MODULE 1

# Definitions and Principles of Language Assessment

Fachrurrazy, M.A., Ph.D.

## INTRODUCTION

This is the first of a series of nine modules of Assessment in Language Teaching for Magister students at the Open University. This first module deals with definitions and principles of language assessment, as a basic knowledge for learning all the other modules in this course. The introduction part of this first module presents the objective, scope, and learning guide for the students. Read them carefully, because they will be useful to help you understand the contents of this module.

### OBJECTIVES

By learning this module, the students are expected to be able to:

1. define assessment terms and purposes,
2. identify trends in testing approaches,
3. identify current issues in testing, and
4. explain the principles of good and effective language assessment.

### SCOPE

Based on the objectives stated above, the topics to be discussed in this module include:

1. definitions of test, measurement, assessment, and evaluation, and their relationship to teaching;
2. various approaches in language testing/assessment, i.e. pre-scientific approach, psychometric-structuralist approach, integrative/pragmatic/ interactive approach, and performance-based approach;

3. current issues in classroom testing;
4. principles of language assessment, i.e. practicality, reliability, validity, authenticity, and Washback.

## LEARNING GUIDE

To get a full understanding of the contents of this module, students are advised to follow the learning guide below.

1. Read the introduction of this module, so that you know what is expected to be mastered in general.
2. Read and learn the activities one by one in details, and do all the exercises and formative tests at the end of each activity.
3. Finishing the tasks, check your responses or answers to the exercises and tests with the answer keys at the end of this module.
4. If your answers or responses are wrong, learn again the related activity. Find why they are wrong, and make corrections.
5. In addition to learning this module, you are also advised to widen and deepen your knowledge by reading the suggested references stated at the end of this module.

## MAIN ACTIVITIES

This module consists of four main topics, namely: (1) definitions of some basic terms of assessment and assessment purposes, (2) approaches in language assessment, (3) current issues in language assessment, and (4) principles of good and effective language assessment. These four topics are discussed in the following Learning Activities 1 to 4 in this module. Pay attention to the examples and illustrations provided in each activity because they will help you understand the contents of this module more easily.

LEARNING ACTIVITY 1

# Definitions and Purposes of Assessment

Two main points are discussed in this activity, i.e. definitions of some basic terms in language assessment and purposes of assessment. Students are expected to be able to understand the details of the two points mentioned above. For that purpose, the students are advised to read through the explanation carefully and do the exercises and the summative test.

## A. DEFINITIONS OF TEST, MEASUREMENT, ASSESSMENT, AND EVALUATION

In the field of teaching and learning, we often find the terms *test, measurement, assessment,* and *evaluation*. In this Magister (S-2) Program, we need to know the meanings and the relationship of all of these terms. Learn the following explanations.

*Test* is a method, a tool or an instrument for measuring students' ability, mastery, or achievement of learning (Brown & Abeywickrama, 2010: 3). The tool or instrument here can be in the forms of questions to be answered by students, true-false items or multiple choice items for students to answer. The questions, true-false items, multiple choice items, or any other forms we make, are tools or instruments which are called tests. Tests are always formal because we prepare and construct them, whether they are written or spoken. The detailed discussion about various test types will be presented in Module 2.

*Measurement* refers to the quantifying of the result of a test. It is usually in the form of figures or scores (Bachman, 1990:18-20; Brown & Abeywickrama, 2010: 4-5). For instance, student A gets 47, student B gets 75 from their test. The scores 47 and 75 are the results of measurement, and still do not mean anything, because there is no interpretation yet whether each of the scores is good or bad, whether it means pass or fail. In some cases, the scores are called *raw scores*.

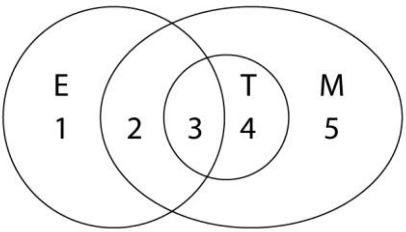
The next term is *assessment* which is claimed to have a wider meaning. It includes *formal tests* and also *informal tests. Informal tests* are usually incidental or unplanned, and can be in the forms of observation and/or comment (Brown & Abeywickrama, 2010: 3). When a teacher approaches

his/her students while they are working in a group, and gives comment "You are on the right track. Go on!", this teacher does an informal test/assessment. Of course, a teacher may also conduct a *formal test* by giving his/her students, for example, a reading text and several comprehension questions to be answered, because this exercise or test belongs to an assessment. Related to assessment, we may encounter the terms *formative assessment* and *summative assessment*. *Formative assessment* is a kind of progress assessment, in which the students are assessed while they are still in their learning progress/process, e.g. daily/weekly test or mid-term test. In our module here, formative test refers to the test after every activity. *Summative assessment* is conducted at the end of a course or program.

*Evaluation* is defined as giving interpretation or judgment to something, which can be student's score or attainment (Bachman, 1990: 22). In the example of measurement above, when we judge that student A fails, because he/she gets only 47, and student B passes, because his/her score is 75, we make a judgment or evaluation. However, we can make an evaluation without measurement. For example, when a student answers our question correctly, and we say "Excellent!", we have made a non-measurement evaluation.

*Teaching* is a process of delivering knowledge or skill to students. Teaching usually involves testing or assessing. This is because in the process of teaching we need to know whether the students have understood what we teach, whether the students have achieved the target of competence, or how far the students have progressed. Therefore, there is a relationship of test, measurement, assessment, evaluation, and teaching. Different experts have different views about their relationships.

Bachman (1990: 23), for instance, proposes a relationship of test, measurement, and evaluation as in the following figure.

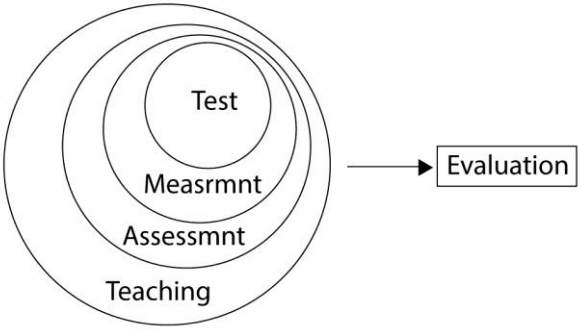



(Adopted with a slight adjustment from Bachman, 1990: 23)

Notes:
E = Evaluation
T = Test
M = Measurement

As seen in the above figure, Bachman does not include the term assessment, probably because assessment was not yet popular at that time. The relationship of test, measurement and evaluation in Bachman's model produces 5 areas. Area 1 is evaluation without test and measurement. This can happen when a teacher makes a description of students' performance for diagnostic purpose. Area 2 is evaluation involving measurement, for example, a teacher ranks his/her students, and then determines their grades. Area 3 is a test which is used for evaluation. This is the common practice in our schools, in which a teacher uses the scores of a test to determine whether each of his/her students reaches the minimum passing level (MPL). Area 4 is a test alone. An example of this is a test administered to students for certain research purpose. Area 5 is measurement only, in which a teacher may group students based on the criteria of male and female, high achievers and low achievers, etc.

Another relationship of test, assessment, and teaching was made by Brown (2004: 5), where test is a part of assessment, and assessment, in turn, is a part of teaching. This relationship was then revised by Brown and Abeywickrama (2010: 6) to include measurement and evaluation. The following figure is the revised model as made by Brown and Abeywickrama.




(Adopted from Brown and Abeywickrama, 2010: 6)

Notes:
Measrmnt = Measurement
Assessmnt = Assessment

From this model, we can see that test is a part of measurement, measurement is a part of assessment, and assessment is a part of teaching. All these parts are used as the bases for evaluation. However, there is also a view that in practice, assessment and evaluation have almost the same meanings, and are used interchangeably (Saukah, 2013: 3). The difference is only that assessment is in the form of description, while evaluation is judgmental. In our course here, we acknowledge the different views as presented above.

## B. ASSESSMENT PURPOSES

A test or assessment can serve a number of purposes, namely: *selection*, *screening*, *placement*, *prognosis*, *diagnosis*, *research*, *program evaluation*, *accountability*, and *achievement*. These 9 purposes of assessment are discussed below.

