MODULE 8

Scoring Interpretation

Dr. Sintha Tresnadewi, M.Pd.
Fachrurrazy, M.A., Ph.D.

INTRODUCTION

This is the eighth module on Assessment in Language Teaching, dealing with scoring interpretation of language tests. This introduction presents the objective, scope, and learning guide for this module. Read them carefully, because they will be useful to help you understand the contents of this module.

OBJECTIVES

By learning this module, you are expected to have knowledge and skill on how to:

1. mark and grade student scores, and
2. interpret student scores.

SCOPE

The topics to be discussed in this module include:

1. the types of scoring and how to determine grading, and
2. how to make interpretation of student scoring and grading.

LEARNING GUIDE

To get a good mastery in learning the contents of this module, you are advised to do the following steps.

1. Read the introduction of this module, so that you know what is expected to be mastered in general.
2. Read and follow the activities one by one in details, and do all the exercises and formative tests.

3. Finishing the tasks, check your answers or responses to the exercises and summative tests with the answer keys at the end of this module. It is advised that you do the tasks first before looking at the keys.
4. If your answers or responses are wrong, learn again the related activity. Find why they are wrong, and make corrections.

MAIN ACTIVITY

The main part of this module consists of two Learning Activity, beginning from the discussion of scoring and grading, and continued to the interpretation of scoring and grading. They are in a continuum, which are inseparable. The main sources for the contents of these two activities are Brown (2005), Brown and Abeywickrama (2010), and Gronlund and Waugh (2009). The examples are adopted, adapted, or created to suit the context of Indonesia.

LEARNING ACTIVITY 1

Scoring and Grading

The topics to be discussed in this activity are scoring and grading. These two topics are presented in one Learning Activity, because both are closely related. The students are expected to be able to master these two topics well. Read them carefully and do the exercise and the summative test that follow.

Scoring or marking is a process of determining the correct and incorrect answers made by students, and are then accumulated to produce a total score. Grading is a process of classifying or grouping the scores for assigning values. The following are the detailed explanations.

A. MARKING/SCORING

There are three further steps that we have to do after we administer a test. The first is *marking* or *scoring* the papers, the second is *grading* or converting the scores to grades, and the third is providing information to the test takers what the grades mean, or telling them how to *interpret* the grades they get. The third (i.e. interpretation) will be presented in the second Learning Activity.

Now let us talk about the first and the second steps.

Marking or *scoring* is the process of awarding a number or a symbol to represent the level of student learning achievement. The most common method is by adding up the number of correct answers on a test, and assigning a number that correlates. Generally, a "score" is a numeric value based on the possible points on a test. If it is out of 100, a typical "score" may be 95, 80, 65, 50, etc.

From the marking point of view, tests fall into one of two categories: *objective* or *subjective*.

The *objective* test has only one correct answer, but the *subjective* test may result in a range of possible answers, some of which are more acceptable than others. Actually, it is not really the tests which are objective or subjective, but the systems by which they are marked.

Let us now look at the difference between *objective marking* and *subjective marking*

1. Objective Marking

Objective marking is possible with multiple choice, binary choice, exact-answer cloze, or matching formats. In all these tests, a list of the keys gives the only correct answers. Thus, the actual marking is easy, that is, the correct answers are simply added up into a score. For example, a test of, say, 40 items, is given to the testees and marked according to the number of correct responses made, usually on the basis of one mark for each correct response, or when using 'weight', certain test items may be marked higher than the others. The total number of such responses is the person's "score". A large number of items in this category can be administered to a group of students in a relatively short time, and the results can be quickly scored by hand or by machine.

Objective marking is also possible for discrete-item test, in which items are independent of each other so that their order can be changed, or one of them can be omitted, without influencing performance on the other items.

2. Subjective Marking

When the tests are integrated by topic, theme, or task, the items are no longer discrete but they form a 'set' in which performance on one item may influence or depend on appropriate performance on another. Besides, these kinds of tests provide more freedom of response, and this introduces greater subjectivity into the scoring. Different scorers may arrive at different scores as they weight elements of the answers differently. In such a case, we deal with subjective marking.

With this kind of test, a teacher has to set a *marking scheme* or *scoring rubric* (see Figure 8.1 below for an example of a marking scheme) at an early

stage in the development of the test. It is even better for a teacher to get students involved in the development of the marking scheme. It provides an opportunity for students to be a part of the thinking process around judging performance and to deepen their understanding of what is required. Once a marking scheme is agreed upon, it affirms that the students do indeed know the criteria and that the teacher knows they know. By doing this, the subjectivity of the judgment can be reduced.

9	Expert user	Has fully operational command of the language: appropriate, accurate and fluent with complete understanding.
8	Very good user	Has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriate words. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.
7	Good user	Has operational command of the language, though with occasional inaccuracies, inappropriate words and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning.
6	Competent user	Has generally effective command of the language despite some inaccuracies, inappropriate words and misunderstandings. Can use and understand fairly complex language particularly in familiar situations.
5	Modest user	Has partial command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field.
4	Limited user	Basic competence is limited to familiar situations. Has frequent problems in understanding and expression. Is not able to use complex language.
3	Extremely limited user	Conveys and understands only general meaning in very familiar situations. Frequent breakdowns in communication occur.
2	Intermittent user	No real communication is possible except for the most basic information using isolated words or short formulae in familiar situations and to meet immediate needs. Has great difficulty understanding spoken and written English.
1	Non user	Essentially has no ability to use the language beyond possibly a few isolated words.
0	Did not attempt the test	No assessable information provided.

