), the formative assessment is:

> The collaborative processes engaged in by educators and students for the purpose of understanding the students' learning and conceptual organization, identification of strengths, diagnosis of weaknesses, areas of improvement, and as a source of information teachers can use in instructional planning and students can use in deepening their understanding and improving their achievement. (p. 6)

Formative assessment is part of the assessment for learning where the assessment process is carried out collaboratively, and the resulting decisions are used to determine "where" students should go. Therefore, the formative assessment does not require a numeric value. In contrast to formative assessment, summative

assessment is carried out to assess the learning process, skills gained, and academic achievement. Usually, a summative assessment is carried out at the end of a lesson or project, semester, or the end of the year. So, summative assessment is under the assessment of learning.

In general, summative assessment has three criteria:

1. The test for the given assignment is used to determine whether the learning objectives have been achieved or not.
2. Summative assessment is given at the end of the learning process so that the summative assessment is an evaluation of learning progress and achievement, evaluation of the effectiveness of learning programs, and evaluation of improvement in goals.
3. Summative assessment uses values in the form of numbers which will later be entered into student report cards.

Purposes of Assessment

The main objectives of the assessment can be divided into three things. First, the assessment aims to be instructional. Assessments are used to collect information about student achievement, both skills, and learning objectives. Thus, to meet the objectives of this assessment, teachers need to use an assessment tool. An example of achieving the purpose of this assessment is when the teacher gives assignments to students to find out whether students have understood the material being taught. The second objective of the assessment is student-centered. This objective relates to the use of a diagnostic assessment, which is often confused with a placement test. Diagnostic assessment is used to determine students' strengths and weaknesses (Alderson, 2005; Fox, Haggerty and Artemeva, 2016)

Meanwhile, the placement test is used to classify students according to their development, abilities, prospects, skills, learning needs. However, both placement tests and diagnostics assessments are aimed at identifying student needs. Finally, the assessment aims for administrative needs. It is related to giving grades to students in number form (e.g., 80) and letters (e.g., A, B) to summarize student learning outcomes. Numbers and letters are used as a form of statement to the public, such as students, parents, and the school. Therefore, assessment is the most

frequently used method and often directly affects students' self-perceptions, less motivation, curriculum expectations, parental expectations, and even social relationships (Brookhart, 2013).

By knowing the purpose of the assessment being carried out, the teacher can make the right assessment decision because the assessment's purpose affects the frequency and timing of the assessment and the assessment method used, and how it is implemented. The most important thing is to consider the objectives of the assessment, effects, and other considerations in carrying out the assessment, both the tools and the implementation process. Thus, teachers can ensure the quality of the assessment class.

Assessment Quality

In implementing assessments in the classroom, teachers must ensure that the assessments carried out are of good quality. For that, teachers need to pay attention to several fundamental aspects of assessment in practice. The first is alignment. Alignment is the level of conformity between assessment, curriculum, instruction, and standard tests. Therefore, teachers must choose the appropriate assessment method in order to be able to reflect on whether the objectives and learning outcomes have been achieved or not.

The second is validity. Validity refers to the suitability of conclusions, use, and assessment results. Thus, high-quality assessments must be credible, reasonable, and based on the results of the assessment.

The third is reliability. An assessment is only said to be reliable if it has stable and consistent results when given to any student with the same level. Reliable is needed to avoid errors in the assessment used.

Next up are the consequences. Consequences are the result of use or errors in using the results of the assessment. Consequences are widely discussed in recent research, focusing on the interpretation of the dark effect test, which is then used by stakeholders (Messick, 1989), which has led to the term washback and is often used in linguistics studies (Cheng, 2014).

Next is fairness. Fairness will be achieved if students have the same opportunity to demonstrate learning outcomes and assessments by producing

equally valid scores. In other words, fairness is to give all students equal opportunities in learning. To achieve fairness, students must know the learning targets, the criteria for success, and how they will be assessed.

The Last is practical and efficient. In the real world, a teacher has many activities to significantly influence the teacher's decision to determine the time, tools, and assessment process. Thus, the question arises whether the resources, effort and time required are precious for the assessment investment? Therefore, teachers need to involve students in the assessing process, for example, correcting students' written drafts together. Besides saving time for teachers, checking student manuscripts Together can train students to be responsible with their own learning.

A teacher needs to understand the testing and assessment experience in order to continue a valid examination. It is because examinations can assist teachers in studying and reflecting on assessments that have been carried out, whether they have been well designed, and how well the assessment tools assess students' abilities. Studying the assessment experience that has been done helps teachers find out and consider construct-irrelevant variances that occur during the assessment process. For example, when the teacher tests students' listening skills. The audio record sound was clear for the students sitting in the front row, but the back row students could not hear the audio. Thus, the student's sitting position and the clarity of the audio record affect the student's score. Therefore, sitting position and audio record sound quality are construct-irrelevant variance that the teacher must consider. Another example of another construct-irrelevant variance is that all students' test results are good because of the preparation or practice for the test, even the level of self-confidence and emotional stability of students.

Philosophy of Assessment

In assessing students, teachers will be greatly influenced by the knowledge, values , and beliefs that shape classroom actions. This combination of knowledge, values , and beliefs is called the philosophy of teaching. Therefore, a teacher needs to know the philosophy of the assessment he believes in. To build a philosophy of assessment, teachers can start by reflecting on their teaching philosophy and

considering the assumptions and knowledge teachers have when carrying out assessments in everyday learning.

The teacher's amount of time preparing the learning plan and implementing it, including assessing the teacher, makes the teacher "forget" and does not have time to reflect on the assessment he has done. Why use this method? Why not use another method? Don't even have time to discuss it with other teachers. The number of administrative activities that the teacher has to do also adds to the teacher's busyness. Several assessments conducted by external schools, such as national exams, professional certificate tests, proficiency tests, have made teachers make special preparations individually. Research conducted by Fox and Cheng (2007) and Wang and Cheng (2009) found that even though students face the same test, the preparation is different and unique. Also, several external factors such as textbooks, students' proficiency, class size, and what teachers believe in teaching and learning English can influence teachers in choosing assessment activities.

Teacher beliefs can be in line with or against curriculum expectations that shape the context for how teachers teach and assess in the classroom (Gorsuch, 2000). When the conflict between teachers' beliefs and the curriculum is large enough, teachers will often adapt their assessment approach to align with what they believe.

In the English learning curriculum history, three educational philosophies form the agenda of mainstream education (White, 1988), classical humanism progressivism, and reconstructionism. White also explained that there are implicit beliefs, values, and assumptions in the three philosophies. Classical humanism holds the values of tradition, culture, literature, and knowledge of the language. This philosophy curriculum's main objective is to make students understand the values, culture, knowledge, and history of a language. Usually, students are asked to translate text, memorize vocabulary, and learn grammar. Because this philosophy highly upholds literature's value, most of the texts used will relate to literature and history. For performance expectations, the new assessment is declared accurate if students get a value of excellence.

Progressivism views students as individual learners so that a curriculum that uses this philosophy will make students the centre of learning. However, the

progressivism curriculum asks teachers to define learning materials and activities. So, the teacher can analyse student needs or evidence that shows student interest and performance to determine the direction and learning activities. Also, this curriculum sees students as unique learners based on their backgrounds, interests, and self-motivation. Therefore, the teacher can negotiate with students about what language learning goals and experiences the students want. This negotiation will later become the basis for teachers in preparing assessments to see differences in developments at the current level with language proficiency, proficiency, and expected performance.

In the progressivism curriculum, language teachers have a role to play (Allwright, 1982): helping students know which parts of language skills need improvement and elaborating strategies for fostering a desire to improve students' abilities. Therefore, all classroom activities depend on daily assessments of the extent to which students achieve agreed-upon learning objectives both individually and in groups.

A curriculum that adopts the philosophy of reconstructionism determines the learning outcomes according to the course objectives. Learning outcomes are the teacher's reference in determining student learning activities and experiences, what students should know and do at the end of the learning process. Therefore, some reconstructionism curricula are mastery-based in which the reference is success or failure, while others take the percentage of student success and compare them with predetermined criteria (such as the Common European Framework of Reference; the Canadian Language Benchmarks) as a reference. The completeness criteria are adjusted to the level of difficulty of the exercises given to students.

In addition to the philosophy of the Language learning curriculum put forward by White, there is another curriculum, namely Post-Modernism or Eclecticism. This curriculum emphasizes uniqueness, spontaneity, and unplanned learning for everyone's reasons, the interaction between students and learning activities is unique. Students in this curriculum are grouped according to their interests, proficiency, age, and others.

Washback

The term washback emerged after Messicks (1989) introduced his theory of the definition of validity in a test. Messick's concept of validity refers to the value generated from a test and how these results affect both individuals (students) and institutions. Messick (1996: 241) says that 'washback refers to the extent to which the introduction and use of a test influences language teachers and learners to do things that they would not otherwise do that promote or inhibit language learning'.

In the following years, Alderson and Wall (1993) formulated several questions as hypotheses that can investigate the washback of a test. Including the following:

1. What do teachers teach?
2. How do teachers teach?
3. What do students learn?
4. How the rate and sequence of teaching?
5. How the rate and sequence of learning?
6. What are teachers' and students' attitudes towards content, methods, and other things in the learning and teaching process?

Washback can implicitly have both negative and positive effects on teachers and students, but it is not clear how it works. Some students may have a more significant influence on a test than other students and teachers. Washback can appear not only because of the test itself but also because of the test's external factors, such as teacher training background, culture in schools, facilities available in the learning context, and the curriculum's nature (Watanabe, 2004a). Therefore, washback does not necessarily appear as a direct result of a test (Alderson and Hamp-Lyons, 1996; Green, 2007). The results showed no direct relationship between the test and the effects produced by the test (Wall and Alderson, 1993, 1996). Wall and Alderson (1996: 219) conclude from the results of their research conducted in Sri Lankan:

> the exam has had impact on the content of the teaching in that teachers are anxious to cover those parts of the textbook which they feel are most likely to be tested. This means that listening and speaking are not receiving the attention they should receive, because of the attention that teachers feel they must pay to reading. There is no indication that the exam is affecting the methodology of the classroom or that teachers have yet understood or been able to implement the methodology of the text books.

Nicole (2008) conducted a study on the effect of local tests on Zurich's learning process using surveys, interviews, and observations. Nicole found that the test involved a wide range of abilities and content, which was also able to help teachers improve their teaching methods. In this case, Nicole as a researcher, simultaneously participates in teaching in collaboration with other teachers in proving that the test has a positive impact on the learning process. The example of this research can be a reference for teachers to learn washback in the context of their respective professions.

In researching the washback effect of tests in familiar contexts, extreme caution should be exercised. Watanabe (2004b: 25) explains that researchers who understand the context of their research cannot see the main features of the context, which are essential information in interpreting the washback effect of a test. Therefore, the researcher must make himself unfamiliar with the context he is researching and use curiosity to recognize the context that is being studied. Then, determine the research scope, such as a particular school, all schools in an area, or the education system. Also, the researcher needs to describe which aspects of washback interest the researcher to answer the question *'what would washback look like in my context?'* (Wall and Alderson, 1996: 197-201).

The next thing that is important to note is what types of data can prove that washback is running as expected (Wall, 2005). Usually, the data obtained follows the formulation of the problem, which can be collected through various techniques, such as surveys and interviews. Interviews provide researchers with the opportunity to dig deeper into the data obtained through surveys. This technique can also be applied in Language classes. Besides, in gathering information about washback, researchers can also make classroom observations to see first-hand what is happening in the classroom. Before making observations, it would be better if the researcher prepares a list of questions or things observed in the classroom. If needed, the researcher can conduct a pilot study to find out whether the questionnaire needs to be developed or updated. Instrument analysis is also needed to detect washback, such as lesson plans, textbooks, and other documents.

In the application of assessments in the classroom, teachers are asked to develop a curriculum and organize learning activities, including assessments, which

cover all the skills and abilities specified in the standard. The test is indeed adjusted to the curriculum standards, but the test will be said to be successful if students can pass the test without taking a particular test preparation program. Therefore, tests shape the construct but do not dictate what teachers and students should do. In other words, tests are derived from the curriculum, and the teacher acts as a curriculum developer so that the methodology and teaching materials can differ from one school to another. So, when the contents of the test and the instructions' contents are in line, the teacher succeeds in compiling the material needed to achieve the learning objectives. Koretz and Hamilton (2006: 555) describe tests with material said to be compatible when 'the knowledge, skills and other constructs measured by the tests will be consistent with those specified in the [content] standards.' However, instead of being called "content standards" for language classes, it is more correctly called "performances standards" or progression. It is because language learning content arranged in performance levels is called a task that is adjusted to the level of difficulty. The following are examples of some of the standards in the Language class.

Table 1.1 Standards for Formatting Writing, language arts, grades 9-12 (WIDA, 2007: 59 in Fulcher, 2010: 284)

	Level 1: Entering	Level 2: Beginning	Level 3: Developing	Level 4: Expanding	Level 5: Bridging
Example Genre: Critical Commentary	Reproduce comments on various topics from visually supported sentences from newspapers or websites	Produce comments on various topics from visually sup- ported para- graphs from newspapers or websites	Summarize critical commentaries from visually supported newspaper, website or magazine articles	Respond to critical commentaries by offering claims and counter- claims from visually supported newspaper, website or magazine articles	Provide critical commentary commensurate with proficient peers on a wide range of topics and sources
Example topic: Note taking	Take notes on key symbols, words of phrases from visuals	List key phrases or sentences from discussions and models (e.g. on the	Produce sentence outlines from discussions, lectures or readings	Summarize notes from lectures or readings in paragraph form	Produce essays based on notes from lectures or readings

	pertaining to discussions	board or from overhead projector)			
Example topic: Conventions and Mechanics	Copy key points about language learning (e.g. use of capital letters for days of week and months of year) and check with a partner	Check use of newly acquired language (e.g. through spell or grammar check or dictionaries) and share with a partner	Reflect on use of newly acquired language or language patterns (e.g. through self-assessment checklists and share with a partner)	Revise of rephrase written language based on feedback from teachers, peers and rubrics	Expand, elaborate and correct written language as directed

Table 1.2 Standards for summative writing, language arts, grades 9-12

(WIDA, 2007: 61 in Fulcher, 2010: 285)

	Level 1: Entering	**Level 2: Beginning**	**Level 3: Developing**	**Level 4: Expanding**	**Level 5: Bridging**
Example genre: Critical commentary	Reproduce critical statements on various topics from illustrated models or outlines	Produce critical comments on various topics from illustrated models or outlines	Summarize critical commentaries on issues from illustrated models or outlines	Respond to critical commentaries by offering claims and counter-claims on a range of issues from illustrated models or outlines	Provide critical commentary on a wide range of issues commensurate with proficient peers
Example topic: Literal and figurative language	Produce literal words or phrases from illustrations or cartoons and word/ phrase banks	Express ideas using literal language from illustrations or cartoons and word/ phrase banks	Use examples of literal and figurative language in context from il- lustrations or cartoons and word/phrase banks	Elaborate on examples of literal and figurative language with or without il-lustrations	Compose narratives using literal and figurative language

The problem that often arises in language learning content standards is that there is no specific target for a particular domain, for example, learning the language used by tour guides in a particular context. Thus, students master the language in general, not referring to the context, domain, or specific skills. Also the level of complexity of content standards raises questions about the relationship of content to the required test form. In other words, the performance test should be based on content standards rather than containing everything so that there is a clear relationship between the meaning of the scores the students achieved and the students' claims of success in "mastering" the standard content. If a student's claim of success in mastering standardized content comes from test scores, then the claim for validity is that of a small sample that can be generalized across content. It is one of the validity problems in shortening the content-based approach (Fulcher, 1999). It means that at any appropriateness of learning content, the question will always arise whether the content standard covers all implementation levels in a comprehensive manner. Even though it is comprehensive, each form of the test will still be adapted to the content.