The first is *selection* assessment. The purpose of selection assessment is to select from the test-takers a number who meet certain criteria. One of the criteria is the use of quota. For example, the Open University will take a quota of 100 students for Magister degree program, and the candidates are 200. We make and administer a test, and then the scores of the test from these 200 candidates are ranked from the highest to the lowest, and then we take the first or highest 100 to pass the test. Selection assessment is usually conducted before a program begins.

*Screening* assessment aims at selecting test-takers who can pass the criterion of minimum passing level (MPL). Therefore, screening assessment is not based on a quota. The test-takers who reach the MPL will pass. This is the case of the final examination in elementary school, junior high school, and senior high school in Indonesia. If, for example, there are 300 students of the $12^{th}$ grade at a senior high school who take the final examination, and all of them can reach the MPL, all of them pass. There is no limitation of the number of students who may pass. Screening assessment is usually conducted at the end of a program. As in the example above, it is conducted at the end of grade $12^{th}$.

There is a possibility to combine selection and screening. As in the case of Open University selection above, the criterion can be a quota of 100 students, on a condition that the score is not lower than 75. Therefore, from 200 candidates there is a possibility that only 80 candidates reach the scores of equal or higher than 75. Therefore, the Open University will accept fewer than the number determined in the quota because others do not reach the score of 75. However, if the score 75 is obtained by more than 100 candidates, we usually take 100 only, based on the quota (except the quota is changed).

*Placement* assessment aims to group students, and is conducted at the beginning of a program. The grouping is usually based on certain criteria, e.g. homogeneity or heterogeneity. For example, a number of students are grouped into high achievers, middle achievers, and low achievers, and put them respectively in 3 separate classes. The grouping is usually for the purpose of ease of treatment or teaching, because they are in a class of homogeneous ability. (Remember, heterogeneous class also has an advantage, can you think of it?).

Sometimes, we want to know whether our students will be successful or not when attending our course; so, we conduct an assessment which is called *prognosis* assessment. The aim of this assessment is to predict whether the students will succeed or not in taking our course. Like placement assessment, this prognosis assessment is also conducted at the beginning of a program.

*Diagnosis* assessment is an assessment which is used to detect students' problem so that the teacher can give the necessary treatment or remedy. For example, when a teacher teaches writing and observes that his/her students have difficulty in writing a simple paragraph. The teacher makes an assessment to find the students' problem. When it is found that the students' problem is, for instance, a lack of vocabulary, then the teacher improves the students' vocabulary before continuing the writing exercises. Diagnosis assessment is usually conducted in the middle of a course or program.

An assessment can also be made for *research* purpose. An S-2 student may construct an assessment and administer it to a class of junior high school students, and collect some data for writing his/her thesis. The junior high school students do not get the result of the assessment, and the result of the assessment does not affect their final grades. The students participate in the assessment just to help the researcher collect data. An assessment for this research purpose can be conducted any time (beginning, middle, or end) of a course or program.

An assessment can be conducted in the middle or end of a course or program for the purpose of course*/program evaluation*. For example, we teach Language Assessment course for one semester. At the end of the semester, we want to know whether the course is effective or not, whether it is relevant to our students or not, or whether it is appropriate or not for the level of our students. We construct an assessment, not for knowing our students' achievement but for evaluating our course; therefore, it is called course/program evaluation.

The next is an assessment for the purpose of *accountability*. For example, the Open University has a cooperative program sponsored by a Provincial Office of Education to train a number of junior high school teachers in the province. The sponsor may ask the Open University for a report of the progress of the training participants, and the university conducts an assessment to know how far the participants of the training have progressed, and then gives a report to the sponsor. The assessment conducted by the university is called an assessment for accountability purpose.

The last is an assessment aiming at knowing students' *progress or achievement*. As discussed earlier, formative assessment is used for the purpose of knowing students' progress during a course, and summative assessment is for knowing students' achievement at the end of a course. Both formative and summative assessments are achievement assessments.

The assessment purposes discussed above can be summarized as in the following table.

| No. | Type of Purpose | Objective | Time |
|---|---|---|---|
| 1. | Selection | to select | before a program |
| 2. | Screening | to filter | end of a program |
| 3. | Placement | to group | beginning of a program |
| 4. | Prognosis | to predict | beginning of a program |
| 5. | Diagnosis | to remedy | middle of a program |
| 6. | Research | to collect data | any time during the program |
| 7. | Program evaluation | to evaluate a program | middle or end of a program |
| 8. | Accountability | to give a report | middle or end of a program |
| 9. | Achievement or progress | to know the attainment | middle or end of a program |

Besides purposes, there are three other types of assessment, based on the materials or contents to be assessed. The first is called *achievement* assessment. In the above explanation, we have mentioned achievement

assessment based on purpose; while here we see it from the point of view of materials which are used in constructing the assessment. It is called achievement assessment when the contents of the assessment are based on what have been taught. After teaching for one semester, for example, a teacher makes a summative assessment for his/her students based on the materials he/she has taught in the semester. This assessment is called achievement assessment because the teacher uses his/her teaching contents to know how far the students have achieved what was taught to them.

The second type is *proficiency* assessment. The contents of this assessment are those which are considered to meet the criteria of proficiency. For example, a proficient English teacher is a teacher who has high mastery of the English language and the teaching methodology. Then, we construct an assessment to measure both the teacher's mastery of the English language and the teaching methodology. The contents of this assessment are not what have been taught to or learned by the teacher previously, but based on the target of proficiency a teacher should attain. In this case, TOEFL (Test of English as a Foreign Language) test can be classified as an English proficiency test because it is based on the criteria of English proficiency, not what has been learned by or taught to the test-takers. The criteria of proficiency in TOEFL are mastery of general English, which covers listening, reading, speaking/writing, grammar, and vocabulary. Another well-known proficiency assessment is IELTS (International English Language Testing System).

The third is *aptitude* assessment. The contents of this assessment are determined based on the criteria of aptitude. Language aptitude test, for example, contains test items for measuring whether someone has an aptitude or talent to learn a language. Therefore, the content of the test can be, among others, measuring whether the test-taker has sensitivity about the distinction of sounds in minimal pairs, or the similar meanings from two different sentence constructions.

## EXERCISE 1

To check your understanding of this first activity, answer all the questions below.

1) Where is the place of assessment in teaching?
2) What is a specific characteristic of assessment as compared with evaluation?
3) What is summative assessment?
4) What is the difference between selection and screening assessment?
5) Why is formative assessment regarded as achievement assessment?

## SUMMARY

From the explanation above we can conclude that test, measurement, assessment, and evaluation have different meanings, but they are related to each other. The relationships may be viewed differently by different experts. A test or assessment can be used for several purposes, i.e. selection, screening, placement, prognosis, diagnosis, research, program evaluation, accountability, or achievement. In another classification, proficiency, achievement, and aptitude tests are types of test or assessment based on the materials to be tested.

## FORMATIVE TEST 1

To check further understanding of the explanation in Learning Activity 1, answer the following questions.

1) What are the advantages and disadvantages of using TOEFL for the final examination of senior high school students in Indonesia?
2) Write an example of a description of a student's speaking performance for diagnostic purpose, as stated in area 1 of Bachman's model.
3) The purpose of placement test is to group students in a homogeneous class or heterogeneous class. Mention the strength and weakness of each grouping.
4) How can a school get benefit from assessment for research purpose?

5) When a teacher describes a student's ability, such as, "For your speaking ability, you are fluent enough, there is no problem with your pronunciation, but you need to pay attention to the word stress and sentence stress", is this a test, assessment, or evaluation? Justify your answer.

If you have finished an exercise, look at the key answers at the end of the module. Evaluate your answers. When you get at least 80% right, you can go to another exercise, but if you don't, review the discussion and examples again. Then, do exercise once more. The following is how to evaluate your exercise and your test.