Figure 8.1
IELTS Marking Scheme
(Source: IELTS Guide for Teachers, www.ielts.org)

B. GRADING

Grading involves quantifying data and assigning value. Grade serves two purposes: 1) notifying students of their achievements, and 2) informing the public of student's performance. Grades assigned to the student's work should represent the extent to which the instructional objectives or the intended learning outcomes have been achieved, and should be in harmony with the grading policies of the school. Some schools have both clearly defined objectives and grading policies; many schools neither.

Assessment of learning during instruction might include the use of objective and essay tests, ratings, papers, and various types of performance assessment. The problem of grading is that of summarizing this diverse collection of information into a single letter grade or brief report. The single grade letter (e.g. A, B, C, D, F) is the most widely used grading system (Gronlund and Waugh, 2009: 192).

1. Methods of Grading

The "scores" can be translated into "grades" through certain conversion, such as, a straight scale (90-100 = A, 80-89 = B, etc). In the case of an individual test, 100-point system might represent the percentage of items correct or the total number of points earned on the test. When used as a final grade, it typically represents a combining of scores from various tests and other assessment results. In the box below, it can be seen a comparison between grades and marks/scores.

Grades and marks/scores compared

- A grade and a mark/score are basically the same thing, or, one is not better than the other. For example, getting a grade of B on a report card will mean that a student received a mark/score between 80 to 90 percent.
- Marks/Scores help students identify their exact attainment whereas grades place students in predefined categories such as A and B.
- Marking system may lead to the perception that high marks/scores, rather than real knowledge, are important. If knowledge enhancement and overall development of students is the prime goal, the grading system is preferable.

C. BASES OF GRADING

There are various ways to determine what to include in the students' grades. A teacher may use one or a combination of two or more of the following (Brown and Abeywickrama, 2010: 320-321).

First, students' grades are determined based on only the results of **formal tests**, such as, quiz, formative test, summative test, or any other scored tasks. As reported in Brown and Abeywickrama, a majority of teachers at the American Language Institute at San Francisco State University agreed that students' grades should be based on formal tests. Gronlund and Waugh (2009: 205) strictly state that "Grades should represent achievement of the intended learning outcomes and be uncontaminated by other factors". This is true when grading is viewed from pure achievement. In practice, a teacher may think of combining it with other factors such as the following, to make it more comprehensive scores.

Second, students' grades are determined based on the intuitive, informal **observation** of the students' language performance. This type of grading is also acceptable for determining students' grades. However, it should be clear about the elements to be observed and how to score them, so that the scores obtained are objective and reliable. In addition, various **alternative**

assessments (see Module 7) are currently accepted as a part for deciding the students' final grades. They can be used to make the grading more valid and reliable.

Third, students' grades are determined based on their oral **participation** in class. If this is taken into account for grading, the students should be informed about it at the beginning of the course/semester. In addition, the method of scoring participation should be determined. A lecturer (name with-held) at the Graduate Program of the State University of Malang (UM) ran his course in the form of seminars. In every group presentation, he took notes the participations of students which are in the forms of asking questions, sharing ideas, helping to answer audience's questions, giving comments, or providing further information. The lecturer counted the participations of every student at the end of the semester, and then converted them into scores to be combined with the scores of the formal tests.

Fourth, students' grades are determined based on their **improvement** during the course period. Improvement or progress in the course can be rewarded and included in students' scores. However, it is suggested that the portion should not be too big, as it is not directly related to the achievement.

Fifth, students' grades are determined based on their **behaviors** in class, e.g. polite, cooperative, ignorant, inattentive, etc. This factor is not relevant to the students' achievement. If this should be considered, it is advised that the students' behaviors are put in a separate notes.

Sixth, students' grades are determined based on whether there is an **effort** to learn. This is the same as the fourth factor above. It should be just a small portion.

Seventh, students' grades are determined based on whether there is **motivation** or not. This is also similar with the fourth and the sixth factors above.

Eighth, students' grades are determined based on their **attendance and punctuality**. This should become a pre-requisite for attending the course. In some universities in Indonesia, the students' attendance should be at least 80%. Attending less than that percentage, a student should be dropped from the course.

Ninth, as an addition, there are possibilities that the students' grades are determined based on the students' kindness (e.g. providing meals or foods for the teacher), or based on the instruction from the school's authority. A colleague (name with-held) told one of the writers of this module that when

he studied at a graduate program, he took a tough course which he felt very hard. The lecturer was very busy everyday and the class was run in the evening. The colleague told that every time before he attended the lecture, he brought some meal as dinner for himself and for his lecturer, and they had dinner together. In his perception, he passed the course because of the meal. A different story was told by a teacher that he was instructed by his school authority to mark-up the students' semester's grades. That is because the semester's grades were used as an element in the formula for determining the passing grade in the National Examination. The authority did not want many students in the region to fail in the exams. This is an intervention and wrong practice, of course, but it happened.