In short, the principle of washback is comprised of the following elements:

a. Positivity influences what and how teachers teach
b. Positivity influence what dan how learners learn
c. Offers learners a chance to adequately prepare
d. Gives learners feedback that enhances their language development
e. Is more formative in nature that summative
f. Provides conditions for peak performance by the learner

Reliability

A reliable test is one that is stable and dependable. If you administer the same test to the same student or paired students on two separate days, the findings should be comparable. The principle of reliability can be summed up as follows (Brown and Abeywickrama, 2018, p. 29):

a. Has consistent conditions across two or more administrations
b. Gives clear directions for scoring/evaluation
c. Has uniform rubrics for scoring/evaluation
d. Lends itself to consistent application of rubrics by the scorer
e. Contains items/tasks that are unambiguous to the test-taker

The topic of test reliability can be best appreciated by taking into account various variables that can lead to their unreliability. We investigate four potential causes of variation: (1) the student, (2) the scoring, (3) test administration, and (4) the test itself.

The Students Reliability Factor

The most common learner-related problem in reliability is exacerbated by temporary unfitness, exhaustion, a "bad day," anxiety, and other physical or psychological causes that cause an observable performance to deviate from one's "real" score. This group also includes considerations such as a taker's test-wiseness and test-taking tactics.

At first glance, student-related unreliability can seem to be an uncontrollable factor for the classroom teacher. We are used to expecting sure students to be stressed or overly nervous to the point of "choking" during a test administration. However, several teachers' experiences say otherwise.

Scoring Reliability Factor

Human error, subjectivity, and racism can all play a role in the scoring process. When two or more scorers provide reliable results on the same test, this is referred to as inter-rater reliability. Failure to attain inter-rater reliability may be attributed to a failure to adhere to scoring standards, inexperience, inattention, or even preconceived prejudices.

Rater-reliability problems are not limited to situations with two or more scorers. Intra-rater reliability is an internal consideration that is popular among classroom teachers. Such dependability can be jeopardized by vague scoring parameters, exhaustion, prejudice against specific "healthy" and "poor" students, or sheer carelessness. When faced with scoring up to 40 essay tests (with no absolute correct or wrong set of answers) in a week, you will notice that the criteria applied to the first few tests will vary from those applied to the last few. You may be "easier" or "harder" on the first few papers, or you may become drained, resulting in an uneven evaluation of all tests. To address intra-rater unreliability, one approach is to read through about half of the tests before assigning final scores or

ratings, then loop back through the whole series of tests to ensure fair judgment. Rater reliability is tough to obtain in writing competence assessments because writing mastery requires various characteristics that are difficult to identify. However, careful design of an analytical scoring instrument will improve both inter- and intra-rater efficiency.

Administration Reliability Factor

Unreliability can also be caused by the circumstances under which the test is performed. We once observed an aural examination being administered. An audio player was used to deliver objects for interpretation, but students seated next to open windows did not hear the sounds correctly due to street noise outside the school. It was a blatant case of unreliability exacerbated by research administration circumstances. Variations in photocopying, the amount of light in various areas of the building, temperature variations, and the state of desks and chairs may all be causes of unreliability.

Test Reliability

Measurement errors may also be caused by the design of the test itself. Multiple-choice tests must be specifically constructed in order to have a range of characteristics that protect against unreliability. E.g., items must be equally complicated, distractors must be well crafted, and items must be evenly spaced in order for the test to be accurate. These reliability types are not addressed in this book since they are rarely appropriately applied to classroom-based assessments and teacher-created assessments.

Test unreliability of classroom-based assessment can be influenced by a variety of causes, including rater bias. It is most common in subjective assessments with open-ended responses (e.g., essay responses) that involve the teacher's discretion to decide correct and incorrect answers. Objective experiments, on the other hand, have predetermined preset answers, which increases test efficiency.

Poorly written test objects, such as vague or have more than one correct answer, can also contribute to unreliability. Furthermore, a test with so many items (beyond what is needed to differentiate among students) will eventually cause test-

takers to become fatigued when they start the later items and answer incorrectly. Timed tests discriminate against students who do not perform well on a timed test. We all know people (and you might be one of them) who "know" the course material well but are negatively influenced by the sight of a clock ticking away. In such cases, it is clear that test characteristics will interact with student-related unreliability, muddying the distinction between test reliability and test administration reliability.

Validity

By far the most complicated criteria of a successful test—and arguably the most important principle—is validity, defined as "the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment" (Gronlund, 1998, p. 226). In somewhat more technical terminology, commonly accepted authority on validity, Samuel Messick (1989), identified validity as "an integrated evaluative judgment of the degree to which objective data and theoretical rationales justify the adequacy and appropriateness of inferences and behaviour based on test scores or other modes of assessment." It can be summed up as follows (Brown and Abeywickrama, 2018, p. 32):

a. Measures exactly what it proposes to measure
b. Does not measure irrelevant or "contaminating" variables
c. Relies as much as possible on empirical evidence (performance)
d. Involves performance that samples the test's criterion (objective)
e. Offers useful, meaningful information about a test-taker's ability
f. Is supported by a theoretical rationale or argument

A valid reading ability test tests reading ability, not 20/20 vision, prior knowledge of a topic, or any other variable of dubious significance. To assess writing skills, ask students to compose as many words as possible in 15 minutes, then count the words for the final score. Such a test might be simple to perform (practical), and the grading would be dependable (reliable). However, it would not

be a credible test of writing abilities unless it took into account comprehensibility, rhetorical discourse components, and concept organization, among other things.

How is the validity of a test determined? There is no final, full test of authenticity, according to Broadfoot (2005), Chapelle & Voss (2013), Kane (2016), McNamara (2006), and Weir (2005), but many types of proof may be used to justify it. Furthermore, as Messick (1989) pointed out, "it is important to note that validity is a matter of degree, not all or none" (p. 33).

In certain situations, it may be necessary to investigate the degree to which a test requires success comparable to that of the course or unit being tested. In such contexts, we might be concerned with how effectively an exam decides whether students have met a predetermined series of targets or achieved a certain level of competence. Another broadly recognized form of proof is a statistical association with other linked yet different tests. Other questions about the validity of a test can centre on the test's consequences, rather than the parameters themselves, or even on the test-sense taker's of validity. In the following pages, we will look at four different forms of proof.

Content-Related Evidence

If a survey explicitly samples the subject matter from which results are to be made, and if the test-taker is required to execute the actions tested, it will assert content-related proof of validity, also known as content-related validity (e.g., Hughes, 2003; Mousavi, 2009). If you can accurately describe the accomplishment you are assessing, you can generally distinguish content-related facts by observation. A tennis competency test that requires anyone to perform a 100-yard dash lacks material legitimacy. When attempting to test a person's ability to speak a second language in a conversational context, challenging the learner to answer multiple-choice questions involving grammatical decisions would not gain material validity. It is a test that allows the learner to talk authentically genuinely. Furthermore, if a course has ten targets but only two are addressed in an exam, material validity fails.

A few examples with highly advanced and complex testing instruments may have dubious content-related proof of validity. It is possible to argue that traditional

language proficiency assessments, with their context-reduced, academically focused language and short spans of discourse, lack material validity because they do not enable the learner to demonstrate the full range of communicative ability (see Bachman, 1990, for a complete discussion). Such critique is based on sound reasoning; however, what such proficiency tests lack in content-related data, they can make up for in other types of evidence, not to mention practicality and reliability.

Another way to perceive material validity is to distinguish between overt and indirect research. Direct assessment requires the test-taker to execute the desired mission. In an indirect test, learners execute a task relevant to the task at hand rather than the task itself. For example, if your goal is to assess learners' oral development of syllable stress and your test assignment is to make them mark (with written accent marks) stressed syllables in a list of written words, you might claim that you implicitly measure their oral production. A direct test of syllable development would necessitate students orally producing target words.

The most practical rule of thumb for achieving content validity in classroom evaluation is to measure results explicitly. Consider a listening/speaking class finishing a unit on greetings and exchanges that involves a lesson on asking for personal information (name, address, hobbies, and others.) with some form-focus on the verb be, personal pronouns, and query creation. The exam for that unit should include all of the above debate and grammatical components and include students in actual listening and speaking results.

Most of these examples show that material is not the only form of evidence that may be used to validate the legitimacy of a test; additionally, classroom teachers lack the time and resources to subject quizzes, midterms, and final exams to the thorough scrutiny of complete construct validation. As a result, teachers must place a high value on content-related data while defending the validity of classroom assessments.

Criterion-Related Evidence

The second type of proof of a test's validity can be seen in what is known as criterion-related evidence, also known as criterion-related validity, or the degree to

which the test's "criterion" has already been met. Remember from Chapter 1 that most classroom-based testing of teacher-designed assessments falls into the category of criterion-referenced assessment. Such assessments are used to assess specific classroom outcomes, and inferred predetermined success standards must be met (80 percent is considered a minimal passing grade).

Criterion-related data is better shown in teacher-created classroom evaluations by comparing evaluation outcomes to results of some other test of the same criterion. For example, in a course unit in which the goal is for students to generate voiced orally and voice-less stops in all practicable phonetic settings, the results of one teacher's unit test could be compared to the results of an independent—possibly a professionally generated test in a textbook—of the same phonemic proficiency. A classroom evaluation intended to measure mastery of a point of grammar in communicative usage will have criterion validity if test results are corroborated by any subsequent observable actions or other communicative in question.

Criterion-related data is often classified into two types: (1) current validity and (2) predictive validity. An evaluation has concurrent validity of the findings are accompanied by other comparable success outside of the measurement. For e.g., true proficiency in a foreign language would substantiate the authenticity of a high score on the final exam of a foreign-language course. In the case of placement assessments, admissions appraisal batteries, and achievement tests designed to ascertain students' readiness to "pass on" to another unit, an evaluation's predictive validity becomes significant. In such situations, the evaluation criterion is not to quantify concurrent ability but to evaluate (and predict) test-probability takers of potential achievement.

Construct-Related Evidence

Build-related validity, also known as construct validity, is the third type of proof that may confirm validity but does not play a significant role for classroom teachers. A construct is any theory, hypothesis, or paradigm that describes observable phenomena in our perception universe. Constructs can or may not be explicitly or empirically measured; their verification often necessitates inferential

evidence. Language constructs include proficiency and communicative ability, while psychological constructs include self-esteem and encouragement. Theoretical structures are used in almost every aspect of language learning and teaching. In the evaluation area, construct validity asks, "Does this test tap into the theoretical construct as defined?" In that their evaluation activities are the building blocks of the object evaluated, tests are, in a sense, operational descriptions of constructs.

A systematic construct validation protocol can seem to be a challenging prospect for most of the assessments you conduct as a classroom teacher. You could be tempted to run a short content search and be pleased with the validity of the test. However, do not be alarmed by the idea of construct validity. Informal construct validation of almost any classroom test is both necessary and possible.

Assume you have been given instructions for how to perform an oral interview. The interview scoring study contains multiple aspects in the final score:

a. Pronunciation
b. Fluency
c. Grammatical accuracy
d. Vocabulary usage
e. Sociolinguistic appropriateness

These five elements are justified by a theoretical construct that says they are essential components of oral proficiency. So, if you were asked to perform an oral proficiency interview that only tested pronunciation and grammar, you would be justified in being sceptical of the test's construct validity. Assume you have developed a basic written vocabulary quiz based on the topic of a recent unit that allows students to describe a series of terms adequately. Your chosen objects may be an appropriate sample of what was discussed in the unit, but if the unit's lexical purpose was the communicative use of vocabulary, then writing meanings fails to fit a construct of communicative language use.

Construct validity is a big concern when it comes to validating large-scale standardized assessments of proficiency. Since such assessments may stick to the maxim of practicability for economic purposes, and since they must explore a small range of expression fields, they will not be able to include all of the substance of a specific area of expertise. Many large-scale standardized exams worldwide, for

example, have not sought to sample oral production until recently, even though oral production is an essential feature of language ability. The omission of oral development, on the other hand, was explained by studies that found strong associations between oral production and the activities sampled on specific measures (listening, reading, detecting grammaticality, and writing). The lack of oral material was explained as an economic requirement due to the critical need to have financially affordable proficiency testing and the high cost of conducting and grading oral output tests. However, with developments in designing rubrics for grading oral production tasks and in automatic speech recognition technologies over the last decade, more general language proficiency assessments have included oral production tasks, owing mainly to technical community demands for authenticity and material validity.

Consequential Validity

In addition to the three currently agreed sources of proof, two other types could be of interest and use in your search to support classroom assessments. Brindley (2001), Fulcher and Davidson (2007), Kane (2010), McNamara (2000), Messick (1989), and Zumbo and Hubley (2016), among others, downplay the possible relevance of appraisal outcomes. Consequential validity includes all of a test's implications, including its consistency in calculating expected parameters, its impact on test-taker's readiness, and the (intended and unintended) social consequences of a test's interpretation and usage.