Formula:

$$\text{Level of mastery} = \frac{\text{The number of the reigh answer}}{\text{The number of the items}} \times 100\%$$

Level of mastery:
90 - 100% = very good
80 - 89% = good
70 - 79% = sufficient
< 70% = Insufficient

LEARNING ACTIVITY 2

# Trends/Approaches in Language Assessment

The purpose of this second learning activity is to identify the various trends or approaches in language assessment. Read the following explanation carefully and do the exercise and the summative test.

It seems that it is not complete if we learn language assessment without also learning the history of its approaches. The history of language assessment approaches, which cannot be separated from the history of language teaching methods, will give us insights for understanding the changes of the methods of assessing language. This activity is, therefore, focused on the discussion of the chronology of language assessment approaches and their relation to the language teaching approaches/methods. The contents of the discussion are taken mainly from Brown and Abeywickrama (2010: 12-16), Heaton (1988: 15-26) and Weir (1990: 1-15). They are as follows.

## A. ESSAY-TRANSLATION APPROACH

We may have known that there have been various methods in teaching English as a foreign language (TEFL). The popular methods, among others, are Grammar Translation Method (GTM), Direct Method (DM), Audio Lingual Method (ALM), and Communicative Language Teaching (CLT). The approaches in language assessment have developed in line with the development of language teaching methods.

The oldest language teaching method known was Grammar Translation Method (GTM). This method had been used up to the beginning of the 20th century. GTM was mainly characterized by the use of translation in language teaching and memorization of grammar rules. The language testing approach at that time followed the characteristics of GTM. It was called Essay-translation approach or Pre-scientific testing stage. This approach was characterized by the use of translation of words, sentences, or paragraphs, from the first language (e.g. Indonesian) to the target language (i.e. English), or the other way around; and the use of grammatical analysis, such as asking

students to identify or define parts of speech, or to mention ways to change singular to plural nouns. The scores of a test are determined by subjective judgment of the teacher. The teacher does not need any specific skill to be able to construct a test.

## B. DISCRETE-POINT APPROACH

The emergence of Audio Lingual Method (ALM), as a development of Direct Method (DM), in the period of post-World War II, changed the approach in language teaching as well as in language assessment. ALM was influenced by the trend of Behaviorism in psychology, i.e. the use of Stimulus-Response-Reinforcement as the teaching and learning strategy. In ALM, TEFL was defined as forming a set of (new) habits. The teaching of grammar was a priority, and was done based on the idea of contrastive analysis, i.e. the teaching points should emphasize the aspects of the target language which are different from those of the first language. In the area of assessment, ALM suggests that language elements (pronunciation, grammar, and vocabulary) and language skills (listening, speaking, reading, and writing) should be tested separately. ALM held the idea that we should teach and also test one thing at a time. The testing approach in the period of ALM was called Psychometric-structuralist approach or Discrete-point approach. This approach emphasizes validity, reliability, and objectivity of a test, and their calculation using statistic formulae. As a consequence, multiple choice test type was preferred, because it was considered most objective and easy for statistic calculation. The discussion and examples of multiple choice and other objective test types are provided in details in Module 2 of this course.

In the 1980s, the Psychometric-structuralist approach was then criticized. It was found that this Discrete-point approach was too much decontextualized, which means that the test of grammar, for instance, should not be mixed with the test of reading comprehension. In reality, the use of grammar depends much on the context of reading. This means that some kinds of test combine several aspects of language. In another example, the test of speaking automatically involves pronunciation, grammar, vocabulary, as well as fluency. Besides, there was a belief that the whole is not always the same as the sum total of its parts (Gestalt theory). From here, there emerged the idea of integrative test.

## C. INTEGRATIVE APPROACH

Integrative test, as the name indicates, combines or integrates two or more aspects of language, such as in the example of testing speaking mentioned above. It was found that several types of test, such as cloze test, dictation, oral interview, essay type, and translation, are inherently integrative. Soon, these kinds of test, especially cloze test and dictation, became popular integrative tests. The construction and formats of some of these integrative tests will be discussed in details in the other modules in this course.

The popularity of cloze and dictation as integrative testing devices led to the idea of unitary competence hypothesis (Weir, 1990: 5) or unitary trait hypothesis (Brown and Abeywickrama, 2010: 14), in which the mastery of language elements and language skills has a correlation. However, later research evidence showed that the hypothesis was proven untrue; therefore, the hypothesis was abandoned (Brown and Abeywickrama, 2010: 14). In addition, the use of cloze and dictation as integrative tests were also criticized, i.e. that cloze and dictation only measure the test-taker's linguistic competence, and not linguistic performance (Weir, 1990: 6). Linguistic performance involves an ability to say what, to whom, where and when.

## D. COMMUNICATIVE TESTING APPROACH

The development of Communicative Language Teaching (CLT) method in the end of 1970s changed the language testing approach. CLT emerged as a response to the weaknesses of ALM, and as a method which accommodates the idea of communicative competence (Richards and Rodgers, 2001: 159-161). As the core concept of CLT, Canale and Swain (1980, cited in Richards and Rodgers, 2001: 160) proposed that the target of language teaching and learning should be communicative competence. This communicative competence consists of four sub-competences, namely: (1) grammatical competence, which refers to the mastery of phonological, grammatical, and lexical aspects; (2) sociolinguistic competence, which refers to appropriate or inappropriate use of language in society; (3) discourse competence, which refers to an ability to interpret messages from wider contexts; and (4) strategic competence, which refers to an ability to use strategies or ways to initiate and maintain communication despite the limited mastery of the language.

## E. PERFORMANCE-BASED APPROACH

In accordance with the concept of CLT, the trend of language assessment changed from discrete-point to communicative assessment. This communicative assessment attempts to accommodate the possible correspondence between the language test performance and the real-world activities. From here we arrive at the idea of performance-based assessment, authentic assessment, task-based assessment, and some other terms (which will be explained as alternative assessment in the following Learning Activity 3). This kind of assessment includes "oral production, written production, open-ended responses, integrated performance (across skill areas), group performance, and other interactive tasks" (Brown and Abeywickrama, 2010: 16).

In performance-based assessment, the test-takers are required to perform certain tasks, and are assessed while they are in the process of doing the tasks. For example, a test-taker is asked to retell a story or to borrow a book from his/her friend, because he/she forgot to bring his/hers. Using these tasks, we can see examples of authentic assessments because the tasks imitate real-life activities. However, performance-based assessment takes longer time in its administration, compared with a paper-and-pencil test. That is why for practical purpose, we still find the wide use of paper-and-pencil tests nowadays. Ideally, we need to move to performance-based assessment.

## F. SPECIFIC INSTRUCTIONAL OBJECTIVES

Setelah memelajari materi pada unit 2, Anda diharapkan dapat:

1. menunjukkan letak suatu benda pada gambar sesuai dengan informasi yang disimak.
2. mengidentifikasi kata-kata berdasarkan gambar yang sesuai dengan deskripsi yang diutarakan secara lisan.

### EXERCISE 2

To check your understanding of this first activity, answer all the questions below.

1) What is the teaching and learning strategy proposed by Behaviorism?

2) What is meant by "unitary trait hypothesis"?
3) When a student says to his/her teacher "I would like to ask a question", instead of "I want to ask a question", what communicative competence is shown by this student?
4) A student who simulates shopping at a supermarket and makes a bargain dialog is considered not an authentic task. Why?
5) Why is (re)telling a story regarded as an authentic task?

## SUMMARY

In this Learning Activity 2 we learn that there are various testing or assessment approaches, starting from essay-translation approach or pre-scientific testing stage, psychometric-structuralist approach, integrative approach, communicative approach, up to performance-based approach. These approaches develop chronologically, and in accordance with the development of language teaching methods.

## FORMATIVE TEST 2

Answer the following questions.

1) Why Discrete-point approach is also called Structuralist approach?
2) Give an example of a contrast between Indonesian and English in the aspect of grammar!
3) Why is dictation considered as an example of integrative test?
4) What is lacking in linguistic competence compared with linguistic performance?
5) We have come to the era of performance-based assessment, but why is paper-and-pencil test still used?

If you have finished an exercise, look at the key answers at the end of the module. Evaluate your answers. When you get at least 80% right, you can go to another exercise, but if you don't, review the discussion and examples again. Then, do exercise once more. The following is how to evaluate your exercise and your test.