D. GUIDELINES FOR SELECTING GRADING CRITERIA

Knowing the various types of grading as discussed above, our question is: What are the criteria for our grading, and how should we select them? Brown and Abeywickrama (2010: 322) suggest the following guidelines.

First, the criteria for grading should be in accordance with the policy of the institution. If a policy in a school is, for instance, to include certain percentage of attendance as a requirement for taking the final test, then a teacher has to include this policy in his/her course, and has to inform it to the students. In this way, it can be avoided a wide variety of policies in the school.

Second, all the components which will be used to determine the final grades should be informed to the students at the beginning of the course/semester, so that they can prepare themselves to expect certain grades. For example, the final grade for the course will be taken from the results of mid-term test (40%), final test (40%), and presentation and participation (20%). The percentages are determined by the teacher, by considering the importance of certain elements. A teacher may, for instance, determine that the final grades of his/her students will be taken 100% from the results of the final/summative test, but indeed this is risky.

Third, if the grading includes points 4 – 8 (improvement, behavior, effort, motivation, and attendance) above, they should be prepared and administered properly. The teacher may prepare a checklist, a chart, a note-taking, or an anecdote (= note of specific incident), and convert them to scores.

Fourth, it should be noted that the portions for points 4 – 8 above should be small, so that the grades still reflect learning achievement. It is not allowed to give a high grade to a student just because the teacher is very much impressed with his/her politeness.

E. GUIDELINES FOR EFFECTIVE AND FAIR GRADING

A good grading should be effective and fair. Gronlund and Waugh (2009: 201-202) propose guidelines to make grading fair and effective. They are as follows.

First, after we have decided to use certain grading procedure, we have to inform it to the students at the beginning of the course/semester (as also stated above), so that the students know what are expected of them in the course. Preferrably, the information is made in writing in the form of Course Outlines. As stated above, it should include the components and the weighting for each component.

Second, Gronlund and Waugh, who are the proponents of achievement assessment, strongly suggest that grading should be based on achievement only. In this case, it can be interpreted that as far as the components of grading are still directly related to learning achievement, they can be included for determining students' grades. For example, in addition to the formal tests, the results of certain kinds of alternative assessment (portfolio, journal, project report, observation, etc.) can be included as components of grading. However, if, by certain considerations, we have to include components which are not directly related to the achievement purpose (such as effort or motivation) we have to give them small portions only.

Third, in relation to the second point above, it is suggested that grading be based on a variety of valid data, such as those from alternative assessments. Some alternative assessments may not be strongly reliable, but eliminating them may lower the validity of grading.

Fourth, when combining scores from several types of assessment, we should consider their relative contributions to the final grades. The results of formal tests (formative and summative) may take the highest weight because they contribute the most to the achievement.

Fifth, select the appropriate methods of grading. We can use pass-fail, A-B-C-D-E/F, excellent-very good-good-fair-poor-very poor, or 4-3-2-1. Whichever we use, we have to complete it with the description of each grade.

Sixth, it is important to review the grades which are in the borderline. Sometimes we have students whose scores are just below the cut-off line, we need to review their achievements, so that if necessary we can help them.

F. TYPES OF GRADING

The types of grading can be divided into two, namely, absolute grading and relative grading (Gronlund and Waugh, 2009: 192-197). An **absolute grading** is a grading which is resulted from comparing a student's language performance to a pre-specified performance standard. An absolute grading usually uses letter grades (A to E) which are defined using a 100-point system, for example:

Table 8.1
Range of Scores in Absolute Grading

Grade	Alternative 1: Points	Alternative 2: Points
A	91 - 100	90 - 100
B	81 - 90	75 - 89
C	71 - 80	60 - 74
D	61 - 70	45 - 59
E or F	0 - 60	0 - 44

The first alternative is a hypothetical example, and the second alternative had been used by IKIP Malang before it changed into a university. Currently, the State University of Malang (Universitas Negeri Malang, 2013: 89) adopts the grade system as follows.

A = 85 - 100
A- = 80 - 84
B+ = 75 - 79
B = 70 - 74
B- = 65 - 69
C+ = 60 - 64
C = 55 - 59
D = 40 - 54
E = 0 - 39

The distribution of the scores is basically arbitrary, but ideally it should be based on (Gronlund and Waugh, 2009: 193) "the teacher's experience with this and past groups of students, knowledge concerning the difficulty of the intended learning outcomes, the difficulty of the tests and other assessments used, the conditions of learning, and the like". As proposed by Brown and Abeywickrama (2010: 331), the letter grades should mean as follows:

A = Excellent
B = Good
C = Adequate
D = Inadequate/unsatisfactory
E or F = failing/unacceptable

Absolute grading has strengths and limitations. Concerning these, Gronlund and Waugh (2009: 195) state:

Strengths

1. Grades can be described directly in terms of student performance, without reference to the performance of others.
2. All students can obtain high grades if mastery outcomes are stressed and instruction is effective.

Limitations

1. Performance standards are set in an arbitrary manner and are difficult to specify and justify.
2. Performance standards tend to vary unintentionally due to variations in test difficulty, assignments, student ability, and instructional effectiveness.
3. Grades can be assigned without clear reference to what has been achieved (but, of course, they should not be).