Bachman and Palmer (2010), Cheng (2008), Choi (2008), Davies (2003), and Taylor (2005) use the word effect to refer to consequential validity, which can be more narrowly defined as the multiple results of evaluation before and after a test administration. Bachman and Palmer (2010, p.30) explain that the effects of test-taking and the use of test scores can be seen at both a macro (the effect on culture and the school system) and a micro level (the effect on individual test-takers).

At the macro stage, Choi (2008) concluded that the widespread usage of standardized exams for reasons such as college entry “deprive[s] students of crucial opportunities to learn and acquire productive language skills,” leading to test users being “increasingly disillusioned with EFL testing” (p. 58).

As high-stakes testing has grown in popularity over the last two decades, one feature of consequential validity has gotten much attention: the impact of test training courses and manuals on results. McNamara (2000) warned against test outcomes that could indicate socioeconomic conditions; for example, opportunities for coaching may influence results because they are "differently available to the students being tested (for example, because only certain families can afford to coach, or because children with more highly trained parents receive support from their parents)."

Another significant outcome of a test at the micro-level, precisely the classroom instructional level, falls into the washback category, which is described and explored in greater detail later in this chapter. Waugh and Gronlund (2012) urge teachers to think about how evaluations affect students' motivation, eventual success in a course, independent learning, research patterns, and schoolwork attitude.

Face Validity

The degree to which "students interpret the appraisal as rational, appropriate, and useful for optimizing learning" (Gronlund, 1998, p. 210), or what has popularly been called—or misnamed—face validity, is an offshoot of consequential validity. "Face validity refers to the degree to which an examination appears to assess the knowledge or skill that it seeks to measure, depending on the individual opinion of the examinees who take it, administrative staff who vote on its application, and other psychometrically unsophisticated observers" (Mousavi, 2009, p. 247).

Despite its intuitive appeal, face validity is a term that cannot be empirically measured or logically justified within the category of validity. It is entirely subjective—how the test-taker, or perhaps the test-giver, intuitively perceives an instrument. As a result, many appraisal experts (see Bachman, 1990, pp. 285-289) regard facial validity as a superficial consideration that is too reliant on the perceiver's whim. Bachman (1990, p. 285) echoes Mosier's (1947, p. 194) decades-old assertion that face validity is a "pernicious fallacy ...[that should be] purged from the technician's vocabulary." in his "post-mortem" on face validity.

Simultaneously, Bachman and other assessment authorities "grudgingly" conclude that test presentation has an impact that neither test-takers nor test creators can disregard. Students might believe, for several purposes, that a test is not measuring what it is supposed to test, which may impact their output and, as a result, cause the previously mentioned student-related unreliability. Students' perceptions of a test's fairness are essential in classroom-based evaluation because they can impact student performance/reliability. Teachers can improve students' perceptions of equal assessments by implementing the following strategies (Brown and Abeywickrama, 2018, p. 38)

a. Formats that are expected and well-constructed with familiar tasks
b. Task that can be accomplished within an allotted time limit
c. items that are clear and uncomplicated
d. directions that are crystal clear
e. tasks that have been rehearsed in their previous course work
f. tasks that relate to their course work (content validity)
g. level of difficulty that presents a reasonable challenge

Finally, the problem of face validity tells us that the learner's psychological status (confidence, fear, etc.) is an essential factor in peak performance. If you "throw a curve" at students on an exam, they will become overwhelmed and anxious. They must have practiced test assignments to be at ease with them before the event. A classroom evaluation is not the time to add new challenges, so you will not know if student complexity is due to the challenge or tested goals.

Assume you administer a dictation exam and a cloze test as a placement test to a group of English as a second language learner. Any students may be frustrated because, on the surface, those assessments do not seem to assess their accurate English skills. They may believe that a multiple-choice grammar test is the best format to use. Some may argue that they did poorly on the cloze and dictation since they were unfamiliar with these formats. While the assessments are superior instruments for selection, students do not believe so.

Validity is a subjective term, but it is critical to a teacher's understanding of what constitutes a successful evaluation. We would do well to remember Messick's (1989, p. 33) warning that validity is not an all-or-nothing proposition and that

different types of validity can need to be added to a test in order to be satisfied with its ultimate usefulness. If you make a point of concentrating on substance and criteria relevance in your language evaluation processes, you will be well on your way to making correct decisions about the learners with whom you deal.

Authenticity

A fourth significant theory of language testing is authenticity, a problematic term to identify, especially in the art and science of assessing and designing tests. Bachman and Palmer (1996) described authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a target language task" (p. 23), and then proposed a strategy for defining specific target language tasks and translating them into relevant test objects.

Authenticity is a term that does not lend itself naturally to scientific description, operationalization, or calculation (Lewkowicz, 2000). After all, who can say whether a job or a language sample is "real-world" or not? Such assessments are often arbitrary, but authenticity is a term that has captivated the attention of various language-testing experts (Bachman & Palmer, 1996; Fulcher & Davidson, 2007). Furthermore, several research forms, according to Chun (2006), fail to replicate real-world tasks.

When you argue for validity in a research exercise, you are essentially saying that this task is likely to be performed in the real world. Many test object styles do not accurately simulate real-world tasks. In their attempt to target a grammatical form or lexical object, they may be contrived or artificial. The arrangement of objects that have no connection to one another lacks credibility. It does not take long to identify reading comprehension passages in proficiency exams that do not correspond to real-world passages.

Authenticity can be presented as follows (Brown and Abeywickrama, 2018, p. 39):

a. Contains language that is as natural as possible
b. Has items that are contextualized rather than isolated
c. Includes meaningful, relevant, interesting topics
d. Provides some thematic organization to items, such as through a story line or episode
e. Offers tasks that replicate real-world tasks

In recent years, there has been a noticeable rise in the authenticity of research assignments. Unconnected, dull, and contrived objects were recognized as a required part of testing two to three decades ago. Everything has changed. It was once thought that large-scale training could not provide productive ability output while remaining under budgetary limits, but several such assessments now include speaking and writing elements. Reading excerpts are drawn from real-world references that test takers are likely to have come across or may come across. Natural language is used in the listening comprehension areas, along with hesitations, white noise, and interruptions. More tests have “episodic” objects, which are sequenced to shape coherent units, chapters, or stories.

Testing and Assessment in Context

Why do tests need to be held? Each test is carried out for a specific purpose because testing is a process to produce fair and correct decisions. In language learning, Carroll (1981: 314) states: 'The purpose of language testing is always to render information to aid in making intelligent decisions about possible courses of action.' However, Caroll's opinion is still too general and needs to be narrowed down further. Davidson and Lynch (2002: 76-78) introduced the term "mandate" to describe where the test objectives are created where the mandate can come from internal or external where the teacher teaches. The internal mandate comes from the teacher or school administration, where the test objectives are tailored to students' and teachers' needs in specific contexts. Usually, the test is used to determine the progress of student achievement, student weaknesses, and group students. Tests are also, sometimes, used to motivate students. For example, when students know they will have an exam on the weekend, they will have an increase in study time compared to a normal day. As the results of research conducted by Latham (1877: 146). 'The efficacy of examinations as a means of calling out the interest of a pupil and directing it into the desired channels was soon recognized by teachers.' Other research conducted by Ruch (1924, p. 3) found that 'Educators seem to be agreed that pupils tend to accomplish more when confronted with the

realization that a day of reckoning is surely at hand.' Generally, tests were able to increase students' motivation in learning to be regarded as fairy tales.

When tests are structured according to local mandate, they must be "ecologically sensitive" and cater to teachers' and students' needs. In other words, the results obtained from this test only apply and give typical locally. Therefore, testing with a local mandate that is ecologically sensitive has different characteristics compared to other tests. For example, a local mandate test will tend to be a formative test where the test acts like a learning process rather than to test the highest achievement. Then, the decisions taken after conducting the test did not have significant consequences for either the teacher or the school but were used to determine what the following learning objective was or determine what lessons the students needed most. The teacher determines the next character, types, and procedures for implementing the assessment and test; even students can convey how they want to be tested. In short, "ecological sensitivity" has a significant impact on the selection and implementation of tests, the decisions taken, and stakeholders' involvement in test design and assessment.

Conversely, the external mandate test refers to why a test is being carried out that comes from outside the context. Usually, the party that conducts the test is the party that is not involved in the learning context and does not directly know the students and teachers. The frequency of motivation to hold external tests is not precise and has a much different function from tests with the internal mandate. The external test aims to determine students' abilities without referring to the student's learning context. So, this test is often called a summative test, which is a test that is carried out at the end of the study period considering that the student has reached the specified standard at that time.

The score obtained through the summative test is considered to provide a 'general' picture of students' abilities outside their learning context. Messick (1989: 14-15) defines generalisability as 'the fundamental question of whether the meaning of a measure is context-specific or whether it generalizes across contexts.' If the formative test results do not have to be general, then in the summative test, the test results are expected to give an idea of the ability of any student who takes the test without being limited to any context. The users of this test score hope that the test

results can represent the students' ability to communicate and adapt to an environment they are not familiar with and are not even present in the test itself. For example, the reading test given is expected to describe students' level of literacy across countries. Another example, a writing test which consists of two questions, is considered capable of representing students' abilities in various writing disciplines.

In the external mandated test, generalization is considered vital because it can show differences in students' abilities between schools, regions, and even countries at a certain level. The external mandate test can be distinguished from an assessment in the classroom regarding its implementation, which has been adjusted to the education and social system values. Students take the test simultaneously at the same place, at the same time, and with seats that are far apart.

The results of this externally mandated test will determine the sustainability of students' education, their long-term prospects, and the work they will do in the future. Thus, the failure of students' inconsistency affects various parties. For example, student failure at the inter-school level will affect reform at the ministerial level by issuing special tests. At the inter-country level, student failure will affect government policies in the field of education. An example of an external mandated test is the Gaokao test conducted in China, where the test results will determine which campus students will study according to the university's passing grade. This test is a test with the most extensive system in the world where the test is carried out in two days, and students will be tested for their proficiency in Chinese, English, mathematics, sciences, and humanities. The exam venue will be closed and guarded by police, and even airplanes will have to take a different route not to cause noise Even though this will cost quite a lot, the Chinese government still carries it out to maintain the concentration of test-takers. Based on the results of research by Haines et al. (2002) and Powers et al. (2002), noise can interfere with concentration and reduce student scores. The difference in student scores due to noise is called the construct irrelevant variance. Another example of irrelevant variance constructs is cheating, using mobile devices (therefore, students are prohibited from bringing mobile devices into the exam room).

No matter how well a test is prepared, there are still unintended consequences. The most common consequence is when teachers and students learn how to answer questions, not master the language being learned. It happens because of the teacher's belief that students can succeed in the test if they learn the technique of answering questions. This effect is part of the washback effect.

Chapter II
Assessing Listening Skill

Competence

The students comprehend how to assess listening skill and can arrange listening skill assessment instrument.

It may seem strange to measure listening independently of speech, given that the two skills are usually practiced together in conversation. However, there are times when no speaking is required, such as when listening to the radio, lectures, or railway station announcements. Often, in terms of testing, there may be cases in which oral testing capacity is deemed impossible for one purpose or another, but a listening test is included for its backwash impact on the growth of oral skills. Listening skills can also be evaluated for diagnostic purposes.

Listening testing is similar to reading testing in several respects because it is a reactive ability. As a result, this chapter will spend less time on topics similar to the testing of the two skills and more time on unique listening issues. The transient existence of spoken language causes particular difficulties in developing listening tests. Listeners cannot usually go back and forth on what is being said in the same manner as a written document might. The one obvious exception, where a tape-recording is made available to the listener, would not constitute a standard listening task for most people.

What the students should be able to do in listening skill should be specify, namely obtain the gist, follow an argument, and recognize the attitude of the speaker. Other specifications are (Hughes, 2003, p. 161-162).:

Informational:

- Obtain factual information;
- Follow instructions (including directions);
- Understand requests for information;
- Understand expressions of need;
- Understand requests for help;
- Understand requests for permission;
- Understand apologies;
- Follow sequence of events (narration);

- Recognise and understand opinions;
- Follow justification of opinions;
- Understand comparisons;
- Recognise and understand suggestions;
- Recognise and understand comments;
- Recognise and understand excuses;
- Recognise and understand expressions of preferences;
- Recognise and understand complaints;
- Recognise and understand speculation.

Interactional:

- Understand greetings and introductions;
- Understand expressions of agreement;
- Understand expressions of disagreement;
- Recognise speaker's purpose;
- Recognise indications of uncertainty;
- Understand requests for clarification;
- Recognise requests for clarification;
- Recognise requests for opinion;
- Recognise indications of understanding;
- Recognise indications of failure to understand;
- Recognise and understand corrections by speaker (of self and others);
- Recognise and understand modifications of statements and comments;
- Recognise speaker's desire that listener indicate understanding;
- Recognise when speaker justifies or supports statements, etc. of other speaker(s);
- Recognise when speaker questions assertions made by other speakers;
- Recognise attempts to persuade others.

Texts

Text should be specified to keep the validity of test and its backwash, such as text type, text form, length, speed of speech, dialect and accent. Text type can be monologue, dialogue, conversation, announcement, talk, instructions, directions, etc. Text forms are such as description, argumentation, narration, exposition, and instruction. Length can be represented in either seconds or minutes. The number of turns taken may be used to specify the length of brief utterances or exchanges. Speed of speech refers to words per minute (wpm) or syllables per second (sps). Dialect can be standard or non-standard varieties, while accents can be regional or non-regional.

The primary thing in arranging exercises to assess students' listening skills is to know the theory of ideas about constructs and how to use them to be carried out in close to the actual context. Historically, there have been three main approaches in measuring students' language skills: the discrete-point, integrative, and

communicative approaches. These three approaches are formed based on the theory of ideas about language and how to understand spoken language and test it.

The theory of practical testing ideas is not always explicit. However, each test is based on a basic theory of how natural constructs are measured. Therefore, some tests were developed based on existing theories, and other tests in some instances were not formed based on existing theories.