Formula:

$$\text{Level of mastery} = \frac{\text{The number of the reigh answer}}{\text{The number of the items}} \times 100\%$$

Level of mastery:
90 - 100% = very good
80 - 89% = good
70 - 79% = sufficient
< 70% = Insufficient

LEARNING ACTIVITY 3

# Current Issues in Language Assessment

Learning Activity 2 above discussed trends in language assessment approaches. Other trends which have become current issues in language assessment are: alternative assessment, computer-based testing, and assessment for 2013 Curriculum in Indonesia. These three issues are discussed in this Learning Activity 3. Read carefully the following explanation and do the exercise and the formative test.

## A. ALTERNATIVE ASSESSMENT

Prior to the last decade of the 20th century, the terms that were often used were *testing*, *measurement*, or *evaluation* (e.g. Bachman, 1990; Heaton, 1988; Hughes, 1989; Oller, 1979; Popham, 1978). Starting from 1990s, there has been a new term, i.e. *assessment* (e.g. Brown, 2004; O'Malley and Pierce, 1996). Along with the emergence of the term assessment, there has been an innovation in language testing or assessment, using various names. They are: alternative assessment, informal assessment, process assessment, authentic assessment, performance-based assessment, classroom-based assessment, and contextualized assessment. Basically, those terms refer to similar concepts, namely, the ideas that conventional or objective tests, such as binary choice or multiple choice, are not enough to assess language mastery of students. The promotion of the above assessment terms was also stimulated by a demand for a match between test tasks and real-life activities in CLT era (see Learning Activity 2). In this module we are introduced to the various terms which show different emphases. Brown and Abeywickrama (2010: 18) show the difference between traditional assessment and alternative assessment in their table as quoted below.

| Traditional Assessment | Alternative Assessment |
|---|---|
| Standardized exams | Continuous long-term assessment |
| Timed, multiple-choice format | Untimed, open-ended responses |
| Decontextualized test items | Contextualized communicative tasks |
| Scores suffice for feedback | Individualized feedback |

| Traditional Assessment | Alternative Assessment |
|---|---|
| Norm-referenced scores | Criterion-referenced scores |
| Focus on discrete answers | Open-ended, creative answers |
| Summative | Formative |
| Oriented to product | Oriented to process |
| Noninteractive performance | Interactive performance |
| Fosters extrinsic motivation | Fosters intrinsic motivation |

Note: The terms in the table above which may need explanation are, first, *norm-referenced* and *criterion-referenced* scores. Norm-referenced scores mean the spread of scores in percentile terms to produce normal distribution; and criterion-referenced scores mean the scores which are compared with certain criteria (e.g. minimum passing level), not to form normal distribution. These terms will be explained in more details in the module of "Scoring Interpretation" (Module 8). The second terms are *extrinsic* and *intrinsic motivation*. Extrinsic motivation is motivation which comes from outside one's self, e.g. a student is motivated to learn harder because his/her parents promise to give a present if he/she succeeds. Intrinsic motivation is motivation which comes from within one's self, e.g. a student learns English because he/she feels that English is important for him/her.

The examples of alternative assessment are portfolio, journal log, conference, self/peer assessment, etc. The detailed description of the types of alternative assessment will be discussed in Module 7 in this course.

As stated above, the innovation in assessment uses various terms. Different experts use different terms to show different emphases. Learn the following explanation.

The term *alternatives in assessment* or *alternative assessment* is used by Brown and Abeywickrama (2010: 123) to refer to the types of assessment other than formal test. Specifically, Brown and Hudson (1998, cited in Brown and Abeywickrama, 2010: 123) state that alternatives in assessments:

1. require students to perform, create, produce, or do something
2. use real-world contexts or simulations
3. are nonintrusive in that they extend the day-to-day classroom activities
4. allow students to be assessed on what they normally do in class every day
5. use tasks that represent meaningful instructional activities
6. focus on processes as well as products
7. tap into higher-level thinking and problem-solving skills
8. provide information about both the strengths and weaknesses of students

9. are multiculturally sensitive when properly administered
10. ensure that people, not machines, do the scoring, using human judgment
11. encourage open disclosure of standards and rating criteria
12. call on teachers to perform new instructional and assessment roles

Note: The term higher-level thinking in the quotation above refers to *analyzing, evaluating*, and *creating* in Bloom's taxonomy (as compared with lower-level thinking, which are *remembering, understanding*, and *applying*).

Another term, *authentic assessment*, which is proposed by O'Malley and Pierce (1996: 4), is defined as "the multiple forms of assessment that reflect student learning, achievement, motivation, and attitudes on instructionally-relevant classroom activities." Brown (2004: 28) adds that authentic assessment is characterized by the use of natural language, contextualized items, meaningful topics, thematic organization, and real-world tasks. In addition to authentic assessment, O'Malley and Pierce (1996: 5) also introduce the term *performance assessment*, which is characterized by the use of constructed response, higher-order/level thinking, authentic tasks, integrated language skills, process and product, and depth versus breadth. Burke (2009: 8) claims that the terms alternative assessment, authentic assessment, standards-based assessment, and performance-based assessment are synonymous. Introducing assessment for the elementary school level, Hill, Ruptic, and Norwick (1998) use the term *classroom-based assessment.* Other terms which can be included here are: *informal assessment* (as opposed to formal assessment), *process assessment* (as opposed to product assessment), *contextualized assessment*, and *task-based assessment*. These innovative assessments challenge teachers to create different methods of assessment.

## B. COMPUTER-BASED TESTING

The second current issue is the use of computer for language assessment. The advance of computer technology and the easy access to the information technology lead to the use of computer and internet in language teaching as well as language assessment. Online course, blended learning, and computer-assisted language learning (CALL) are examples of the use of computer and internet in language teaching. These computer-based teaching models are

automatically followed by computer-based testing. CBT (computer-based testing) and iBT (internet-based testing) in TOEFL, as well as CAT (computer-adaptive test) are examples of computer-based testing. It can be predicted that in the future, more and more computer-based teaching and testing will be developed.

Computer-based testing, of course, has advantages and disadvantages. The advantages of computer-based testing are offered by Douglas and Hegelheimer (2008, cited in Brown and Abeywickrama, 2010: 20), namely:

- a variety of easily administered classroom-based tests
- self-directed testing on various aspects of a language (vocabulary, grammar, discourse, one or all of the four skills, etc.)
- practice for upcoming high-stakes standardized tests
- some individualization, in the case of CATs
- large-scale standardized tests that can be administered easily to thousands of test-takers at many different stations, then scored electronically for rapid reporting of results
- improved (but imperfect) technology for automated essay evaluation and speech recognition

Note: High-stakes test is "an instrument that provides information on the basis of which significant decisions are made about test-takers" (Brown and Abeywickrama, 2010: 349).

The disadvantages of computer-based testing, as also mentioned by Douglas and Hegelheimer (2008, in Brown and Abeywickrama, 2010: 20), are as follows.

- Lack of security and the possibility of cheating are inherent in unsupervised computerized tests.
- Occasional "homegrown" quizzes that appear on unofficial Web sites may be mistaken for validated assessments.
- The multiple-choice format preferred for most computer-based tests contains the usual potential for flawed item design.
- Open-ended responses are less likely to appear due to (a) the expense and potential unreliability of human scoring or (b) the complexity of recognition software for automated scoring.
- The human interactive element (especially in oral production) is absent.
- Validation issues stemming from test-takers approaching tasks as test tasks rather than as real-world language use.

When the disadvantages can be overcome, surely computer-based testing will replace or, at least, become an alternative for the traditional paper-and-pencil tests.