Relative grading is when a student's performance is compared to the performance of the other members in his/her group. In a relative grading, the results of the students' test are ranked from the highest to the lowest. Then, we determine the percentage of the students who will get A, B, C, D, and E. See the examples in Table 8.2 below.

Table 8.2
Percentage of Students in Relative Grading

Grade	School A	School B	School C
A	20%	25%	50%
B	30%	25%	30%
C	35%	25%	20%
D	10%	15%	0%
E	5%	10%	0%

As seen in this table, each grade is determined by the percentage of the students, regardless of their scores. As an illustration, we can compare the absolute grading and relative grading, in a class containing 40 students, with a range of scores between 42 and 91. The scores gotten by each student are ranked as follows.

S-1 = 91	S-11 = 80	S-21 = 69	S-31 = 56
S-2 = 90	S-12 = 80	S-22 = 68	S-32 = 54
S-3 = 88	S-13 = 78	S-23 = 68	S-33 = 53
S-4 = 88	S-14 = 76	S-24 = 67	S-34 = 52
S-5 = 86	S-15 = 75	S-25 = 65	S-35 = 50
S-6 = 86	S-16 = 75	S-26 = 65	S-36 = 50
S-7 = 85	S-17 = 74	S-27 = 63	S-37 = 47
S-8 = 84	S-18 = 72	S-28 = 60	S-38 = 45
S-9 = 83	S-19 = 70	S-29 = 58	S-39 = 44
S-10 = 82	S-20 = 70	S-30 = 56	S-40 = 42

Note: S-1, S-2, ... = Student 1, Student 2, and so on

If we use Alternative 2 of the absolute grading (see above) for these 40 students, we will find the distribution of the students as follows.

A (90-100) = 2 students
B (75-89) = 14 students
C (60-74) = 12 students
D (45-59) = 10 students
E (0-44) = 2 students

Now, if we compare it with the distribution of students using School C in the relative grading, the distribution of the 40 students will be as follows.

A (50%) = 20 students, ranging from scores 91 to 70
B (30%) = 12 students, ranging from scores 69 to 54
C (20%) = 8 students, ranging from scores 53 to 42
D (0%) = 0 student
E (0%) = 0 student

In this relative grading, the basis of grading is the percentage of students, not the range of scores. As seen above, the grades are A, B, and C only, regardless of the scores. If C is determined as the lowest passing level, it means that all the students pass. As in the absolute grading, relative grading also has strengths and limitations. They are (Gronlund and Waugh, 2009: 195) as follows.

Strengths

1. Grades can be easily described and interpreted in terms of rank in a group.
2. Grades distinguish among levels of student performance that are useful in making prediction and selection decisions.

Limitations

1. The percent of students receiving each grade is arbitrarily set.
2. The meaning of a grade varies with the ability of the student group.
3. Grades can be assigned without clear reference to what has been achieved (but, of course, they should not be).

The strengths of the absolute grading and relative grading are different, but there are similarities in their limitations. The ranges of scores in absolute grading and the percentage of students in relative grading are determined arbitrarily. Due to being arbitrary, the determination of score ranges and the percentage of students is done by the teacher or the school, and may not be related to the concept of achievement. These limitations can be minimized by the appropriate knowledge and experience of those who make the decision.

In distributing the grades, sometimes we are faced with conditions (Brown and Abeywickrama, 2010: 325), such as:

1. whether we need to give grade A to a big number of students when the scores achieved are not satisfactory,

2. our well-supported impression that the students did not prepare themselves well before taking the test,
3. our intention to include 'effort' or 'improvement' of the low achievers, or
4. our suspicion that we made a too difficult test, or we did not teach very clearly.

In the case of the third and the fourth points, we may consider modifying the distribution of students. How did you usually determine your students' grades?

G. SOME CULTURAL FACTORS ABOUT GRADING

Brown and Abeywickrama (2010: 329-330) mention about cross-cultural factors in grading. The factors are cited here for the purpose of their use as a consideration for determining students' grades.

First, previously it was unheard of students to be required to make a self-assessment or peer-assessment. Once, one of the writers of this module asked his students to make a self-assessment at the end of a course, the students refused with a reason that it was improper for students to assess themselves. In another occasion, a group of teachers in an in-service training were required to make a peer-assessment. They made an agreement among them that everyone should give high scores for their peers, which was surely not fair. It seems that it takes time before we can practice self-assessment and peer-assessment.

Second, it is common in Indonesia that no student questions the grading criteria used by the teacher. However, it should not mean that the criteria for grading not to be specified.

Third, there used to be a perception that a teacher who could make difficult tests was a superior teacher. Students were 'not allowed' to get a perfect grade (since grade A is reserved for God, or for very few extraordinary students only). Students should be satisfied with Bs. Now, this inappropriate perception should be changed.

Fourth, there is a belief that one single final examination is sufficient for determining the final grade. As stated earlier, a single source for the final grade is risky. There can be some factors (e.g. sickness or limited time to study) which make students not prepared or ready for the examination, and in

turn produce an unreal performance in the test. Several and a variety of sources for the final grades will make them more valid and reliable.

Fifth, it was once perceived that it was not necessary to prepare students to do their best in the test. This perception needs a change. Currently, iBT TOEFL requires the test-takers to do the test through internet. Consequently, they should be prepared to be able to use internet to do the test. Preparing students for a test is important, as far as not telling them the test answers.