The Discrete-Point Approach

In the heyday of the audio-lingual method in language learning, with structuralism as the linguistic paradigm and behaviourism as the psychological paradigm, discrete-point became the language testing approach most commonly used by language teachers. The most famous figure as a consultant for this approach is Lado, who defines language as part of a habit. Lado emphasized that language is a habit that is often used without the need for awareness to use it (Lado, 1961). The discrete-point approach's basic idea is that language can be identified based on language elements, and these elements can be tested. Language testing developers choose the most essential element as a representation of language knowledge because of the many language elements.

According to Lado, listening comprehension is a process of understanding sound language. To test students' listening skills, the technique used is to play or sound the words to students and check whether students understand what they hear, especially the essential parts of the sentences spoken (1961: 208). Furthermore, Lado explained that the parts that need to be considered or tested in the listening test are the phonemes segment, stress, intonation, grammatical structure, and vocabularies. The types of tests that can be used are multiple-choice, pictures, and true/false. Also, what needs to be considered in compiling test listening, the context used should not be too much; it is enough to help students avoid ambiguity and nothing more (1961, 218). Thus, according to Lado, a listening test refers to a test of students' ability to recognize language elements orally.

Discrete-point is a test that is done by selecting the correct answer. The types of tests commonly used in this test are true/false and multiple-choice, where most people think they are the same form of questions. The concept of multiple-choice

in the concrete-point test became the basic idea for the creation of the TOEFL. Although currently, the TOEFL focuses more on comprehension and inference, it still maintains a multiple-choice format. For the listening test itself, the discrete-point test tasks were phonemic discrimination task, paraphrase recognition, and response evaluation.

Phonemic Discrimination Tasks

The phonemic discrimination task is an example of a most often used test in the discrete-point approach to the listening test. This type of test is done by asking students to listen to one isolated word, and students have to determine which word they hear. Usually, the words used are words that differ only by one phoneme or are often called minimal pairs, such as 'ship' and 'sheep,' 'bat' and 'but.' so that students need to know the language able to answer these questions.

For example, students will listen to a recording and choose the words they hear.

Students hear:

They said that they will arrive in Bucureşti next week.

Students read:

*They said that they will **arrive/alive** in Bucureşti next week.*

Students do not get any clue except the explanation that what is being tested is phonetic information. This test is not natural if it refers to the actual conditions when a conversation occurs. Both the speaker and listener will use context in understanding the message conveyed. Nowadays, this test is no longer used, but it can still be used if the student or test taker is a native speaker of the language being tested and has particular problems distinguishing similar sounds (for example, Japanese people find it challenging to distinguish bunya / l / from / r /).

Paraphrase Recognition

Basically, the discrete-point test focuses on a tiny part of a speech, but students or test takers must understand the part being tested and the overall utterance in the listening test.

Example:

Test-takers/ students hear:

Willey runs into a friend on her way to the classroom.

Test-taker read:

a. *Willey exercised with her friend.*
b. *Willey runs to the classroom.*
c. *Willey injured her friend with her car.*
d. *Willey unexpectedly meets her friend.*

The example problem above focuses on the idiom 'run into,' and the other words are just a context for the idiom. Although each choice gives a different meaning between "run" and "run into," to answer the question, students must understand other words.

Response Evaluation

In this type of test, not only one item is tested. Students are required to understand many items on the questions given to be able to answer the questions correctly. Students will hear a question and choose the correct answer to the answer options that have been provided in writing. Example:

Students hear:

How much time did you spend in London?

Students read:

a. *Yes, I did.*
b. *Almost $300.*
c. *About three days.*
d. *Yes, I must.*

The correct answer is (c) 'about three days'. In this test, the focus points being tested are whether the students understand how much time's expression. In option (a) 'yes, I did' be confounding students' understanding of the use of the word 'did' in the question. Option (b) 'almost $ 300' is to confuse students' understanding of using the word 'how much'. So, this question will no longer only test one discrete point but many points.

Another example that looks similar to the form of the question above but is presented differently as follows (Buck, 2001, p. 65)

Students hear:

Male 1: are sales higher this year?

Male 2: a) they're about the same as before.

b) no, they hired someone last year.

c) they're on sale next month.

The questions above are not presented in writing, but orally, both questions and answers. Therefore, it is not the linguistic aspect that is tested in this question. However, the students' ability to understand the meaning of statements uttered by males 1.If students understand the language well, then there are no difficulties for students in answering the questions above because for the two distractors in the answer option is an answer that is not related to the question given. For assessment, discrete-point items are usually assessed by giving a value of one for everyone correct answer, then adding up all the correct answer.

Other techniques in assessing listening skill are:

1. Multiple choice
2. Short answer
3. Gap filling
4. Information transfer
5. Note taking
6. Partial dictation
7. transcription

Chapter III
Assessing Speaking Skill

Competence

The students comprehend how to assess speaking skill and can arrange speaking skill assessment instrument.

The fundamental issue with measuring oral ability is the same as it is with testing writing ability. We want to assign tasks that constitute a representative sample of the population of oral tasks that we expect students to be able to complete. The assignments can evoke behaviour that accurately reflects the students' abilities. Then, the behavioural samples will be scored in a valid and reliable manner.

Representative Tasks

At the specified content of the Cambridge CCSE Test of Oral Interaction, there four levels at which a certificate is awarded (Hughes, 2003, p. 113-116).

Operations

Expressing:	likes, dislikes, preferences, agreement/disagreement, requirement, opinions, comment, attitude, confirmation, complaints, reasons, justifications, comparisons
Directing:	instructing, persuading, advising, prioritising
Describing:	actions, events, objects, people, process
Eliciting:	information, directions, clarification, help
Narration:	sequence of events
Reporting:	description, comment, decisions and choice

Types of text discussion
Addressees 'Interlocuter' (teacher from candidate's school) and one fellow candidate
Topics Unspecified
Dialect, Accent and Style also unspecified
Skills
Informational skills
Candidates should be able to:

- Provide personal information
- Provide non-personal information
- Describe sequence of events (narrate)
- Give instructions
- Make comparisons
- Give explanations

- Present an argument
- Provide required information
- Express need
- Express requirements
- Elicit help
- Seek permission
- Apalogise
- Elaborate an idea
- Express opinions
- Justify opinions
- Complain
- Speculate
- Analyse
- Make excuses
- Paraphrase
- Summarise (what they have said)
- Make suggestions
- Express preferences
- Draw conclusions
- Make comments
- Indicate attitude

Interactional skills
Candidates should be able to:

- Express purpose
- Recognise other speakers' purpose
- Express agreement
- Express disagreement
- Elicit opinions
- Elicit information
- Question assertions made by other speakers
- Modify statements or comments
- Justify or support statements or opinions of other speakers
- Attempt to persuade others
- Repair breakdowns in interaction
- Check that they understand or have been understood correctly
- Establish common ground
- Elicit clarification
- Respond to requests for clarification
- Correct themselves or others
- Indicate understanding (or failure to understand)
- Indicate uncertainty

Skills in managing interactions
Candidates should be able to:

- Initiate interactions
- Change the topic of an interaction
- Share the responsibility for the development of an interaction
- Take turns to other speakers
- Come to a decision
- End an interaction

Types of text

- Presentation (monologue)
- Discussion

- Conversation
- Service encounter
- Interview

Other speakers (addressees)
- May be of equal or higher status
- May be known or unknown

Topics Topics which are familiar and interesting to the candidates

Dialect Standard British English or Standard American English

Accent RP, Standard American

Style Formal and Informal
Vocabulary range Non-technical except as the result of preparation for presentation

Rate of speech Will vary according to task

Choose Appropriate Techniques

Three general techniques can be used in assessing speaking skill: interview, interaction with friends, and responses to audio-recorded or video-recorded stimuli.

Interview

The interview is perhaps the most popular format for assessing oral interaction. However, in its conventional style, it has at least one potentially serious drawback. The tester-candidate partnership is usually such that the candidate talks as if to a superior and cannot take the initiative. Consequently, only one type of speech is elicited, and several roles (such as asking for information) are absent from the candidate's results. However, this issue can be avoided by incorporating a combination of elicitation methods into the interview case. Some techniques in interview:

1. Questions and requests for information

   Yes/No questions can be avoided in general, except maybe at the start of the interview when the student is already warming up. Requests of the following types can evoke the performance of different operations (of the kind specified in the two sets of requirements above):

   'Can you tell what your opinion on ....?'

   'Can you describe why ....?'

2. Pictures

   Ask the students to choose one picture and describe it.

3. Role play

   The students can be asked to assume a role in a particular situation and check how they use the language functions.

4. Interpreting

In this techniques, the students will pretend to be an interpreter. This technique can be conducted by asking two students come to front of the class, one of the students acts a native speaker and does a monologue, while the other acts as interpreter.

Interaction With Friends

One benefit of letting candidates communicate with one another is that it can evoke language suitable for interactions between equals, which the test requirements might require. It may also elicit higher results because applicants may feel more secure than working with a superior, seemingly omniscient interviewer.

However, there is a dilemma. One candidate's success is likely to be influenced by the performance of the others. For example, an assertive and disrespectful candidate can overpower and deny another candidate the opportunity to demonstrate his or her abilities. If candidates have to communicate with one another, the pairs should be carefully paired wherever possible. In general, I would caution against letting more than two candidates interact, as greater numbers raise the likelihood of a hesitant candidate struggling to demonstrate their ability. Some techniques can be used:

1. Discussion

   This technique is done by set the students to be a couple, then ask them to discuss a topic which need a decision.

2. Role play

   For this technique, two students are as to do a specific role and the teacher as an observer of the role play.

Responses to Audio- or Video-Recordings

Uniformity in elicitation procedures can be accomplished by providing all candidates with the same computer-generated or audio/video-recorded stimuli (to which the candidates answer into a microphone). This format, known as 'semi-direct,' could increase dependability. It can also be cost-effective if a language laboratory is available so many applicants can be evaluated simultaneously. The apparent drawback of this format is its inflexibility: there is no means to follow up on candidates' answers. The techniques can be applied in this part are describe situations, remarks in isolation to respond to, and simulated conversation.

Valid and Reliable Scoring

Similar to assessing writing skill, assessing speaking also can be use holistic and analytic rating scales. The criteria need to be assessed are (Hughes, 2003, p.127):

Accuracy	Pronunciation must be clearly intelligible even if some influences from L1 remain. Grammatical/lexical accuracy is high though grammatical errors which do not impede communication are acceptable.
Appropriacy	The use of language must ne generally appropriate to function and to context. The intention of the speaker must be clear and unambiguous.
Range	A wide range of language must be available to the candidate. Any specific items which cause difficulties can be smoothly substituted or avoided.
Flexibility	There must be consistent evidence of the ability to 'turn-take' in a conversation and to adapt to new topics or changes of direction.
Size	Must be capable of making lengthy and complex contributions where appropriate. Should be able to expand and develop ideas with minimal help from the interlocutor.

It has been suggested that holistic and analytic measures can be used to verify each other. The American FSI (Foreign Service Institute) interview protocol, for example, allows the two testers involved in each interview to both allocate students to a level holistically and score them on a five-point scale on each of the following: accent, grammar, vocabulary, fluency, and comprehension. All scores are then weighted and added together. The resulting score is then entered into a table that translates the scores into the holistically defined levels. The converted score should result in the same amount as the candidate's initial assignment. If this

is not the case, the testers would have to rethink whether their initial assignments were correct. The weightings and conversion tables are focused on studies that found a substantial consensus between holistic and analytic ratings. I will testify to the effectiveness of this method because I used it myself while checking bank employees. I've included the ranking scales and weighting table for the reader's convenience. However, keep in mind that they were designed for a specific reason and could not be assumed to perform well in a radically different case without alteration. It's also worth noting that using a native-speaker norm to assess success has recently come under fire in several language testing circles.

The five-point scale can be described as follows (Adams and Frith in Hughes, 2003, p.131-133)

Proficiency Descriptions

Accent

1. Pronunciation frequently unintelligible.
2. Frequent gross errors and very heavy accent make understanding difficult, require frequent repetition.
3. “Foreign accent” requires concentrated listening, and mispronunciations lead to occasional misunderstanding and apparent errors in grammar or vocabulary.
4. Marked “foreign accent” and occasional mispronunciations which do not interfere with understanding.
5. No conspicuous mispronunciations, but would not be taken for a native speaker.
6. Native pronunciation, with no trace of “foreign accent.”

Grammar

1. Grammar almost entirely inaccurate except in stock phrases.
2. Constant errors showing control of vert few major patterns and frequently preventing communication.
3. Frequent errors showing some major patterns uncontrolled and causing occasional irritation and misunderstanding.
4. Occasional errors showing imperfect control of some patterns but no weakness that causes misunderstanding.
5. Few errors, with no patterns of failure.
6. No more than two errors during the interview.

Vocabulary

1. Vocabulary inadequate for even the simplest conversation.
2. Vocabulary limited to basic personal and survival areas (time, food, transportation, family, etc.).
3. Choice of words sometimes inaccurate, limitations of vocabulary prevent discussion of some common professional and social topics.
4. Professional vocabulary adequate to discuss special interests; general vocabulary permits discussion of any non-technical subject with some circumlocutions.
5. Professional vocabulary broad and precise; general vocabulary adequate to cope with complex practical problems and varied social situations.
6. Vocabulary apparently as accurate and extensive as that of an educated native speaker.

Fluency

1. Speech is so halting and fragmentary that conversation is virtually impossible.
2. Speech is very slow and uneven except for short or routine sentences.
3. Speech is frequently hesitant and jerky; sentence may be left uncompleted.
4. Speech is occasionally hesitant, with some unevenness caused by rephrasing and groping for words.
5. Speech is effortless and smooth, but perceptively non-native in speed and evenness.
6. Speech on all professional and general topics as effortless and smooth as a native speaker.

Comprehension

1. Understands too little for the simplest type of conversation.
2. Understands only slow, very simple speech on common social and touristic topics; requires constant repetition and rephrasing.
3. Understands careful, somewhat simplified speech when engaged in a dialogue, but may require considerable repetition and rephrasing.
4. Understands quite well normal educated speech when engaged in a dialogue, but requires occasional repetition or rephrasing.
5. Understands everything in normal educated conversation except for very colloquial or low-frequency items, or exceptionally rapid or slurred speech.
6. Understands everything in both formal and colloquial speech to be expected of an educated native speaker.

WEIGHTING TABLE

	1	2	3	4	5	6	(A)
Accent	0	1	2	2	3	4	_____
Grammar	6	12	18	24	30	36	_____
Vocabulary	4	8	12	16	20	24	_____
Fluency	2	4	6	8	10	12	_____
Comprehension	4	8	12	15	19	23	_____
						Total	_____

Note the relative weightings for the various components.