## C. ASSESSMENT IN 2013 CURRICULUM

The third issue which needs to be taken into account in the area of assessment is the 2013 Curriculum in Indonesia. This current curriculum is characterized by the use of core competences and basic competences as the standards of syllabus contents (Kemdikbud, 2013: 53). There are four core competences which should be achieved in the teaching and learning activities. These four core competences are: spiritual competence (*Kompetensi Inti 1* = KI-1), social competence (KI-2), knowledge competence (KI-3), and skill competence (KI-4). Spiritual competence deals with the vertical relation between the students and their God, such as doing religion rituals or admiring God's creations. Social competence deals with horizontal relation among human beings, e.g. being cooperative, honest, discipline, etc. Knowledge competence, for English subject, deals with reading comprehension, grammar mastery, vocabulary mastery, etc. Skill competence deals with ability to make written report, perform oral presentation, carry out project task, etc. Each of these competences is elaborated in basic competences. From the basic competences, teachers determine the indicators or learning objectives, teaching contents, teaching and learning process, and then the assessment.

In 2013 Curriculum, teachers are required to assess three domains, namely, affective domain, which covers spiritual competence (KI-1) and social competence (KI-2); cognitive domain, which is knowledge competence (KI-3); and psychomotoric domain, which is skill competence (KI-4). Please keep in mind that skill competence here does NOT refer to the four language skills (listen, speak, read, and write), although they share the same term (i.e. skill) Teachers are also required to conduct assessment which includes the use of formal assessment and informal/alternative assessments. These are challenges for English teachers in Indonesia.

Other challenges in the field of assessment for English teachers are the possibility of the implementation of the concept of learning styles (i.e. auditory, visual, and kinesthetic styles) in the classroom, or the implementation of the concept of multiple intelligences (i.e. linguistic,

spatial, logical-mathematical, musical, kinesthetic, natural, intra-personal, inter-personal, and existential intelligences) for students. When these concepts can be implemented in language classrooms, we have to think of the kinds of assessment which suit the different learning styles or different intelligences of our students.

## EXERCISE 3

To check your understanding of this first activity, answer all the questions below.

1) When motivated the promotion of the term assessment?
2) Why alternative assessment is also called performance assessment?
3) Brown uses the term "alternative assessment", what terms are used by O'Malley and Pierce to refer to the same idea?
4) What is the difference between computer-based testing and internet-based testing?
5) Which of the four core competences belongs to Bloom's psychomotoric domain?

## SUMMARY

In this Learning Activity 3 we discuss current issues which affect assessment. They are alternative assessment, computer-based assessment, assessment in the 2013 Curriculum in Indonesia, and assessment for different learning styles and different intelligences. In this Learning Activity, teachers are just reminded that in the future they may need to create assessments which suit these developments.

## FORMATIVE TEST 3

Answer the following questions.

1) Computer-based testing leads to "individualization in testing". What does it mean?

2) What is meant by "real-world contexts" in the characteristic of alternative assessment?
3) Mention one possible solution to the "lack of security" in the use of computer-based testing.
4) In what domain, according to the 2013 Curriculum, does "the grammar learning" belong to?
5) What are meant by auditory style, visual style, and kinesthetic style?

If you have finished an exercise, look at the key answers at the end of the module. Evaluate your answers. When you get at least 80% right, you can go to another exercise, but if you don't, review the discussion and examples again. Then, do exercise once more. The following is how to evaluate your exercise and your test.

Formula:

$$\text{Level of mastery} = \frac{\text{The number of the reigh answer}}{\text{The number of the items}} \times 100\%$$

Level of mastery:
90 - 100% = very good
80 - 89% = good
70 - 79% = sufficient
< 70% = Insufficient

LEARNING ACTIVITY 4

# Principles of Language Assessment

Ideally, every time we make a test, the test should be good and effective. This Learning Activity provides a discussion of how to make a good and effective test, which include practicality, reliability, validity, authenticity, and positive Washback (Brown, 2004: 19-30; Brown & Abeywickrama, 2010: 25-39). These characteristics of test are discussed below. Read and understand them.

## A. PRACTICALITY

The first characteristic of an effective test is *practicality*. Practicality relates to the considerations of cost of a test, time allotment, test administration, human resource, test construction, and test scoring (Brown, 2004: 19). A good test should be relatively low in *cost*. It should be affordable by the students or test-takers. Requiring our senior high school students to take a TOEFL test (which costs more than $100 for each student) for a mid-semester test will be too expensive for the current financial condition of parents in Indonesia. A test which is prepared in a power point display for the whole class can be cheaper compared with the use of paper, but it may not be practical because it is difficult for the students who need to think longer or faster, or for the students who want to look back again at the previous items. Find a test which is low in cost, but does not sacrifice the quality of the test.

Another consideration is *time allotment* for doing the test. Approximately, between half an hour and two hours will be appropriate length of time. A test which should be finished by secondary school students in 5 hours will be too long. The students may get tired after the first two hours, and so they may hardly concentrate for the rest of the time. The loss of concentration will lead to an unreliable result of the test.

Next is the *administration* of the test. Make the test administration as simple as possible by, for example, using ordinary classroom and its available facilities. A test which requires students to do it in a special computer laboratory with internet facility, will not be practical if the facility

is not fully available at the school. When all the required facilities are available, of course we may construct any test method as we like.

*Human resources* are also important to be considered. The test administrators should have abilities to prepare and administer the test. If the test administrators or proctors do not have the skill to handle a test, or they need special training to administer, for instance, an internet-based test, it will not be practical. It is better to use simpler administration of the test. A test administrator who is a test constructor should have ability to construct a good test or to adopt/adapt from the available tests.

For the *test construction*, we may, for instance, use essay-type test, which is easy to construct but takes time to score, or use multiple-choice test, which takes longer time to construct but is easy to score. In this case, if the number of students or test-takers is big, e.g. 300 students, multiple-choice test will be more practical, because it takes a long time only in constructing the test but the scoring can be fast. If the number of students or test-takers is small, e.g. fewer than 50 students, essay-type test can be used. In this case, constructing essay-type test does not take a long time, and the scoring is still manageable.

The last consideration for practicality is *scoring*. Too subjective scoring will tend to have a problem of reliability. This problem is explained further in the following sub-topic of reliability. Related to a consideration of practicality in scoring is, for example, when a test should be scored by a special machine which is available far away from the test location and takes a long time to process the scoring. This will be not practical. In the same way, if for instance a test which is conducted in only 15 minutes for a student, but requires 5 raters to score, it is not practical. For practicality purpose, the number of raters for this example should be reduced.

## B. RELIABILITY

The second characteristic of a good test is reliability. Reliability means consistency, i.e. consistency in relation to students or test-takers, raters or scorers, test administration, and the test itself. There are several factors which affect assessment reliability.

To get reliable scores from the *test-takers*, we need to be sure that the test-takers are in good physical and mental conditions when taking the test. A test-taker who is unfit, fatigue, or in bad mood at the time of taking the test,

may not be able to concentrate, and therefore cannot show his/her best or real performance. In other words, the result of his/her test may not be reliable. The test-takers who are not familiar with the procedure of doing the test will not be able to reach optimal performance in the test either. This, in turn, makes the result of the test unreliable. Unreliable test results may also be shown when in a group of test-takers some of them are familiar with the test procedure that they can do the test faster and more easily, while others who are not familiar with the test procedure do the test in confusion and uncertainty.

The *raters* or *scorers* of a test should possess reliability. They should be consistent in scoring a test. There are two kinds of rater reliability, i.e. intra-rater reliability and inter-rater reliability. *Intra-rater reliability* means consistency within the rater/scorer himself/herself. If a writing test done by student A is scored 80 today, and a week later, the same rater still gives 80 (or a bit higher or lower) as he/she re-scores the same test for this student, it means that the rater is consistent or has intra-rater reliability. On the other hand, if student A is scored 80 today, and a week later he/she is scored 60 or 90 by the same rater, we say that the rater is not consistent; in other words, the rater does not have or has low intra-rater reliability.