H. ALTERNATIVES TO LETTER GRADING

Letter grading is not the only method of grading. As we have alternative assessments, we also have alternative gradings. Instead of giving letters A to E to the students' test papers, we may give comments on the quality of students' works. Brown and Abeywickrama (2010: 332-337) suggest the following alternative gradings.

For formative feedback, such as a test, paper, report, or other formal tasks, besides figures or letter grades, grading may include:

1. a teacher's marginal and/or end comments
2. a teacher's written reaction to a student's self-assessment of performance
3. a teacher's review of the test in the next class period
4. peer-assessment of performance
5. self-assessment of performance
6. a teacher's conference with the student

For summative assessment, with slight modifications the grading may include:

1. a teacher's marginal and/or end of exam/paper/project comments
2. a teacher's summative written evaluative remarks on a journal, portfolio, or other tangible product
3. a teacher's written reaction to a student's self-assessment of performance in a course
4. a completed summative checklist of competencies, with comments
5. narrative evaluations of general performance on key objectives
6. a teacher's conference with the student

In addition to these feedback gradings, we may try to use self-assessment. Since students in Indonesia are not yet accustomed to doing self-

assessment, Brown and Abeywickrama (2010: 333) recommend the use of guided self-assessment, such as:

1. checklists
2. a guided journal entry that directs the student to reflect on the content and linguistic objectives
3. an essay that self-assesses
4. a teacher-student conference

The next possibility is to use narrative evaluations. The strength of this evaluation is that it is individualized. The limitations are that it is difficult to be quantified, it takes much time for a teacher to write, and if it is accompanied with a letter grade, there is a tendency that students pay little attention to the narration. An example of narrative evaluation is presented below (quoted from Brown and Abeywickrama, 2010: 335).

Course: Grammar Instructor: Grade: A

Mayumi was an outstanding student in her grammar class this semester. Her attendance was perfect, and her homework was always turned in on time and thoroughly completed. She always participated actively in class, never hesitating to volunteer to answer questions. Her scores on the quizzes throughout the semester were consistently outstanding. Her test scores were excellent, as exemplified by the A+ she received on the final exam. Mayumi showed particular strengths in consistently challenging herself to learn difficult grammar; she sometimes struggled with assignments, yet never gave up until she had mastered them. Mayumi was truly an excellent student, and I'm sure she will be successful in all her future endeavors.

In the current curriculum (2013 Curriculum) in Indonesia, an attempt has been made to use narrative evaluation; however, it is still in a very simple form. For example, a student is given a comment that "at the end of this semester his/her basic competencies in English have been achieved (or have not been achieved)".

Still another alternative is checklist evaluation. This is simpler than narrative evaluation, because the teacher just puts a check-mark in the

appropriate column and a short comment. Below is an example for a mid-semester evaluation checklist.

Student's name : _____
Class : _____
Subject : _____

	Excellent progress	Satisfactory improvement	Needs improvement	Unsatisfactory progress
Listening	_____	_____	_____	_____
Speaking	_____	_____	_____	_____
Reading	_____	_____	_____	_____
Writing	_____	_____	_____	_____

Comments: _____
_____
_____

Suggestion for the rest of the semester:
_____
_____
_____

The last alternative is conference. It is a one-on-one meeting between the teacher and a student. The teacher can ask a student's perception about his/her progress or problem in his/her study. Then, the teacher leads the student to find alternative solutions to the problem(s). However, with a big number of students, conference will take time.

I. SOME PRINCIPLES IN GRADING

The practices of grading differ widely from one teacher to another or from one school to another school; however, it is not necessary to make them uniformed. Brown and Abeywickrama (2010: 337) further mention some principles which can be used as guidelines for designing a grading system in a school or an institution, namely:

a. grading is not necessarily based on a universally accepted scale

b. grading is sometimes subjective and context-dependent
c. grading of test is often done on a "curve"
d. grades reflect a teacher's philosophy of grading
e. grades reflect an institutional philosophy of grading
f. cross-cultural variation in grading philosophies needs to be understood
g. grades often conform, by design, to a teacher's expected distribution of students across a continuum
h. tests do not always yield an expected level of difficulty
i. letter grades may not "mean" the same thing to all people
j. alternatives to letter grades or numerical scores are highly desirable as additional indicators of achievement

So, we may design our own grading system, based on our needs.

EXERCISE 1

Answer all the questions in this exercise.

1) Mention one strength and one weakness of objective marking.
2) Mention one strength and one weakness of subjective marking.
3) For classroom situation, in what circumstances will you employ objective tests, and in what circumstances will you employ subjective tests?
4) Besides writing, what other skill needs subjective marking procedure?
5) Why should the inclusion of factors, such as behavior, effort, motivation, and attendance, be considered carefully in grading achievement tests?

SUMMARY

From this first Learning Activity we learned that marking/scoring is a process of getting a total raw score of a student's test. Marking or scoring can be objective. Grading is giving meanings to scores, commonly using A – E with descriptions in the form of 100-point system. The bases of grading can be formal test, alternative assessment, participation, improvement, behavior, effort, motivation, attendance, or others. The criteria to select the bases of grading can be institutional policy, clarity of components, or other considerations. Grading is fair

and effective when it is informed to the students, focused on achievement, based on a variety of data, using weighting if necessary, and using letter grades and their descriptions.