The total of weighted scores is then looked up un the following table, which converts it into a rating on a scale 0-4+.

CONVERSION TABLE

Score	Rating	Score	Rating	Score	Rating
16-25	0+	43-52	2	73-82	3+
26-32	1	53-62	2+	83-92	4
33-42	1+	63-72	3	93-99	4+

As analytic scales of this kind are used instead of holistic scales, the question of what pattern of scores (for a particular candidate) should be considered acceptable emerges (as with writing testing). It is essentially the same dilemma as persons failing to match holistic definitions. Once again, deciding what deficiencies to meet the expected level on specific criteria is appropriate based on experience.

Chapter IV
Assessing Reading Skill

Competence


The students comprehend how to assess reading skill and can arrange reading skill assessment instrument.

The testing of reading ability seems deceptively easy if compare to testing oral ability. You take a passage and ask few questions about it, and voila! Although you can create a reading test easily, it may not be a proper test and may not measure what you want it to measure.

The fundamental issue is that practicing receptive skills does not always, or generally, result in overt behaviour. When people write and speak, there is always little to see or hear when they read and listen. The challenge for the language tester is to devise activities that will require the applicant to practice reading (or listening) skills and result in behaviour that demonstrates the successful application of those skills. This issue is divided into two sections. First, there is confusion over the abilities used in reading and that language tests are interested in measuring for different reasons; these have been hypothesized, but some have been unequivocally proven to occur. Second, even though we trust a specific ability, determining whether an object has succeeded in calculating it is challenging.

The proper solution to this issue is not to use the simple approach to reading testing described in the first paragraph as we wait for proof that the abilities we believe exist. We think these abilities exist because, as readers, we are conscious of at least some of them. We are aware that, depending on our reading goal and the type of text, we can read in various ways. On one occasion, we could read slowly and deliberately, word by word, to pursue a philosophical statement. Another time, we could jump from page to page, pausing just a few seconds on each to get the gist of something. Another time, we could skim down a column of text, looking for a specific piece of material. Undoubtedly, experienced readers are experts at adapting their reading style to the intent and content. As a result, I see no reason why these various types of reading should not be included in a test's requirements.

When we focus on our reading, we become aware of other abilities we possess. Few of us know the meaning of any word we come across, but we will frequently infer the meaning of a word from its context. Similarly, as we listen, we are constantly inferring about objects, stuff, and activities. If we read that someone spent an evening in a bar and then staggers home, we can conclude that he staggers because of what he drank (I realize that he may have been an innocent footballer who was hit on the ankle in a game and then went to the pub to drink lemonade, but I did not say that any of our inferences were correct).

It would be counterproductive to continue providing samples of our known reading skills. The argument is that we are aware of their existence. The fact that not all of them have been validated by study does not justify excluding them from our requirements, and therefore from our studies. The question is whether including them in our test would be beneficial. The response may be assumed to depend, at least in part, on the intent of the exam. It is a screening evaluation that seeks to define in depth the strengths and shortcomings in learners' reading skills. If it is an achievement test, and the improvement of these abilities is a course goal, the response must be yes once more. If it is a placement test, where a rough indicator of reading ability is necessary, or a mastery test, where an 'overall' measure of reading ability is sufficient, the response may be no. However, the response 'no' raises another concern. What would we test if we do not put these abilities to the test? Any one of the questions listed in the first paragraph must be measuring something. If our things are going to test something, indeed based on validity, in a test of overall abilities, we can test a selection of all the skills involved in reading that are important to our intent. It is what I would suggest.

The weasel words in the previous sentence are, of course, relevant to our intent. There may be a justification for using objects that measure the ability to differentiate between letters in a screening test for beginners (e.g., between b and d). However, these are usually measured indirectly by higher-level objects. The same can be said for syntax and vocabulary. They are all checked implicitly in a reading exam, but grammar and vocabulary elements are only tested in grammar acid vocabulary examinations, in my opinion.

To be compliant with our general specification framework, we will refer to the skills that readers perform when reading a text as 'operations.' Following are checklists (not intended to be exhaustive) that the author believes the reader of this book will find helpful. Take note of the distinction between expeditious (quick and efficient) reading and slow and cautious reading based on variations in meaning. In the past, there has been a trend in studies to give expeditious reading less weight than it merits. As a result of this, many pupils have not been taught to learn efficiently and effectively. It is a significant drawback when they study abroad and are forced to learn thoroughly in very short amounts of time. Another case of hazardous backwash!

The expeditious and careful reading operations can be described as follows (Hughes, 2003, p.138-139)

Expeditious reading operations

Skimming
The candidate can:

- Obtain main ideas and discourse topic quickly and efficiently;
- Establish quickly the structure of text;
- Decide the relevance of a text (or part of a text) to their needs.

Search reading
The candidate can quickly find information on a predetermined topic.

Scanning
The candidate can quickly find:

- Specific words or phrases;
- Figures, percentages;
- Specific items in an index;
- Specific names in a bibliography or a set of references.

Careful reading operations

- Identify pronominal reference;
- Identify discourse markers;
- Interpret complex sentences;
- Interpret topic sentences;
- Outline logical organization of a text;
- Outline the development of an argument;
- Distinguish general statements from examples;
- Identify explicitly stated main ideas;
- Identify implicitly stated main ideas;
- Recognize writer’s intention;
- Recognize the attitudes and emotions of the writer;
- Identify addressee or audience for a text;
- Identify what kind of text is involved (e.g. editorial, diary, etc.);
- Distinguish fact from opinion;
- Distinguish hypothesis from fact;
- Distinguish fact from rumour or hearsay.

Make inferences:

- Infer the meaning of an unknown word from context.
- Make propositional informational inferences, answering questions beginning with *who*, *when*, *what*.
- Make propositional explanatory inferences concerned with motivation, cause, consequence and enablement, answering questions beginning with *why, how*).
- Make pragmatic inferences.

Reading process can be presented as follows:

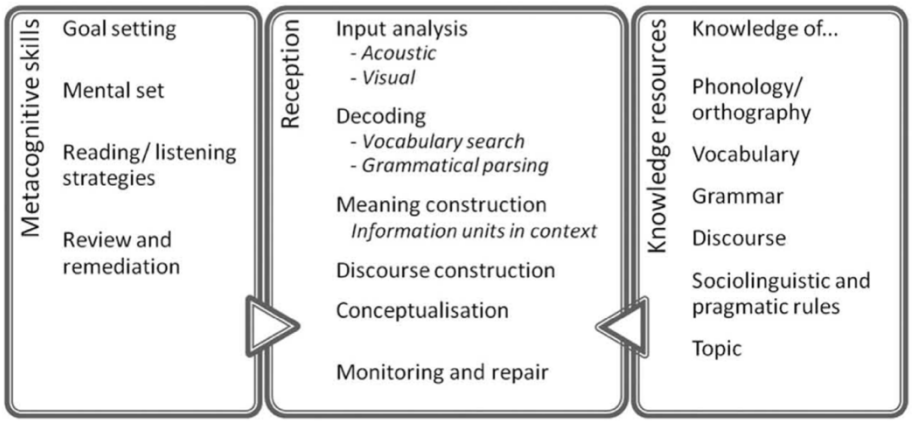


Figure 4.1 An outline model of receptive language process by Weir (2005) and Field (2008) (cited in Green, 2014, p. 101)

When developing a reading ability assessment, developers must consider the various types of reading that the assesses would need to do in the target domain. What methods, skills, and sources of knowledge will be used? Figure 4.1 can be a reference to decide what kind of task should be used.

The left column (metacognitive skills) describes how readers handle the reading process. A student, for example, can determine what kinds of information he wants to get from a text and set himself the goal of extracting this information. He decides how to learn in order to get the knowledge he needs as quickly as possible. He considers what hhe already knows about the subject and will formulate questions that he wants the text to answer: he creates a mental collection to communicate with the text. He selects a promising source, skimming through a textbook on a subject he has researched to see if it provides knowledge he does not already know. He assesses his understanding and learns that he does not

comprehend what the author says in the chapter. So he takes some reparative approaches. He returns to the beginning of the Chapter and reads it slowly to strengthen his understanding, perhaps with the assistance of a dictionary or encyclopaedia: this is what Enright et al. (2000) refer to *as reading to learn*.

Texts

Texts that candidates are supposed to be able to handle can be classified according to a variety of criteria, including type, form, vocabulary range, length, topic, style, graphic features, readability or difficulty, intended readership, and grammatical structure.

Type: Textbooks, handouts, documents (in newspapers, journals, or magazines), poems/verse, flyers, letters, encyclopaedia entries, forms, diaries, charts, dictionary entries, schedules, posters, postcards, timetables, novels (extracts), short stories, surveys, guides, computer aid programs, notices, and signs.

Form: description, exposition, argumentation, instruction, and narration.

Graphic features: charts, tables, diagrams, illustrations, and cartoons.

Topic: non-specialist, non-technical.

Style: formal, informal.

Intended readership: specific or general.

Length: length refers to the number of words which is according to the level of the students, and whether it is expeditious or careful reading.

Readability: this part is measure the difficulty of the text. Using it is depended on the institution.

Range of vocabulary: it refers to list of words.

Range of grammar: it refers to list of structures or grammar which is found in the course book.

Techniques

It is crucial that the methods used interfere with reading as least as possible and do not put a substantially challenging job on top of reading. It is one reason why asking candidates to write responses, particularly in the text's language, should be avoided. They may be able to read perfectly well, but writing disabilities may

preclude them from showing this. Among the possible solutions to this dilemma are:

1. Multiple choice
2. Short answer
3. Gap filling
4. Information transfer
5. Cloze test

Chapter V
Assessing Writing Skill

Competence


The students comprehend how to assess writing skill and can arrange writing skill assessment instrument.

Given the decision to specifically assess writing ability, we can state the testing issue for writing in general terms. It is divided into three sections:

1. We must assign writing assignments that are adequately reflective of the population of tasks that we should require students to complete.
2. The assignment should evoke valid writing samples (— for example, samples that accurately reflect the students' abilities).
3. the writing samples must and will be scored correctly.

Representative Tasks

To determine if the tasks we assign indicate the tasks we want students to be able to complete, we must first define the tasks that they should be able to complete. The test requirements should include this information. The task framework specification includes the following elements: operation, text type, addressees, text length, topics, dialect, and design.

For example, writing task level 1 in Cambridge Certificates in Communicative Skill in English (CCSE) handbook has the complete set of specification in the table as follows:

Table 5.1 Specification set of Writing Test in CCSE Handbook

Operations	*Expressing:* thanks, requirements, opinions, comment, attitude, confirmation, apology, want/need, information, complaint reasons, justifications *Directing:* ordering, instructing, persuading, advising, warning *Describing* actions, events, objects, people, processes *Eliciting* information, directions, service, clarification, help, permission

	Narration sequence of events *Reporting* description, comment, decisions
Types of text	Form, letter (personal, business), message, fax, note, notice, postcard recipe, report, set of instructions.
Addressees of texts	Unspecified, although 'the target audience for each piece of writing made clear to the candidate'
Dialect and length	Unspecified

The CCSE Certificate in Writing specifications (as they exist in the Handbook) presumably account for a large proportion of the writing activities that students in general language courses with communicative purposes are required to accomplish. As a result, they can be helpful to readers of this book who are in charge of testing and writing on those courses. Institutional testers should classify the elements that refer to their specific case under each heading. There will be points where more clarity is required, and somewhere extra elements are required. There is no excuse to feel constrained by this structure or its content, but these requirements can serve as a good starting point for various testing purposes.

In terms of content validity, the optimal exam will allow applicants to complete all applicable possible writing assignments. Our best measure of a candidate's ability will be the overall score earned on the test (the sum of the scores on each of the various tasks). We would not consider any of a candidate's grades equal, even though they were ideally scored on the same scale if this were ever possible. People can excel in certain things while failing at others. If we cannot include any task (which is usually the case) and thus choose only the task or tasks that a candidate is excellent (or bad) at, the result is likely to be somewhat different. It is why we make an effort to choose a representative set of activities. Moreover, the more tasks we assign (within reason), the more reflective of a candidate's abilities (and therefore the more valid) the entirety of the samples (of the candidate's ability) we receive. It should also be noted that if a test contains a diverse and representative sample of parameters, the test is more likely to have a beneficial backwash effect. For instance the CCSE level 1 version for May/June 2000 (Hughes, 2003, p. 86-88).

This test of writing is about working in a Summer Camp for Children in America. Look carefully at the information on this page. Then turn to the next page.

AMERICAN SUMMER CAMPS FOR CHILDREN

VOLUNTEERS WANTED FOR AUGUST 2000

We are looking for people to work as Helpers in our Summer Camp in Florida. You will be responsible for organising games and activities for groups of children.

There is no salary, but travel and living expenses will be paid.

Write to us for more information and an application form:

American Summer Camps for Children
450 Sunny Dale Avenue
Florida 70401
USA

Fax: 1-836-704-9732

Task 1

You saw the advertisement for Helpers. You write a letter to American Summer Camps at the address in the advertisement.

In your letter:

- find out about
  - the start and finish dates
  - the hours of work
  - the type of accommodation
- ask for an application form

Task 2

American Summer Camps for Children sent you an application form. Fill in the APPLICATION FORM below

AMERICAN SUMMER CAMPS FOR CHILDREN

SECTION A: Please use CAPITALS for this section.

FAMILY NAME: (Mr/Mrs/Ms)_____

FIRST NAME(S): _____

AGE: _____ DATE OF BIRTH: _____

NATIONALITY: _____

SECTION B

TICK (✓) the age group of children you would most like to work with.
(NB: Choose only ONE group)

9–10 ☐ 11–13 ☐ 14–16 ☐

Why did you choose this particular age group?

_____

_____

SECTION C

In about 30 words, say why you think you would be especially good at organising games and activities for children.

_____

_____

_____

Signature: _____ Date: _____

Task 3

You are now working in the American Summer Camps for Children in Florida. You write a postcard to an English-Speaking friend.