The problem of reliability in a rater is not found in scoring objective-type tests (e.g. true-false, matching, or multiple choice), because there is a clear correct/wrong answer, which is provided in the answer key. The scoring is done just by counting the number of the correct answers. The problem of reliability may occur in scoring subjective-type test (e.g. essay writing or speaking assessment) in which the scoring relies on the rater's subjective considerations. The results of scoring may tend to be unreliable if the rater is tired, should score a large number of test papers, work for a long time without breaks, or does the scoring with no rubric or guide to the correct answer. To avoid unreliability, when doing the scoring a rater should be fit, in good health, and in a comfortable place. If the test papers are too many, there needs to have more raters. There should also be periodic breaks in scoring time. If the test is in the form of essay questions, there should be answer key or guide to the correct answer. If a rater has to score a long essay or a speaking performance, he/she needs to have a rubric as a scoring guide. An analytic scoring rubric will be better than a holistic scoring rubric, because in analytic scoring rubric there are detailed points and their descriptions to guide the scoring.

To be reliable, a rater needs to train himself/herself in using the scoring rubric. In the training, he/she can use benchmarks, which are samples of standard qualities and suggested scores for the test-takers written response. The rater trains himself/herself to score a number of test-takers' works and compare them with the benchmarks. The closer the scores to the benchmarks, the more reliable the rater.

Another way to train the rater's reliability is by comparing two sets of scores made by the same rater. For example, using a rubric a rater scores 50 essays, and several days later he/she re-scores the same 50 essays, by trying to not remember the previous scores. Then, the two sets of scores are compared using correlation formula, e.g. product-moment statistic formula. The result of statistic calculation may show that the two sets of scores are highly correlated, moderately correlated, or lowly correlated. When the two sets of scores are highly correlated, it means that the rater has a high intra-rater reliability. When they are lowly correlated, it means that the rater has low intra-rater reliability. In this latter case, the rater's consistency is low. The rater needs to train himself/herself again until he/she gets high correlation.

Besides intra-rater reliability, another kind of rater reliability is *inter-rater reliability*, which means consistency between two or more raters. Inter-rater reliability is needed when two raters score different sets of essays independently. This may occur when there are 100 essays, and one rater is not able to score alone. This rater needs another rater to help scoring the essays. The 100 essays are divided into two; rater A gets essays 1 – 50, and rater B gets essays 51 – 100. Then, both raters score the essays independently. We hope that these two raters have the same perception on the quality of the essays they score. We do not want to find one rater is more lenient in giving scores and the other is stricter. In this case, the two raters need to make themselves have the same perception, or have inter-rater reliability.

There are two ways to attain the same perception or inter-rater reliability. First, the two raters score the same essay. After that, they compare and discuss the scores they have given to the essay, especially the scores which are different. They analyze whether any of the raters is too lenient or too strict in giving the scores. It is hoped that the two raters reach the same perception and decide to agree with a certain score. If this exercise is practiced repeatedly, the two raters will have inter-rater reliability. The

second way is that the two raters score the same 50 essays independently. Then, the set of 50 scores made by rater A is compared with the set of scores made by rater B, using correlation formula, like in statistic calculation for intra-rater reliability above. If the result shows that the two sets of scores have high correlation, it means that the two raters have high inter-rater reliability. In other words, the two raters have similar perception; therefore, they may score the essays assigned to each of them independently. Please note, however, that this second way is not the same as the case in which two raters score the same sets of essays, and then they combine the pairs of scores and then the scores are divided by two. For example, student X gets 70 from rater A, and 80 from rater B. Then, the score for student X is 70 + 80, divided by 2, which is 75. This can be fair but not effective because both raters have to score all and the same essays.

The next consideration for reliability is the *test administration.* A test administration is reliable if the procedure of administration is in accordance with what has been designed. A listening test which is conducted using sounds from a tape-recorder is reliable in administration if the same quality of sounds can be heard equally by all the test-takers. However, if the sounds from the recording can be heard clearly by some test-takers and not clearly by some other test-takers, the administration of the listening test is not reliable. Another example of unreliable test administration is when two groups of test-takers do the test in different places. One group do the test in a classroom with good chairs and tables, and another group do the test in an auditorium which is provided with chairs only and the test-takers uncomfortably have to use cartons to write. In this example, the test administration is not reliable because the two groups are not treated in equal comforts. Other things which can affect administration reliability are noise, time limit, seat condition, room temperature, quality of copied test papers, proctor's behaviors, etc.

The last consideration of reliability is related to the *test* itself. Unreliable scores can be due to bad quality of test, such as: unclear instruction, ambiguous answer, bad item construction, or clues to the correct/wrong answer. Unclear instruction can be found when in matching items the instruction does not state whether each of the responses (in the right column) can be used only once or more than once. Ambiguous answer occurs when in multiple choice items there are two or more correct choices. An example of bad item construction is an open-ended question asking about test-taker's

opinion, in which any answer can be correct. Clue to the correct answer is found when the longest option in multiple choice item is the correct answer. Besides all of these, of course, very unreliable results of a test can be caused by cheating done by the test-takers while doing the test, knowing the answers to the test beforehand, or unfair practice of the proctors, who inform the answers of the test to the test-takers.

In addition, there are still other ways to measure a test reliability, i.e. test-retest reliability, equivalent forms reliability, split-half reliability, Cronbach alpha reliability, and Kuder-Richardson reliability (Djiwandono, 2008: 171-185). *Test-retest reliability* is obtained by repeating the same test to the same students. In this case, we make a set of test, then administer it to a number of students, and record the scores. After several days, we administer again the same test to the same students, and record the scores. Then, the scores from the first test and the scores from the second test are correlated using Product-moment correlation formula. If the result of the statistic calculation shows high correlation, it means that the test set that we made is reliable. In using test-retest technique, it should be noted that the time between the two administrations of the test should not be too short, that the students still remember their answers in their first test, nor too long, that the students get improvement in their language mastery.

*Equivalent forms reliability* is obtained when we make two equal sets of test, i.e. having the same purpose, objectives, scope, type of test, and number of items. Then, the two sets of test are administered to the students, and the scores are correlated like in the test-retest procedure mentioned above. If the result is correlated, it means that the two sets of test are reliable.

*Split-half reliability* can be measured when we make a set of test and administer it to a number of students, then the scores are separated, i.e. a set of scores from the odd numbers of items and another set of scores from the even numbers of items. The two sets of scores are correlated again like the above procedure. If the result shows that there is a correlation, it means that the set of test is reliable. This split-half technique is based on the assumption that the test items in the set of test have gradual difficulties; therefore, the pairs of odd and even numbers are equal in difficulty levels.

*Cronbach-alpha reliability* is measured like in split-half procedure, but instead of using product-moment formula it uses Cronbach-alpha formula. There is another variant of Cronbach-alpha formula, i.e. a formula which is used to measure reliability of the scores of essay test.

The last is *Kuder-Richardson (K-R) reliability*. This K-R reliability requires one administration of a test. The answers made by the students doing the test are scored in dichotomy, i.e. a correct answer is scored 1 and a wrong answer is scored 0, then the scores are calculated using K-R formula. There are two versions of K-R formula, one is the K-R20 formula and another one is K-R21 formula which is a simpler formula used for teacher-made tests.

## C. VALIDITY

The third principle of a good and effective test is *validity*. Validity is usually defined as a test or assessment which is used to measure what is supposed to be measured. This section discusses some aspects related to validity, i.e. content-related validity, criterion-related validity, construct-related validity, consequential validity, and face validity (Brown & Abeywickrama, 2010: 29-36). They are elaborated below.

*Content-related validity* refers to the validity of the content of a test in relation to its objective. For example, in the teaching-learning process we teach Language Assessment using Heaton's (1988) book about assessing language skills, but for the summative test we use the test materials from O'Malley and Pierce (1996), which is about authentic assessment, then our test is not valid. When we teach narrative texts to our students, and then the test materials are in the form of argumentative texts, our test is not valid. However, if we teach a legend of *Malin Kundang* to our students, and the test uses a legend of *Tangkuban Prahu*, our test is still valid, because both legends belong to the same narrative type texts.

Sometimes, unconsciously we make mistakes in content validity. For example, we want to make a vocabulary test with the following item.

1. *You have to wash your hands with ....*
   - *a. soup*
   - *b. soap*
   - *c. shop*
   - *d. sop*

This test item looks like a vocabulary test, but in fact it is a spelling test, because the test-takers are just required to recognize the correct spelling of word *soap*. The correct vocabulary test item should be as follows.