The types of grading can be absolute or relative. There are also some cultural factors in grading regarding self-assessment, clarity of grading criteria, difficult test, single test, and student preparation. We can also use alternatives to letter grading, such as, comments, simple self-assessment, narrative evaluation, and the use of checklist. There is a number of principles for creating grading, e.g. not necessary universal, can be objective or subjective, tendency to be curved, based on teacher's belief, and so on.

FORMATIVE TEST 1

Answer all the following questions.

1) Why is using a single test (e.g. final-semester test) risky for determining students' grades?
2) Mention one strength and one weakness of absolute grading.
3) What is wrong when a teacher says to his/her students, "If you do not pay attention to my explanation today, I will give you a test tomorrow"?
4) Why is it difficult for students in Indonesia to conduct a self-assessment?
5) What is the importance of preparing students before doing the real test?

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

Scoring Interpretation

This Learning Activity is focused on scoring interpretation. After doing this activity, students are expected to be able to make interpretations of the students' scores.

What does a score mean? When, say, Anton got a score of 30 on his reading test, what does this score mean and how should we interpret it?

Standing alone, the figure/score has no meaning at all and is completely uninterpretable. At the most superficial level, we do not even know whether this figure represents a perfect score of 30 out of 30 or a very low percentage of the possible score, such as 30 out of 100. Even if we do know that the score is 30 out of 40, or 75%, what then?

When we obtain scores from a language test, we need to be able to report these in ways that are meaningful and useful for test users. There are several groups to whom we will potentially need to report the test scores (Bachman, 2005):

1. Test takers who may want to know how they rank in their group, whether their score is at the acceptable standard, or what their relative strengths and weaknesses are.
2. Teachers, who may want to use the results to make decisions about diagnosis and progress of their students, and for assigning grades.
3. School administrators who may want to use the test results to help inform decisions about resource allocation and curriculum development.
4. Parents who may want to find out how well their children are progressing in school.
5. Funding agencies who may sponsor individual students and require feedback on their progress, or who may fund special programs and are interested in obtaining feedback on the effectiveness of the program.
6. Test developers themselves who may need to record scores in ways that are meaningful to them, to help them better interpret and use the results of the test for feedback, as part of the test development.

There are two approaches to score interpretation: 1) a *norm-referenced* score interpretation, or 2) a *criterion-referenced* score interpretation.

Norm-referenced score interpretation compares test takers to a sample of peers. The goal is to rank students as being better or worse than other students. Norm-referenced test score interpretation is associated with traditional education. Students who perform better than others pass the test, and students who perform worse than others fail the test.

Because norm-referenced scores provide information about the relative standing of individuals in a group, they are particularly appropriate for situations where the decisions to be made are relative, where we want to select or reward those individuals who have scored in the top portion of those who took the test. (See Figure 8.2 for an example of a norm-referenced interpretation). Thus, norm-referenced scores can be used for the following specific purposes (Bachman, 2005):

1. Comparing the performance of different individuals on the same test.
2. Comparing the performance of a given individual on different tests, or on different forms of the same test.
3. Comparing the performance of individuals with that of norm groups.
4. Reporting test results to individuals not familiar with the test itself.

A popular method for interpreting scores using norm-reference used to be using a normal-curve or bell-shaped distribution, which looks like the following.

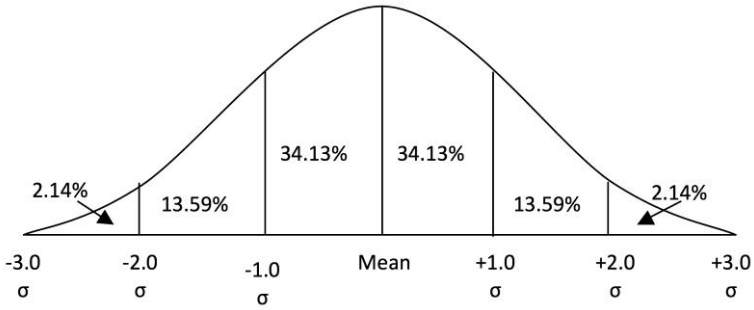

Figure 8.2
Normal-curve Distribution

Norm-reference is based on an assumption that a group of students fall in a normal distribution, in which the majority is average and a small number is above average and below average. When we calculate the students' scores using norm-reference, we begin with the raw scores gotten by the students.