On your postcard tell your friend:

- where you are
- why you are there
- two things you like about the Summer Camp write your POSTCARD here

POSTCARD

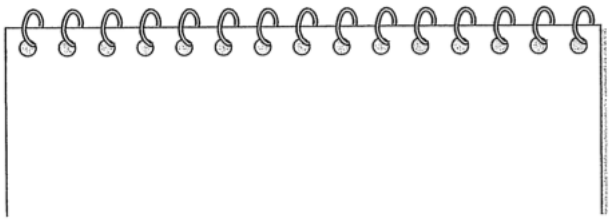
Ms Jane Collins
23 High Street
Everytown, Cambs.
England

Task 4

You have arranged to go out tonight with Gerry and Carrie, two other Helpers at the Summer Camp in Florida. You have to change your plans suddenly, and cannot meet them. You leave them a note.
In your note:

- apologise and explain why you cannot meet them
- suggest a different day to go out.

Write you NOTE here



This demonstrates that the examiners made a concerted effort to construct a diverse sample of assignments. What is also evident is that with so many possible activities and so few objects, the test's material relevance is eventually called into question. A single iteration of the exam cannot provide comprehensive coverage of the number of practicable activities. There is no simple solution to this dilemma. The only analysis will inform us whether a candidate's success on a tiny group of chosen tasks will result in somewhat close ratings to those awarded for performance on another small, non-overlapping set.

It is not nearly as difficult to choose representative writing assignments at an English medium university. Content validity is less of an issue than for the much broader CCSE exam. Since there is no substantial variability under the heading of 'operations,' a test requiring the pupil to write four responses could span the whole set of assignments, assuming that variations in the subject did not apply. In reality, the writing part of each version of the test contained two writing tasks, making each version of the test contain 50% of all tasks. Topics were selected that were supposed to be familiar to all students, and facts or reasons were given.

Valid and Reliable Scoring

1. Set tasks which can be reliably scored

   Several of the recommendations made to achieve a representative score would also help with accurate scoring.

2. Set as many tasks as possible

   The more points there are for each candidate; the more accurate the final score can be.

3. Restrict candidates

   The larger the limitations placed on the contestants, the more strictly equivalent their results would be.

4. Give no choice of tasks

   Making candidates complete all tasks often facilitates comparisons between candidates.

5. Ensure long enough samples

   Elicited writing samples must be long enough for reliable judgments to be rendered. This is especially critical when seeking diagnostic knowledge. For example, to collect accurate statistics on students' organizational capacity in writing, the pieces must be long enough for the organization to emerge. Given a set time limit for the research, it is almost unavoidable friction between the need for the duration and the need for as many samples as possible.

6. Create appropriate scales for scoring

   The scales used in rating performance are expected to be included in the requirements under the heading 'criteria' performance standards. There are two basic scoring approaches: holistic and analytic.

7. Holistic scoring

   Holistic scoring (also known as 'impressionistic' scoring) entails assigning a single score to a piece of writing based on an overall impression. This type of scoring has the advantage of being extremely quick. Experienced scorers will judge a one-page piece of writing in a matter of minutes, even less. This means that each work can be scored several times, which is lucky because it is also required! Harris (1968) cites studies in which the

reliability coefficient was just 0.25 when each pupil composed one 20-minute composition — scored only once. Holistic scoring, in which four independent qualified scorers score each student's work, will result in high scorer reliability if well-conceived and well organized. There is nothing magical about the number four; it is merely that testing has repeatedly proven that when writing is scored four times, the scorer reliability is acceptable.

The TOEFL assessment component for writing skill can be used in assessing students' writing task (Hughes, 2003, 96-97).

TEST OF WRITTEN ENGLISH
Scoring guide

Readers will assign scores based on the following guide. Though examinees are asked to write on a specific topic, parts of the topic may be treated by implication. Readers should focus on what the examinee does well.

[6] **Demonstrates clear competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors.**
A paper in this category
- Effectively organized and well developed
- Is well organized and well developed
- Uses clearly appropriate details to support a thesis or illustrate ideas
- Displays consistent facility in the use of language
- Demonstrates syntactic variety and appropriate word choice

[5] **Demonstrates competence in writing on both the rhetorical and syntactic levels, though it will probably have occasional errors.**
A paper in this category
- May address some parts of the task more effectively than others
- Is generally well organized and developed
- Uses details to support a thesis or illustrate an idea
- Displays facility in the use of language
- Demonstrates some syntactic variety and range of vocabulary

[4] **Demonstrates minimal competence in writing on both the rhetorical and syntactic levels.**
A paper in this category
- Addresses the writing topic adequately but may slight parts of the task
- Is adequately organized and developed
- Uses some details to support a thesis or illustrate an idea
- Demonstrates adequate but possibly inconsistent facility with syntax and usage
- May contain some errors that occasionally obscure meaning

[3] **Demonstrates some developing competence in writing, but it remains flawed on either the rhetorical or syntactic level, or both.**
A paper in this category may reveal one or more of the following weaknesses:

- Inadequate organization or development
- Inappropriate or insufficient details to support or illustrate generalizations
- A noticeably inappropriate choice of words or word forms
- An accumulation of errors in sentence structure and/or usage

[2] **Suggests incompetence in writing.**
A paper in this category is seriously flawed by one or more of the following weaknesses:

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

[1] **Demonstrates incompetence in writing.**
A paper in this category

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

However, the headings are too general due to many institutions using this scoring rubrics. The good point is that this rubric provides six levels of linguistic feature indication which is useful in scoring and for the test score users

8. Analytic scoring

   Analytic scoring methods include a different score for any of a variety of aspects of an assignment. John Anderson developed the following scale based on an oral capacity scale found in Harris (1968) cited in Hughes, (2003, 101-102).

Grammar

6. Few (if any) noticeable errors of grammar or word order.
5. Some errors of grammar or word order which do not. However, interfere with comprehension.
4. Error of grammar or word order fairly frequent: occasional re-reading necessary for full comprehension.
3. Errors of grammar or word order frequent; efforts of interpretation sometimes required on reader's part.
2. Errors of grammar or word order very frequent; reader often has to rely on own interpretation.
1. Errors of grammar or word order so severe as to make comprehension virtually impossible.

Vocabulary

6. Use of vocabulary and idiom rarely (if at all) distinguishable from that of educated native writer.
5. Occasionally uses inappropriate terms or relies in circumlocutions; expression of ideas hardly impaired.
4. Uses wrong or inappropriate words fairly frequently; expression of ideas may be limited because of inadequate vocabulary.
3. Limited vocabulary and frequent errors clearly hinder expression of ideas.
2. Vocabulary so limited and so frequently misused that reader must often rely on own interpretation.
1. Vocabulary limitations so extreme as to make comprehension virtually impossible.

Mechanics

6. Few (if any) noticeable lapses in punctuation or spelling.
5. Occasional lapses in punctuation or spelling which do not. However, interfere with comprehension.
4. Errors in punctuation or spelling fairly frequent; occasional re-reding necessary for full comprehension.
3. Frequent errors in spelling or punctuation; lead sometimes to obscurity.
2. Errors in spelling or punctuation so frequent that reader must often rely on own interpretation.
1. Errors in spelling or punctuation so severe as to make comprehension virtually impossible.

Fluency (style and ease of communication)

6. Choice of structures and vocabulary consistently appropriate; like that of educated native writer.
5. Occasional lack of consistency in choice of structures and vocabulary which does not, however, impair overall ease of communication.
4. 'Patchy', with some structures or vocabulary items noticeably inappropriate to general style.
3. Structures or vocabulary items sometimes not only inappropriate but also misused; little sense of ease communication.
2. Communication often impaired by completely inappropriate or misused structures or vocabulary items.
1. A 'hotch-potch' of half-learned misused structures and vocabulary items rendering communication almost impossible.

Form (organization)

6. Highly organized; clear progression of ideas well linked; like educated native writer.
5. Material well organized; links could occasionally be clearer but communication not impaired.
4. Some lack of organization; re-reading required for clarification of ideas.
3. Little or no attempt at connectivity, though reader can deduce some organization.
2. Individual ideas may be clear, but very difficult to deduce connection between them.
1. Lack of organization so severe that communication is seriously impaired.

SCORE:

Gramm:_____+ Voc_____+ Fluency_____+ Form =_____

(TOTAL)

Analytic scoring has a range of benefits. First, it addresses the issue of unequal subskill growth in individuals. Second, scorers are forced to accept facets of results that they would otherwise overlook. Third, the fact that the scorer is required to have several scores seems to make the scoring more accurate. However, it is unlikely that scorers will judge each factor independently of the others (a phenomenon known as the 'halo effect,' possessing (in this case) five 'shots' at measuring the student's results could contribute to more excellent reliability.

Each of the components is assigned an equal weight in Anderson's scheme. Other schemes (such as those of Jacobs et al. (1981), below) represent the relative significance of the various factors as viewed by the tester (with or without statistical support) in weightings assigned to the various components. Grammatical accuracy, for example, could be given more weight than spelling accuracy. The cumulative score of a nominee is the sum of the weighted scores.

The biggest drawback of the analytic approach is the amount of time required. Scoring can take longer, except with practice, than the holistic approach. Depending on the situation, the analytic approach or the holistic method would be the more cost-effective way of achieving the desired degree of scorer reliability.

A second disadvantage is that focusing on the various aspects can divert attention away from the overall impact of the writing. Since the number of its parts is often more significant than the sum of its parts, a composite score may be very accurate but not true. Indeed, the aspects that are scored separately (the 'parts'), probably based on the theory of linguistic output that most appeals to the author of any given analytic context, do not reflect the whole, 'right' collection of such aspects. To avoid this, scorers are often expected to include an extra, impressionistic score on each composition, with significant differences between this and the analytic total being investigated.

Chapter VI
Testing for Young Learner

Competence

The students comprehend how to assess young learner English skill and can arrange English skill assessment instrument.

What we know about language learning has a wide range of consequences for evaluating foreign and second languages. Teachers and assessors must understand the social and cognitive mechanisms at work as children adapt to the evaluation criteria set before them to assess language learning. Effective language assessment develops children's abilities to use language in its broadest sense; assessment can also promote and monitor children's ability to enter new discourses relevant to the language they are studying, if they are primarily social communication discourses for present and future encounters with native speakers, and/or discourses of linguistic literacy. The effective appraisal occurs in an environment in which children's first language and first language cultures are recognized and built.

Children's better capacity to comprehend and use formulae in the early stages of schooling necessitates selecting specific forms of activities in those early stages, where children will perform using their established formulae and vocabulary. Such a task will be familiar, regular, and most likely repetitive. One example is early morning whole-class rituals in which children check the day, date, and temperature. Simple games are another choice. More rule-based assessment exercises can be used when children are more fluent in the language and competent enough to do explicit language-focused evaluation work. Children can tolerate linguistic usage in which they are expected to go above the predictable and routine; they can be asked to tell someone what they did during the weekend, explain a shared experience, or write a story on a selected animal as they advance. When new guidelines emerge, testing must be tailored to the context of language rules, terminology, and meaning that children can handle; however, teachers and assessors must continue to track the ongoing production of formulae as they continue to play an essential role in active language usage.

Assessment and feedback must elicit optimistic feelings in children about language learning, themselves, and others. Since children carry various perspectives and motivations to their learning, individualized needs evaluation and related targeted input while instruction helps to improve achievement and, as a result, encouragement. Teaching self-assessment techniques and encouraging self- and peer-assessment in the classroom allows children to participate in the continuous deep learning needed for effective international and second language learning and develop their language learning strategies.

Understanding the role children's first language plays in their foreign and second language learning is needed for practical evaluation. It is mainly accomplished by the teacher's and assessor's recognition of the first language throughout the evaluation phase (e.g., their appreciation of the usage of the first language in some cases to help children understand what is required of the assessment procedure), and their acceptance of children's use of the first language when their second language "falls down". Such acknowledgments in classroom and external evaluations can be made with proper preparation and without compromising the evaluation results' integrity.

Furthermore, decisions regarding children's foreign or second language learning that are made without respect for the nature of their success in their first language are likely to be ill-informed, and the subsequent behaviour, such as placement and interference, maybe insensitive and even detrimental. Assessment exercises aimed at determining young learners' abilities to use the language must demonstrate the language use experiences that children partake in within a successful language learning atmosphere. Since a large portion of evaluation in elementary schools occurs in the classroom and during the day-to-day business of learning that constitutes the program, it assumes that a large portion of assessment occurs through activities that children are actively involved in the classroom. Children demonstrate their language skills by doing tasks they are familiar with and tasks that are likely to pique their curiosity and desire to use the language. Also, in more formal assessment activities involving language use, such as a one-on-one interview with the teacher or a brief image summary, it is essential to format the assessment task to illustrate the types of learning tasks that maximize their

engagement and involvement in language use. The types of tasks that do this are those that represent the most successful ways they learn the language.

It is self-evident that the way children learn better will be mirrored in the way they are tested, and understanding how young learners learn a language is also essential for those interested in language testing of young learners.

Teachers and assessors need knowledge of language learning at all stages of the evaluation process, including when they choose or create assessment activities, assess the quality of children's results, and provide input and reports on that performance. At all of these points, the children's growth and long-term achievement can be influenced positively or negatively.

The Effect of Curriculum in Language Assessment

Language learning in schools is often rooted in a program developed by the state, district, educator, or classroom teacher, and this has a significant effect on the essence of language learning. A fixed textbook may also be used to develop the curriculum. The way a curriculum or textbook is written out and sequenced represents the curriculum writer's, teacher's, or textbook developer's understanding of language learning. Assessment should represent the curriculum's aims and priorities, but the curriculum's embedded understandings can also inform language learning. If the existing curriculum stresses the study of grammar and vocabulary in isolation, teachers, and assessors may find it difficult, if not impossible, to assess children's language usage abilities. On the other hand, when the program is intended to encourage language usage, learners may have chances to use language meaningfully, so measuring language skill by language use exercises is the most effective and agreed method of assessing language learning.

Some curricula precisely specify goals or results in information and skills that are essential to language learning. The aims extend beyond the immediate comprehension, comprehension, and skills of language learning to include relevant, connected fields such as the development of understanding of how children approach life (intercultural interpretation), the development of language awareness, and the development of knowing-how-to-learn skills.

These objectives define what children need, according to this framework, in order to properly learn the language and learn beyond language, and they define the criteria for an evaluation in the language learning program. School authorities also define goals and learning objectives in standards documents. Such curricula for young learners can be broader in nature (though this is uncommon), describing only the structures and terminology to be taught. Thus, the framework within which teachers and assessors serve and the textbooks that follow that curriculum determine the scope and essence of language learning and, as a result, have a significant impact on what is taught and evaluated.