*2. You have to wash your hands with ....*
   *a. sand*
   *b. soap*
   *c. mud*
   *d. grass*

In this item the test-taker has to choose a word whose meaning is suitable with the context stated in the stem; therefore, this item is valid as a vocabulary test.

Related to content validity we have to know two other terms, i.e. direct test and indirect test. *Direct test* is when we test directly what to be tested. For example, if we want to know whether a test-taker knows exactly the position of primary stress in the word *develop*, we have to ask the test-taker to pronounce the word and check whether he/she puts the stress correctly (i.e. on the second syllable) or not. However, sometimes it is difficult or not practical to use direct test due to the time limit or large number of test-takers. In this case, we can use *indirect test*. With the above example, we may make a written test by writing *de-vel-op* (in separate syllables) and ask the test-taker to determine whether the stress is on the first, second, or third syllable. This is called indirect test. Surely, the best test is the direct test. Indirect test has one weakness, i.e. in the above example, the test-taker may know that the stress is on the second syllable, but when he/she really pronounces the word, it can happen that he/she pronounces it unconsciously with the stress on the first syllable.

*Criterion-related validity* deals with whether a test reaches certain criteria. Criterion-related validity has two kinds, namely, concurrent validity and predictive validity. Our test has a *concurrent validity* if its results are supported by other valid tests. For example, in our knowledge TOEFL test is a valid proficiency test. We make another set of proficiency test, and then it is administered to our students, who have taken a TOEFL test. The result of our test is compared with the result of the TOEFL test, using correlation (e.g. product-moment) statistic formula. If there is a high correlation between the two tests, it means that the test that we make has concurrent validity (with TOEFL test).

A test has a *predictive validity* if it can predict the success of test-takers in the future (see prognosis purpose of a test as explained in Activity 1in this module). For example, we have a program to train teachers at S-2 level, and

so we make a test with the purpose to know whether the participants will be successful or not in their study at S-2 level. The test is administered at the beginning of S-2 program. By the end of S-2 program we score the success of the participants. These scores are compared with the scores of the test that we made and administered at the beginning of the program. If the result of the comparison shows that there is a correlation between the two scores, i.e. the participant who gets good score from the test at the beginning of the program also gets good score for his/her success, or the other way around, then we can conclude that the test at the beginning of the program has predictive validity. When a test has predictive validity, we can say that the higher the result of the test the higher the possibility to succeed in the program.

Next is *construct validity*, which means that a test should be valid by its construct. Construct refers to theory, hypothesis, or model of something (Brown & Abeywickrama, 2010: 33). Reading test is valid if it matches with reading construct, and speaking test needs to be valid by its construct. Now what is reading construct and what is speaking construct? As we know, the purpose of reading test is for comprehension; thus, reading comprehension should cover comprehension of main idea, explicitly stated information, implied information, vocabulary meaning, and cohesive devices. These elements of comprehension are the construct of reading. When a reading test has included all these elements, we can say that the reading test is valid by construct. The same thing happens with speaking. The purpose of speaking test is to measure the productive oral mastery, which is the construct of speaking. This construct of speaking includes fluency, pronunciation, content, organization, grammar, and diction. When a speaking test measures all these, we can say that the test is valid by construct. This also means that when we test speaking, and the focus is only on the length of speech, it can be said that the test lacks construct validity.

*Consequential validity* refers to the impact of a test to the test-takers. When we determine that the final exam, for example, should be conducted through internet, the consequence is that the test-takers should be prepared to be able to use internet-based test. Otherwise, our test will not be valid because the test-takers may be troubled by the inability to use internet. The problem of consequential validity may also occur when we use certain type of test, and some of the test-takers who can afford to pay for the coaching of the test will do the test better than those who do not get the coaching. In this case, the test has a problem of consequential validity, since it is not fair for all test-takers. This happens in Indonesia in facing the national examination,

where a number of financially more able students attend learning guidance in private learning institutions.

The last kind of validity is *face validity*, which concerns the appearance of the test. We may think that a written test does not look suitable for testing speaking, or multiple choice grammar test seems unsuitable as a test for writing. In these two examples, i.e. written test for speaking and grammar test for writing, the tests lack face validity. The lacks of the tests lie in the incomplete constructs of speaking and writing. The correct face validity is when speaking is tested through speaking and writing is tested through writing.

## D. AUTHENTICITY

Authenticity can mean the degree of closeness of the test tasks to the real-life tasks in the target language (Bachman & Palmer, 1996: 23). Regarding the features, Brown and Abeywickrama (2010: 37) mention that authentic assessment:

1. contains language that is as natural as possible
2. has items that are contextualized rather than isolated
3. includes meaningful, relevant, interesting topics
4. provides some thematic organization to items, such as through a story line or episode
5. offers tasks that replicate real-world tasks

An example of *natural language* in an oral interaction can be seen in the following dialog.

*A. What's your name?*
*B. Sintha*
*C. Where are you from?*
*D. Malang*

In such a dialog, sometimes a teacher requires his/her student to answer the above questions using complete sentences, such as in the following.

*A. What's your name?*
*B. My name is Sintha*
*C. Where are you from?*
*D. I am from Malang*

The complete answers made by B in this example do not reflect the natural English as used by the native speaker. For authentic assessment we have to use natural English, as required in CLT method.

An example of *contextual tes*t is when we test vocabulary. Rather than asking:

- Write the meaning of "trivial"

it is better to have the following item:

- *"The students think that the test is difficult, but the teacher regards it as trivial." The underlined word means ....*

The first item uses isolated word, but the second item is contextualized, and helps the test-takers find the answer. Another example of contextual test item is when a teacher asks:

- *Muthia, if you meet your teacher in a super-market at 7 p.m., how would you greet him/her?*

A test item should be *meaningful*. An example of meaningful item is as follows:

Teacher : *Budi, a friend of yours, Amelia, held a birthday party last week. Actually, you wanted to come, but she did not invite you. What would you say when you meet her?*

The expected answer for this item is: *"If you had invited me, I would have come."*

The following example is not meaningful.

*Teacher : Repeat after me. "Dian and Renza study English."*
*Students : Dian and Renza study English.*
*Teacher : Change to "past".*
*Students : Dian and Renza studied English.*
*Teacher : Change to "continuous".*
*Students : Dian and Renza are studying English.*

In this example, the drill is not meaningful, because even though the students can use past form and continuous form correctly, they may not know how and when to use the sentence forms. A teacher often focuses on form rather than on meaning; therefore, this drill is not meaningful.

The next feature of authenticity is *thematic* organization of assessment. Rather than using unconnected sentences to assess the use of tenses, it is better to use a passage or story, which can provide context for the use of certain tense form.

The last feature of authenticity is that a test task should imitate *real-life* task. A dicto-comp is an example of real-life task, because in dicto-comp students are asked to hear something told by the teacher while the students take notes, and then they rewrite what has been told to them. This practice imitates the real-life, where a secretary takes note the instruction told to him/her, and then rewrite what is expected from the instruction. Another example of real-life feature of authentic assessment is a reading test whose text is selected from current issues adopted or adapted from newspapers, magazines, brochures, etc.

## E. WASHBACK

The fifth or last principle of a good and effective test is *washback* or *backwash*. Washback can be defined as the effect of test or assessment on teaching, learning, learner, or government and society. Washback can be positive or negative. For example, since there is a writing test in the national examination, teachers who were previously reluctant to teach writing, then they teach writing. Knowing that the test is always challenging to the students, then the students are motivated to learn and make better preparation for the test. These are examples of positive washback. However, when teachers know that the national examination always uses multiple choice test items, then in the teaching and learning activities the teachers drill their students on how to do multiple choice test, forgetting teaching students the process of learning, this is an example of negative washback. Or, knowing multiple choice examination, students are busy preparing the effective strategy for cheating. This is the worst negative washback.

Washback is different from feedback. *Feedback* is figures, letters, comments or suggestions given to students' works so that the students know the quality of their works. However, good feedbacks can become positive washback. For example, when returning a student's work on writing, the teacher writes: *"I like your writing. The content shows that you know much about the topic. The only thing you need to improve is spelling."* When this feedback encourages the student to improve his/her spelling mastery, then the teacher's feedback has positive washback to the student.