Suppose after a test, it is found among the 50 students that the highest score obtained is 65 and the lowest score is 15. From all the scores, we count the mean score (M) which is, for example, 40. From here we will count the standard deviation (SD) of the scores (using SD formula). Suppose it is found that the SD is 10. So, the scores of M + 1SD is 50, M + 2SDs is 60, and M + 3SDs is 70 (in the example = 65). Likewise, the scores of M – 1SD is 30, M – 2SDs is 20, and M – 3SDs is 10 (in the example = 15). According to the normal distribution, the percentage of students between -1SD and +1SD will be about 68% (= 34 out of 50 students). These will be called average students. The percentage of students between +1SD and +2SD is 14% (= 7 students), and the percentage of students between +2SD and +3SD is 2% (= 1 student). The percentages of students between -1SD and -2SD and between -2SD and -3SD are also 14% (= 7 students) and 2% (= 1 student) respectively. If we distribute these to grades A – E, they look as follows:

Grade	% of Students	Number of Students	Scores
A	2%	1	61 – 65
B	14%	7	51 – 60
C	68%	34	31 – 50
D	14%	7	21 – 30
E	2%	1	15 – 20

Notes:
In this example:

- R (range of scores) : 15 – 65
- M (mean) : 40
- SD (standard deviation) : 10
- N (number of students) : 50

In this illustration, the numbers of students who get A, B, C, D, and E are 1, 7, 34, 7, and 1 respectively. Whatever the range of the scores, the distribution of students who get A to E will be the same. If, for instance, the highest score is 95 and the lowest score is 65, using norm-referenced calculation, the distribution of students will be the same. We can predict, then, that whatever the range of the scores, there will be 2% of students who get A, 14% of students who will get B, and so on. If, for example, the passing grades are A – C, we know that about 84% or 42 (out of 50) students will

pass. This is similar to the principle of relative grading which we discussed earlier. One weakness of norm-reference or relative grading is that two students from different classes, who each gets grade A, may have different competencies.

Another way to interpret scores, based on norm-reference, is by using **percentile**. For example, what percentile would a score of 30 represent? The score 30 falls into 2% + 14% = 16% (in the area of -2SDs). This means that the position of the student who gets 30 is equal to or above 16% of the number of students. Another example, a student who gets the score 60 falls into 2% + 14% + 68% + 14% = 98%, which means that the position of this student equals to or is above 98% of the total students. In other words, it can also be said that this student is below the 2% of best students.

If percentile scores indicate how a score obtained by a student is related (in percentage) to other scores in the group, **standard scores** "represent a student's score in relation to how far the score varies from the test mean in terms of standard deviation units" (Brown, 2005: 123). Two types of standard scores are presented below, namely, z scores and T scores (Brown, 2005: 123-126, and Gronlund and Waugh, 2009: 215-218).

The **z scores** indicate the distance of a score, in term of SD, below or above the mean score (M). The z score can be obtained using the formula: **student's score minus mean score, divided by standard deviation.** For example, a student's score is 60, M is 40, and SD is 10, the z score of this student is 60 minus 40, divided by 10, equals 2 or +2.00. This means that the score of the student (= 60) is 2 SDs above the mean (M). On the other hand, if a student's score is 25, the z score of this student is: 25 minus 40, divided by 10, equals -1.5, which means that the score is 1.5 SDs below the M.

The score from z score calculation is rather difficult to be understood by common people, because (a) it can be positive or negative, (b) the score is small, ranging from -3.00 through 0.00 to +3.00, and (c) the score includes decimals (Brown, 2005: 125). One technique to overcome these possible difficulties is by transforming z score into **T score**. The formula for T score is simple, namely: T = 10z + 50. So, using the above examples, a student whose z score is +2.00, his/her T score is 10(+2) + 50 = 70; and another student whose z score is -1.5, his/her T score is 10(-1.5) + 50 = 35. These scores (70 and 35, in 100-point scale) will make sense for common people.

In addition to z score and T score, Gronlund and Waugh (2009: 217-219) include stanine (standard nine) as a standard score. In stanine, students' scores are distributed to nine groups, with the percentages as follows.

Stanine 1	= 4%	the lowest stanine scores
Stanine 2	= 7%	
Stanine 3	= 12%	
Stanine 4	= 17%	
Stanine 5	= 20%	the center of distribution
Stanine 6	= 17%	
Stanine 7	= 12%	
Stanine 8	= 7%	
Stanine 9	= 4%	the highest stanine scores

This stanine distribution is actually similar to the norm-reference distribution, except that in the stanine distribution there are nine groups of scores, whilst in the norm-reference distribution there are five groups of scores. In addition, the percentages of each group are different.

Criterion-referenced score interpretations require tests with the following characteristics (Bachman, 2005):

1. The tests need to be based on a clearly defined criterion domain
2. The types of tasks to be included in the test need to be clearly described
3. Test tasks need to be selected to represent the criterion domain
4. The number of tasks or scale levels need to be sufficient to make dependable inferences about ability.

From the characteristics of criterion-referenced test above, criterion-referenced scores clearly need to be interpreted in term of the specific criterion domain or domains on which they are based. This is similar to absolute grading which we discussed earlier. This also implies that if a test covers multiple domains, separate scores need to be provided for each domain. (See Figure 8.3 for an example)

For example, in a writing test that includes tasks intended to measure the test takers' control of grammar, vocabulary, cohesion, and organization in writing, separate ratings and associated descriptions need to be provided for each of these areas.

How to Interpret IELTS

Test takers receive scores on a band scale from 1 to 9. A profile score is reported for each skill. The four individual scores are averaged and rounded to produce an overall band score. Overall band scores and scores for each component (Listening, Reading, Writing and Speaking) are reported in whole bands or half bands.

Overall band scores

Test takers receive a Test Report Form including or listing their overall band score and their sub-scores on each of the four components: Listening, Reading, Writing and Speaking. Each of the component scores is equally weighted. The overall band score is calculated by taking the mean of the total of the four individual component scores.