Language Knowledge

We are now shifting our focus from language learning processes to the essence of language ability. How can we describe language skills in order to 'capture' them in assessment? How do we analyse a child's language use in an appraisal challenge to determine if it is suitable for the case, whether it can accomplish what it sets out to do, and its strengths and shortcomings? This section's framework is complicated; nevertheless, children's language ability is no less complex than adults' language ability. I would say that teachers and assessors of young learners must have a thorough understanding of the essence of language ability.

Language learners need organizational skills to arrange and generate their own spoken and written texts and comprehend the texts of others. To arrange individual utterances or sentences, they include grammatical knowledge, which, according to Bachman and Palmer (1996), includes vocabulary, phonology, graphology, and syntax. To form texts by merging utterances or sentences, they need textual knowledge comprised of knowledge of cohesion and knowledge of the rhetorical or conversational organization. Cohesion knowledge is needed for creating or comprehending the relationship between sentences in written texts or utterances in conversations. Making or comprehending organizational growth in written texts or conversations requires knowledge of the rhetorical or conversational organization. For example, we know that in their ideal form, English written narratives have a beginning, a climax, and a resolution.

Assessing Language through Task

Parents and teachers both expect their children to pass the requisite assessments in certain teaching cases where external assessment is used. Most parents and teachers will expect exams to improve their children's language skills, but this is not always the case. Since external language assessments may significantly impact the essence of language teaching and learning in the classroom, aligning standardized tests with language usage is highly desirable. Language instructors may help children learn to be language consumers by providing them with the right learning experiences and requirements. Also, evaluation should be organized so that it promotes the growth of language usage; this is accomplished by testing mainly by language use activities. Language usage tasks provide evidence to teachers and assessors on a child's capacity to use language in communicative ways.

Recent performance evaluation advancements have produced new assessment guidance that instructs the approach to assessment through language use tasks. As a result, the first part of this chapter examines the assumptions and features of performance evaluation. Following that are some additional concepts of successful task-based evaluation of young learners. These assumptions support the evaluation methodology used in this book. Children can demonstrate their ability to use language by sharing meaning according to their intentions and unexpected ways depending on the situation by language use tasks.

Principles and frameworks for selecting language use testing assignments, whether for the classroom or external assessments, are needed. How do teachers and assessors choose the most appropriate evaluation assignments for young students? What types of testing exercises provide children with the best learning experiences and the best chance to demonstrate their abilities? Any children will be disadvantaged if testing assignments are chosen incorrectly. Any children can need assistance when performing activities; are there any ways that appraisal tasks may be pre-analysed to ensure that changes can be made to ensure the child's best performance? This chapter provides principles and frameworks for selecting evaluation activities.

Performance Assessment

The word ‘performance assessment' is used here as an umbrella term to refer to a group of related assessment methods, including ‘alternative' and ‘authentic' assessment. Performance evaluation is described as evaluation that “involves either the observation of actions in the real world or the simulation of a real-life activity” (Weigle, 2002). Assessment via selected-response items is avoided in these approaches. A selected-response object is as follows, in which children are asked to choose the correct expression. It is also known as a discrete point evaluation item because it is designed to test only one aspect of language knowledge (in this case, knowledge of proper use of personal possessive pronouns).

Choose the correct word to fill the blanks. You can use the words twice.

my	her	his	our	your	their	I	she	he	you

Rosé buys a new book. _____ book is one of the rare books in the world.
John has a wide yard. He lets people play kites in _____ yard.
I have one sister and one brother. _____ siblings live with my parents in Toronto.

Teachers in success assessments prefer to avoid using appraisal elements like the one above that explicit target vocabulary for language. Instead, performance tests allow learners to use the vocabulary for real-world reasons and in real-world or practical circumstances, and they evaluate their achievement in doing so. Grammar and vocabulary skills in children are measured as part of their success in real-world or practical challenges rather than individually in discrete-point measurement pieces. Teachers will observe and assess the overall success (did they complete the task?) as well as the elements of language usage, such as vocabulary and grammar, within the task performance (to what degree did they use a variety of vocabulary? How accurate was the performance?). Teachers and assessors make assessments on results by contrasting students' performance to the average performance of all learners, rather than comparing students' performance to the average performance of all learners.

The concepts of performance evaluation extend beyond testing and into the teaching and learning processes. The following are a summary of the assumptions and features of success assessment:

- Students are active participants rather than passive subjects
- Evaluation and guidance occur simultaneously and continuously
- Processes as well as products are evaluated
- Development and learning need to be recognized and celebrated
- Multiple indicators and sources of evidence are collected over time
- Results of the assessment are used to plan instruction, improve
- classroom practice, and optimize children's learning
- The assessment process is collaborative among parents, teachers,
- children, and other professionals as needed

(Jalongo, 2000. P. 287)

Thus, success assessment includes an emphasis on children's skills in real-world activities, as well as exposure to more comprehensive assessment features that promote, among other things, constructive engagement, attention to learning processes, and involvement of parents and children in the assessment process. Young learners learn better through concrete and practical experiences, and proof of their language learning is more likely to be present in language use testing exercises close to those of the child's natural world. In recent years, performance assessment has had a significant impact on assessment thinking; this influence has been incredibly intense in classroom-based assessment, where there have been more ways to apply some of the concepts than, for example, in external testing. However, there is also a movement to integrate aspects of performance-based evaluation into more structured assessment contexts, such as large-scale research.

Language Use Tasks

We will now take a closer look at the meaning and several instances of language usage activities. Tasks, traditionally described as teaching activities with a pedagogical goal (Purpura, 2004), have recently been distinguished by their capacity to evoke engagement and meaning negotiation, as well as involve learners in dynamic meaning-focused activities (Nunan, 1989, 1993; Berwick, 1993; Skehan, 1998). The concept of a language usage assignment used in this book reflects this focus on the communicative aims of tasks. A language-use task is described as 'an operation in which individuals use language to achieve a specific aim or objective in a specific situation' (Bachman and Palmer, 1996, p. 44).

Language usage exercises are goal-oriented, meaning that the learner understands what is expected of them and is situation-specific. Each instance of language use is almost entirely new (Bachman and Palmer, 1996, p. 44). Language use tasks may include listening, chatting, reading, writing, or combining these activities.

Children's language involvement in language usage activities requires a degree of spontaneity and creativity; they make their own sense, creating meaning or comprehending meaning, depending on the intent and conditions of the case. Children's imagination and spontaneity stem from their 'language resource,' that is, the language and language laws they have internalized. There may be unanalysed chunks of vocabulary or new rule-based constructions. Children use this vocabulary to achieve a communicative goal properly for the language use sense. Language usage functions can be carried out in a primary and assisted manner or a more extensive, complicated, and autonomous mode. We may not generally expect inventive language usage in language usage activities to be precise, comprehensive in vocabulary use, or acceptable, but we expect these characteristics to improve as practice increases. Language usage functions do not have to be loud or time-consuming (factors that are avoided in some teaching situations).

In the following example, a language testing exercise for beginning early language learners includes children filling in blanks in sentences to help them write about a story they have learned. They are instructed to fill in the blanks with their own words. The challenge differs from the previous task example's discrete-point object in that it has holes for children's own words. Children can write the story they've learned in their own words, with help from the part-sentences given, and the instructor can anticipate some spontaneous linguistic usage in the holes. This simple assignment asks children to write a story (the purpose) in the narrative language (the situation). The challenge should be made more open, with children having to pick what happened in the plot.

The following task shows an example of a language testing task for more experienced language learners. Children are asked to write down what happened during a scientific experiment that they saw. The questions direct them through the mission while also including a learning structure for a procedural genre.

Children are given a stapled booklet with eight pages, with a simple story written by the teacher through the pages. There are blanks in the sentences on some pages. Children are asked to fill in the blanks, and to draw a picture for each page, illustrating what is happening in the story. The children have already heard the story: it has been read to them several times, and talked about in classroom activities.

Page1: The little girl's name was Jane.
Page 2: One day she found a ……..
Page 3: She felt very …… about this.
Page 4: Then she ……..

(McKay, 2006, p. 101)

Children have observed a simple science experiment. They take notes. They are then asked to write what happened in the science experiment. They have been given a paper with headings to guide them.

Name: _____

What were we trying to find out? Describe the equipment that was used. What did we do?
What happened?
What did we find out?

(McKay, 2006, p. 102)

Language Use Tasks In the Classroom

There are almost limitless language usage tasks that can be used to test young language learners. Many language teaching activities in the classroom may be used for evaluation. Teachers can monitor children's success during the task, conduct on-the-spot evaluation while they teach, or set up the task for structured assessment with requirements made clear to the children at the start of the task. An on-the-run assessment is an assessment that is blended into the hectic pace of teaching. Williams (1984) provides a list of classroom language use activities in Table 6.1 that evokes the young learner language classroom and highlights the features of appropriate tasks for young language learner evaluation. Many, if not all, of the classroom teaching and learning activities, or portions of them, could be suitable for evaluation. Although these activities are designed for children in the lower

elementary grades, modifications that move towards more content-based learning would be appropriate for upper elementary students.

Table 6.1 ***Examples of classroom language use tasks for young learners (Williams, 1984, p. 209)***

Doing puzzles and solving problems	Writing and solving riddles	Using maps Growing plants
Measuring and weighing things	Conducting surveys (e.g., food, birthdays, traffic surveys)	Making things (e.g., witches, spacemen, stranded on an island)
Following and writing recipes	Interviewing people (e.g., parents, people in the neighbourhood, different occupations)	Inventing games (e.g., board games, writing the instructions)
Inventing and designing things (my ideal . . . A machine to . . . fashions)	Planning things (e.g., an outing, a party)	Reading and designing brochures
Choosing (e.g., films, clothes)	Writing letters (for real purposes)	Filling in forms
Designing and recording a TV programme	Finding out (e.g., what things are made of, what materials are used for, how things grow, whether objects float or sink)	Using songs and rhymes
Studying the local environment (e.g., plants, birds, buildings)	Making charts and graphs	
Listening to stories (a particularly motivating form of language input, and recommended as a daily activity)	Painting, drawing and talking about what we are doing	

Teachers will choose from various language use activities based on the children's proficiency level, their needs, and the program demands. Tasks can include problem-solving, information gap tasks (where children must find out information to complete the task), opinion gap tasks (where children must find out someone else's opinion to complete the task), effective gap tasks (where children must find out what others are feeling to complete the task), picture-based tasks, games, literature-based tasks, and drama tasks.

Many games and drama activities are appropriate for classroom appraisal, in which the teacher watches and records the children's success as the task's pacing progresses. Personalized appraisal exercises are suitable for young learners because

the subject is relevant to their own needs and lives. Personalized activities include expressive writing about oneself and one's family and friends, questionnaires and polls, and individual interviews about thoughts and ideas. Literature-based activities are ideal for all students (Falvey and Kennedy, 1997). Children's tales are typically used in literature-based activities. Children can, for example, read stories aloud (assessing their ability to read this level of text aloud), draw pictures based on a part of the story (assessing various constructs such as comprehension of the sequence of events in the story, comprehension of description), write questions about a story or a poem (assessing comprehension as well as ability to write questions), or finish an unfinished project (different levels of understanding are assessed based on whether the questions are literal or interpretive). The following is a straightforward writing assignment in which children are asked to read a poem and answer a question.

What is the wizard doing in bed? Write down what you think.

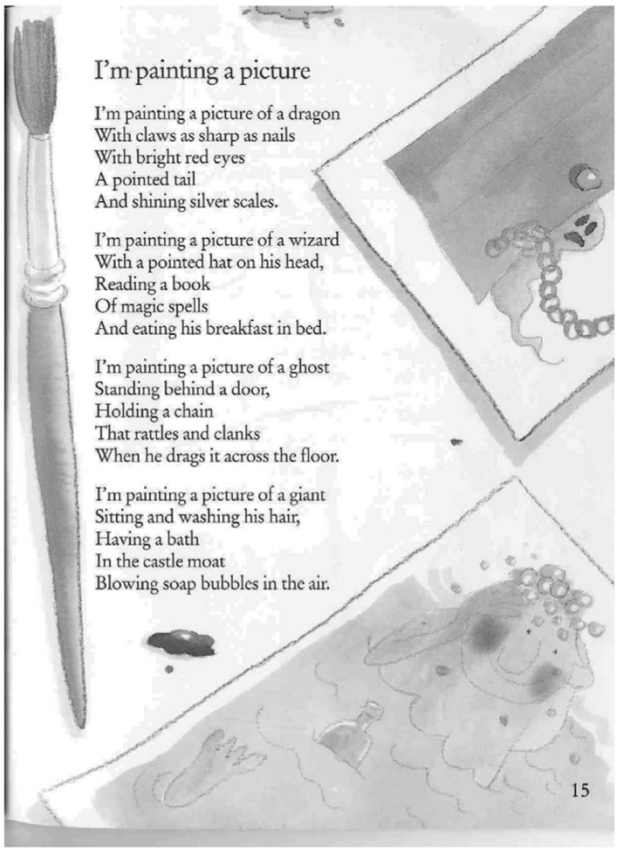


Figure 6.1 An example of a literature-based response (Poem by Foster and Lewis, 1996)

Language use assessment can be embedded in a variety of instructional activities in the classroom. There is at least one language use appraisal task (see teaching action in bold) in Cameron's (2001) teaching task system in Table 4.5 that allows the teacher to measure children's language use. Early language learners are asked to write their own sentences about Hani's weekend in this evaluation mission.

Table 6.2 an example of embedded language use assessment task in classroom (Cameron, 2001, p. 34)

TASK Say sentences about Hani's weekend			
	Preparation	CORE	*Follow up*
Language learning goals	Activate previously learnt lexis. Practise past forms of verbs.	Oral production of sentences from grid.	Written production of Hani's sentences. Composition of own sentences.
Teaching activities	Teacher-led: (1) Use of single pictures to prompt recall of lexis. (2) Divide board into two and recall/practise past forms. (3) Pairs practise with single pictures.	(1) Whole class introduction of grid and teacher modelling of sentences (2) Pair production of own sentences about Hani's weekend e.g. P1 points to aboxandP2 says sentence.	(1) Teacher writes key words on board, next to pictures. (2) Teacher models writing sentence from grid. (3) **Pupils write own sentences about Hani's weekend.** (4) Pair checking of accuracy.