## EXERCISE 4

To check your understanding of this first activity, answer all the questions below.

1) In his/her own class, is it allowed for a teacher to answer student's question about the test? Explain.
2) What is analytic scoring?
3) What is the problem when writing test is made in the form of rearranging sentences using multiple choice format?
4) Why is intra-rater reliability important?
5) What is the difference between washback and feedback?

## SUMMARY

In this Learning Activity, we learn that a good and effective assessment should be characterized by practicality, reliability, validity, authenticity, and positive washback. Practicality deals with cost of the test, time limit in doing the test, ease of administration, human resources, test construction, and ease of scoring. Reliability refers to student factor, intra- and inter-rater reliability, test administration reliability, and reliability of the test itself. Validity can be in the form of content validity, criterion-related validity (concurrent validity and predictive validity), construct validity, consequential validity, and face validity. Authenticity is characterized by the use of natural language, contextualized test items, meaningful topic, thematic, and real-world tasks. Lastly, washback means the impact of assessment on teaching, learning, learner, or government and society. It may be difficult to meet all these characteristics, but it is suggested that a test should consider as many of the characteristics as possible.

## FORMATIVE TEST 4

Answer the following questions.

1) Why is the loss of concentration by the test-takers in doing a test leads to unreliable test results?

2) In a university entrance test, very often a proctor is prohibited to answer any question from the test-takers. Why?
3) Give an example of proctor's behavior in a test room which causes unreliability.
4) Give a reason why mechanical/substitution drill is considered not meaningful.
5) What kind of washback may happen when teachers let their students cheat in the final examination?

**Note:**
For further reading about the contents of this module, you are recommended to read:

- Brown (2004)
- Brown and Abeywickrama (2010)
- O'Malley and Pierce (1996)
- Weir (1990)

See the details of these sources in the list of references at the end of this module.

If you have finished an exercise, look at the key answers at the end of the module. Evaluate your answers. When you get at least 80% right, you can go to another exercise, but if you don't, review the discussion and examples again. Then, do exercise once more. The following is how to evaluate your exercise and your test.

Formula:

$$\text{Level of mastery} = \frac{\text{The number of the reigh answer}}{\text{The number of the items}} \times 100\%$$

Level of mastery:
90 - 100% = very good
80 - 89% = good
70 - 79% = sufficient
< 70% = Insufficient

# Key to Answers

Below are keys to the exercises and summative tests. Very important in the keys are the key ideas; therefore; your answers may use different wording. In some questions you may have different answers. If you are not sure of your answers, you may contact the tutors/instructors in the Open University.

**Exercise 1**

1) Assessment is part of teaching.
2) Assessment is descriptive and evaluation is judgmental.
3) Summative assessment is an assessment conducted at the end of a course or program, to measure the students' attainment.
4) Selection assessment is conducted before a program begins, to select candidates to fulfill a quota; and screening assessment is conducted at the end of a program, to select the test-takers who pass the MPL.
5) Formative assessment is regarded as achievement assessment because it measures the progress of students in attaining the learning objective(s).

**Exercise 2**

1) The strategy is Stimulus-Response-Reinforcement.
2) Unitary trait hypothesis holds that mastery of language elements and language skills correlates.
3) It is sociolinguistic competence.
4) It is not authentic because in supermarket there is no bargain.
5) Telling or retelling a story is a common practice in real-life situations.

**Exercise 3**

1) It was promoted by the unsatisfactory use of conventional test.
2) Because the assessment is based on what is acted or performed by the test-takers.
3) O'Malley and Pierce use the terms authentic assessment and performance assessment.
4) Computer-based testing uses a computer program, and internet-based testing uses internet web.
5) KI-4 or skill competence.

**Exercise 4**

1) Yes, if the student asks for clarification of the test instruction or the meaning of certain test items.
2) Analytic scoring is the scoring using detailed points and their descriptions to guide the rater in scoring.
3) It has a problem with face validity; writing test should show the test-taker's ability to produce a piece of writing.
4) Because the scoring made by a rater should be objective and fair.
5) Feedback is a letter, figure, comment, or suggestion given to the test-takers work to show its quality, while washback is the impact of a test to the test-taker, teacher, teaching and learning process, etc.

**Formative Test 1**

1) Advantages: TOEFL is readily available, easy to score, can be used for a large number of students. Disadvantages: expensive, may not match with the teaching materials.
2) For example, Student A has a problem in pronouncing some consonant clusters, e.g. /-gz/, /-bd/, and /-pt/.
3) The strength of homogeneous class is that it is easy for a teacher to teach because the students' ability is relatively the same; but its weakness is that the students in the low-ability will feel inferior or demotivated in learning. The strength of heterogeneous class is that the low-ability students can learn from the high-ability students; but its weakness is that it is difficult for a teacher to teach the students with widely varied abilities.
4) The school should require the researcher to give a copy of the research report to the school, and the teacher in the school should use the research findings for improving his/her students.
5) It is an assessment which is non-test. It describes the student's ability in speaking. It is not an evaluation, because it does not give judgment to the student.

**Formative Test 2**

1) Because it emphasizes the teaching of structure or grammar.
2) For example, English has various verb forms to show time of occurrence, whereas Indonesian does not have (or, there can be other answers)

3) Because dictation involves listening ability, spelling, vocabulary recognition, comprehension, and expectancy grammar.
4) Linguistic competence does not show the test-taker's ability to demonstrate his/her ability to use what to say, to whom, when and where.
5) Just for practicality purpose, i.e. limited time and simpler test administration.

**Formative Test 3**

1) It can be done in the test-taker's own time and place.
2) Real-world contexts mean the imitation of activities in real-life.
3) For example, the test can be opened only by an authorized test administrator. (There can be other possible answers).
4) It belongs to cognitive domain.
5) Auditory style means the learning style through listening, visual style means the learning style through seeing, and kinesthetic style is the learning style through doing.

**Formative Test 4**

1) Because the results of the test may not show the real abilities of the test-takers
2) The reason is that in the university entrance test the proctor is prevented to give wrong answer, even if it is only a clarification question.
3) For example, a proctor stands beside a test-taker, which makes the test-taker not able to concentrate. (There can be other possible answers.)
4) Because most of mechanical/substitution drills focus on form/pattern, and not on meaning or use.
5) There can be a negative washback, in which the students will not prepare for the test seriously, knowing that they can cheat.

# References

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

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

Brown, H.D. 2004. *Language Assessment: Principles and Classroom Practices.* White Plains, NY: Pearson Education.

Brown, H.D. and Abeywickrama, P. 2010. *Language Assessment: Principles and Classroom Practices (2nd edition).* White Plains, NY: Pearson Education.

Burke, K. 2009. *How to Assess Authentic Learning (5th edition).* Thousand Oaks, CA: Corwin.

Heaton, J.B. 1988. *Writing English Language Tests (New Edition)*. London: Longman.

Hill, B.C., Ruptic, C., and Norwick, L. 1998. *Classroom Based Assessment*. Norwood, MA: Christopher-Gordon.

Hughes, A. 1989. *Testing for language teachers*. Cambridge: Cambridge University Press.

Kemdikbud. 2013. *Peraturan Menteri Pendidikan dan Kebudayaan R.I. No. 81A tentang Implementasi Kurikulum. Lampiran IV: Pedoman Umum Pembelajaran*. Jakarta: Kemdikbud.

Oller, J.W. 1979. *Language Tests at School: A pragmatic Approach*. London: Longman.

O'Malley, J.M. and Pierce, L.V. 1996. *Authentic Assessment for English Language Learners: Practical Approaches for Teachers*. White Plains, NY: Addison Wesley.

Popham, W.J. 1978. *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Richards, J.C. and Rodgers, T.S. 2001. *Approaches and Methods in Language Teaching (2nd edition)*. Cambridge: Cambridge University Press.

Saukah, A. 2013. *Penilaian Pembelajaran Bahasa*. Malang: UM Press.

Weir, C.J. 1990. *Communicative Language Testing*. New York: Prentice Hall.