Children would be bringing the words they learned and practiced in the training challenge into practice. There are also resources for on-the-spot assessments to observe children's improving vocabulary and syntax and their ability to form the sentences that are being practiced. Language use assignments provide teachers with chances to measure children's abilities to use language in the classroom.

Selecting Appropriate Assessment Tasks and Procedures for Young Learners

What are the guiding criteria for selecting assessment tasks? Are any assessment activities more effective than others? Language evaluation activities may be chosen by the classroom language instructor, the textbook author, or others such as other teachers in the school, test creators in the education department, and commercial testing firms. These language assessment tasks can stand alone or be part of a more extensive assessment procedure that includes assessment across several tasks. Assessment processes include, for example, instructor observation, resumes, and self-assessment.

Some first-base principles to guide the selection of assessment tasks and procedures

The following are some fundamental concepts to consider when choosing activities and methods for assessing young learners. These first principles are mainly derived from the curriculum and evaluation of young learners. They are then supplemented by more basic concepts and logical structures drawn from the evaluation sector.

1. Choose activities and practices that are appropriate for the characteristics of young learners: The tested characteristics of the learners will be known to the instructor and assessors, who will then be able to choose assignments and practices to fit these characteristics. Teachers and assessors must consider various considerations depending on their understanding of the task's intent and the characteristics of the learning scenario.
2. Examine the learners' most important language-use abilities: Teachers must ensure that children's skills are assessed for them to be effective in their language learning. The curriculum typically determines the breadth of expertise, talents, and abilities that must be learned and measured. A communicative education will require the opportunity to use the target language as a core purpose, as well as other similar goals such as sociocultural understanding, learning-how-to-learn skills, and language awareness.

If no curriculum exists, teachers must make their own decisions on the appropriate skills to test. Teachers should do a needs report, a survey of the types of language skills that children need at school, in the neighbourhood, and in potential language use programs (such as a class excursion to a target language-speaking environment). Language ability theoretical models, such as the Bachman and Palmer (1996) system of communicative language ability and the Common European Framework (Council of Europe, 2001), will advise teachers about the elements of language proficiency that must be tested.

3. Make assessment decisions that ensure the assessment is accurate and credible, as well as having a positive effect: Many questions must be answered about the evaluation tasks used in the classroom and external assessments. Is the role enough for all children? Is it measuring what it seems to be assessing? Is the task's scoring appropriate? Will the mission have a good effect, such as on schooling and the children's future development? The following section outlines methods for analyzing processes and operations in response to these critical problems.
4. 'Bias for the better,' but keep the expectations high: The best appraisal tasks and strategies are those that enable children to succeed to their full potential. Swain (1985) coined the phrase "bias for better" to express this concept. We must do everything in our power to provide children with the opportunity to achieve their full potential. Is it possible that a child's poor success is due to conditions in the task or practice that inhibit the child from demonstrating what he or she is capable of? Was the child given enough time? Were the orders issued in a clear and understandable manner? Were there any background sounds, such as children playing outside, that caused him or her to lose concentration? Were there any culturally related parallels in the role that the child was unfamiliar with? Was the mission or process adequately motivational for this particular child? The analytical method in the following segment contains several more questions that will assist teachers and assessors in giving children the best opportunity possible.

5. Engage students critically: Assessment exercises can be sufficient to reach the appropriate developmental and proficiency level for the children involved, but they may be deficient in academic difficulty for the children. A fundamental task is often expected and can be combined with more intellectually complex tasks. The problem here is that teachers are being asked to doubt their overall work selection – is there a level of academic difficulty for the children in at least some of the tasks?
6. Draw from multiple sources of information: When making decisions about children's skills, it is critical to consult multiple sources of knowledge, particularly in high-stakes scenarios. Where appropriate, teachers can gather data from various activities chosen to observe the desired range of behaviour. Where possible, they can use multiple techniques, such as observation, portfolios, self-assessment, quizzes, and assessments, to ensure that they obtain the most precise and comprehensive image of the child's ability. Making a judgment about a child's success based on a single source of knowledge is "dangerous, if not reckless" (Brown and Hudson, 1998). External testers are restricted to gathering knowledge from a single test, including up to six activities. As a result, the experiments are meticulously planned and thoroughly tested. Regardless, external assessments cannot capture the range and scope of the child's understanding and abilities. They are intended for specific reasons and cannot gather knowledge on the entire range of the child's skills in the same manner as the classroom teacher does, and has access to many sources of information during the process of her teaching and curriculum-based evaluation.

REFERENCES

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C. and Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing* 13, 3, 280–297.

Alderson, J. C. and Wall, D. (1993). Does washback exist? *Applied Linguistics* 14, 2, 115–129.

Alderson, J. C. and Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing* 13, 3, 280–297.

Allwright, R. (1982). Perceiving and pursuing learners' needs. In M. Geddes & G. Sturtridge (eds), *Individualisation* (pp. 24–31). Oxford: Modern English Publications.

Bachman, L. (1990). *Fundamental consideration in language testing*. NY: Oxford University Press.

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

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Berwick, R. (1993). Towards an educational framework for teacher-led tasks. In G. Crookes and S. M. Gass (eds.), *Tasks in a Pedagogical Context: Integrating Theory and Practice* (pp. 97–124). Clevedon, Avon: Multilingual Matters.

Black, P. & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80 (2), 139-148.

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability*, 21(1), 5–31.

Brindley, G. (2001). Assessment. In R. Carter & D. Nunan (Eds), *The Cambridge guide to teaching English to speakers of other languages* (pp. 137-143). Cambridge: Cambridge University Press.

Broadfoot, P. (2005). Dark alleys and blind bends: Testing the language of learning. *Language Testing*, *22*, 123-141.

Brookhart, S. M. (2013). Grading. In J. H. McMillan (ed.), *Research on classroom assessment* (pp. 257–272). Los Angeles, CA: Sage.

Brown, H. D., & Abeywickrama, P. (2018). *Language Assessment Principles and Classroom Practice 3rd Edition*. Philadelphia: Pearson Education.

Brown, J. D. and Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly 32*(4), 633–75.

Buck, G. (2001). *Assessing Listening*. UK: Cambridge University Press.

Cameron, L. (2001). *Teaching Languages to Young Learners*. Cambridge: Cambridge University Press.

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. Reprinted in Allen, H. B. and Campbell, R. N. (eds) (1965), *Teaching English as a Second Language: A Book of Readings* (pp. 313–330) New York: McGraw Hill.

Carroll, J. B. (1981). Twenty-five years of research on foreign language aptitude. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83-118). Rowley, MA: Newbury House.

Chappelle, C. A. & Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (Ed). *The companion to language assessment*, (Vol. 3, pp. 1079-1097). NY: John Wiley & Sons.

Cheng, L. (2008). Washback, impact, and consequence. In N. Hornberger (Ed), *Encyclopedia of language and education* (2nd ed., pp. 2479-2494). NY: Springer.

Cheng, L. (2014). Consequences, impact, and washback. In A. J. Kunnan (ed.), *The companion to language assessment* (pp. 1130– 46). Chichester: John Wiley & Sons. doi:10.1002/9781118411360. wbcla071.

Choi, I. C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, *25*, 39-62.

Chun, C. (2006). Commentary: An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly*, *3*, 295-306.

Cizek, G. J. (2010). An introduction to formative assessment: His- tory, characteristics, and challenges. In H. Andrade & G. Cizek (eds), *Handbook of formative assessment* (pp. 3–17). New York: Tay- lor and Francis.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Davidson, F. & Lynch, B. (2002).*Testcraft: A teacher's guide to writing and using language test specifications*. New haven, CT: Yale University Press.

Davies, A. (2003). Three heresies of language testing research. *Language Testing*, *20,* 355-368.

Enright, M., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P. and Schedl, M. (2000) *TOEFL 2000 Reading Framework: A Working Paper (TOEFL Monograph MS-17)*. Princeton, NJ: Educational Testing Service.

Falvey, P. and Kennedy, P. (eds.), (1997). *Learning Language through Literature*. Hong Kong: Hong Kong University Press.

Foster, J. and Lewis, J. I. (1996). *You Little Monkey and Other Poems for Young Children*. Oxford: Oxford University Press.

Fox, J. & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first (L1) and second (L2) language test takers. *Assessment in Education*, 14(1), 9–26.

Fox, J., Haggerty, J. & Artemeva, N. (2016). Mitigating risk: The impact of a diagnostic assessment procedure on the first-year experience in engineering. In J. Read (ed.), *Post-admission language assessment of university students*. Cham: Springer International. DOI: 10.1007/978-3-319-39192-2

Fulcher, G. (1999). Assessment in English for academic purposes: putting content validity in its place. *Applied Linguistics* 20, 2, 221–236.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education, An Hachette UK Company.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment*. NY: Routledge.

Green, A. (2007). *IELTS Washback in Context: Preparation for Academic Writing in Higher Education*. Cambridge: Cambridge University Press.

Green, A. (2014). *Exploring language assessment and testing language in action*. NY: Routledge.

Gorsuch, G. (2000). EFL educational policies and educational cultures: Influences on teachers' approval of communicative activities. *TESOL Quarterly*, *34*(4), 675–710.

Grounlund, N. E. (1998). *Assessment of students achievement* (6th ed). Boston, MA: Allyn & Bacon.

Haines, M. M., Stansfeld, S. A., Head, J. and Job, R. F. S. (2002). Multilevel modelling of aircraft noise on performance tests in schools around Heathrow Airport London. *Journal of Epidemiology and Community Health*, 56, 139–144.

Harris, D. P. (1968). *Testing English as a second language*. New York: McGraw Hill.

Hughes, A. (2003). *Testing for language teachers* (2nd ed). Cambridge: Cambridge University Press.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfield, V. F., & Hughey, J. B. (1981). *Testing ESL composition: a practical approach*. Rowley, Mass: Newbury House.

Jalongo, M. R. (2000). *Early Childhood Language Arts*. Boston: Allyn and Bacon.

Kane, M. (2010). Validity and fairness. *Language testing*, *27*(2), pp. 177-182.

Kane, M. (2016). Validity as the evaluation of the claims based on test score. *Assessment in education*, *23*(2), 309-311.

Koretz, D. M. and Hamilton, L. A. (2006). Testing for accountability in K-12. In Brennan, R. L. (ed.), *Educational Measurement*. 4th edition. Westport, CT: American Council on Education/ Praeger, 531–578.

Lado, R. (1961). *Language testing: the construction and the use of foreign language use*. London: Longman.

Latham, H. (1877). *On the Action of Examinations Considered as a Means of Selection*. Cambridge: Dighton, Bell and Company.

Lewkowicz, J. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, *17*, 43-64.

McKay, P. (2006). *Assessing young language learners*. UK: Cambridge University Press.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick' legacy. *Language Assessment Quarterly*. *3*, pp. 31-51.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational measurement* (3rd edition, pp. 13–103). New York: Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing* 13, 3, 241–256.

Mousavi, S. A. (2009). *An encyclopedic dictionary of language testing* (4th Ed). Tehran: Rahnama Publications

Nicole, N. (2008). *Washback on classroom practice for teachers and learners*. Unpublished MA dissertation, University of Leicester.

Nunan, D. (1989). *Designing Tasks for the Communicative Classroom*. Cambridge: Cambridge University Press.

Nunan, D. (1993). Task-based syllabus design: selecting, grading and sequencing tasks. In G. Crookes and S. M. Gass (eds.), *Tasks in a Pedagogical Context: Integrating Theory and Practice* (pp. 55–68). Clevedon, Avon: Multilingual Matters.

Powers, D. E., Albertson, W., Florek, T., Johnson, K., Malak, J., Nemceff, B., Porzuc, M., Silvester, D., Wang, M., Weston, R., Winner, E. and Zelazny, A. (2002). *Influence of Irrelevant Speech on Standardized Test Performance*. TOEFL Research Report 68. Princeton, NJ: Educational Testing Service.

Purpura, J. (2004). *Assessing Grammar*. Cambridge: Cambridge University Press.

Ruch, G. M. (1924). *The Improvement of the Written Examination*. Chicago: Scott, Foresman and Company.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Swain, M. (1985). Large-scale communicative language testing: A case study. In Y. P. Lee, A. C. Y. Fok, G. Lord and G. Low (eds.), *New Directions in Language Testing* (pp. 35–46). Oxford: Pergamon Press.

Taylor, L. (2005). Washback and impact. *ELT Journal*, *59*(2), 154-155.

Taylor, C. S. & Nolen, S. B. (2008). *Classroom assessment: Supporting Teaching and Learning in Real Classrooms* (2nd ed). New Jersey: Pearson Education.

Wall, D. (2005). *The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*. Cambridge: Cambridge University Press.

Wall, D. and Alderson, J. C. (1993). Examining washback: the Sri Lankan impact study. *Language Testing, 10*(1), 41–69.

Wall. D. and Alderson, J. C. (1996). Examining washback: the Sri Lankan impact study. In Cumming, A. and Berwick, R. (eds), *Validation in Language Testing*. Clevedon: Multilingual Matters, 194–221.

Wang, H. & Cheng, L. (2009). Factors affecting teachers' curriculum implementation. *The Linguistics Journal*, *4*(2), 135–66.

Watanabe, Y. (2004a). Teacher factors mediating washback. In Cheng, L., Watanabe, Y. and Curtis, A. (eds), *Washback in Language Testing*. Mahwah (pp. 129–146). NJ: Lawrence Erlbaum.

Watanabe, Y. (2004b). Methodology in washback studies. In Cheng, L., Watanabe, Y. and Curtis, A. (eds), *Washback in Language Testing* (pp. 19–36) Mahwah, NJ: Lawrence Erlbaum.

Waugh, C. K., & Gronlund, N. (2012). *Assessment of students achievement* (10th ed.). White Plains, NY: Pearson.

Weigle, S. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, England: Palgrave MacMillan.

White, R. (1988). *The ELT curriculum: Design, innovation and management*. Oxford: Basil Blackwell.

Williams, M. (1984). A framework for teaching English to young learners. In C. Brumfit, J. Moon and R. Tongue (eds.), *Teaching English to Children*. Harlow, Essex: Longman.

Zumbo, B. D., & Hubley, A. M. (2016). Bringing consequences and side effects of testing and assessment to the foreground. *Assessment in education: Principles, Policy, & Practice*, *23*(2), 299-303.