Overall band scores are reported to the nearest whole or half band. The following rounding convention applies; if the average across the four skills ends in .25, it is rounded up to the next whole band, and if it ends in .75, it is rounded up to the next whole band.

Thus, a test taker achieving 6.5 for Listening, 6.5 for Reading, 5.0 for Writing and 7.0 for Speaking would be awarded an overall band score of 6.5 (25 + 4 = 6.25 = Band 6.5).

Likewise, a test taker achieving 4.0 for Listening, 3.5 for Reading, 4.0 for Writing and 4.0 for Speaking would be awarded an overall band score of 4.0 (15.5 ÷ 4 = 3.875 = Band 4.0).

On the other hand, a test taker achieving 6.5 for Listening, 6.5 for Reading, 5.5 for Writing and 6.0 for Speaking would be awarded band 6 (24.5 ÷ 4 = 6.125 = Band 6).

Figure 8.3
IELTS Score Interpretation
(Source: IELTS Guide for Teachers, www.ielts.org)

- CURRENT ISSUES ABOUT ASSESSMENT IN INDONESIA

Since the last quarter of the 20th century there have been a pro and contra about the National Examination (NE) in Indonesia, and recently there was an idea to use the scores of NE for a university entrance. These two issues are discussed by Saukah (2013: 49-54), and are summarized below.

First, in Indonesian elementary and (junior and senior) secondary schools, all students are required to take an NE at the end of each school level. The scores gotten by the students are used for determining the success or failure to leave school. The policy to use this single determinant is reacted in pro and contra.

Those who are pro to the NE give reasons that:

1. without NE, students will be reluctant to study for the school examination (SE), because they know that they will pass.
2. with NE, the students' mastery of subject competencies can be compared for mapping the quality of education in Indonesia.
3. if passing is only determined by the SE, there is a possibility that certain schools arrange so that all students will pass, without considering mastery of competencies.

And, those who are contra to the NE give reasons that:

1. using NE to determine the passing is not fair, since the qualities of education in Indonesia are not equal. Students who do not pass may be due to the lack of facilities or low quality of their teachers, not because they are not serious in learning.
2. the NE becomes a high-stake, which may give negative impacts. The teaching-learning process becomes like in the private learning course, where students are drilled with tests. The students may feel psychologically pressured, and do irrational practices, such as, praying in graves or asking for help from a "clever" person, which are not directly related to the mastery of learning competencies.
3. when the NE is conducted at the same time nationally, it needs a complex management which is costly and has a risk of "leaking". The highest risk is the loss of honesty in the students, teachers, and principals for the sake of the school prestige.

The solutions to these problems need to be found. Saukah (2013: 53) proposes alternative solutions, i.e. that NE should not be used as a single factor to determine the students' passing, or NE is used only for mapping the qualities of schools in Indonesia, which consequently is not conducted every year.

The second issue is the idea to use the scores of NE for a university entrance. Saukah (2013: 49-51) makes the following analysis. The purpose of

NE is to measure students' **achievement** of the competencies of subject-matters taught at school, while the purpose of university entrance test is **to predict** whether the prospective university students will succeed or not in studying at the university. From the interpretations of the scores, it can be seen that the NE and the university entrance test are different. The NE scores are interpreted using **criterion-reference method**, while the scores of the university entrance test are interpreted using **norm-reference method**. Therefore, Saukah (2013: 51) concludes that the idea to use NE scores for university entrance is incorrect.

EXERCISE 2

To check your understanding of presentation in this first activity, please answer oll of the following questions.

1) Seen from the purposes of norm-referenced scores, what kind of test is appropriate with the purpose? Choose one of the following and explain why you choose your answer:
   A. Diagnostic test
   B. Achievement test
   C. Proficiency test
   D. End-of-a-unit test

2) Seen from the purposes of criterion-referenced scores, what kind of test is appropriate with the purpose? Choose one and explain why you choose your answer:
   A. Placement test
   B. Achievement test
   C. Proficiency test
   D. Selection test

3) The test scores are called "feedback" when they are used for:
   A. the test-takers who want to know how they rank in their group
   B. the teacher who uses the scores for students remedy
   C. the school administrators as a consideration for buying facilities
   D. the parents who may want to find out how well the progress of their children

4) The test scores are aimed as "accountability" when the are used for:
   A. the test-takers who want to know how they rank in their group
   B. the teacher who uses the scores for students remedy
   C. the school administrators as a consideration for buying facilities
   D. the parents who may want to find out how well the progress of their children

5) What are the differences between the National Examination (NE) and the university entrance test?

SUMMARY

From this second Learning Activity, we can conclude that the interpretation of scores and grades can use norm-reference or criterion-reference. Norm-reference score interpretation, which is similar to relative grading, deals with normal distribution, percentile, z score, and T score. Criterion-reference score interpretation is similar to the absolute grading. In this activity, it is also presented some issues about pro and contra about the National Examination (NE) in Indonesia and the use of NE scores as a university entrance.

FORMATIVE TEST 2

As a formative test, answer the questions below.

1) If the M of scores is 50, the SD is 10, and the z score of a student is +1.3, what is this student's score?
2) If a student's z score is +2.35, what is his/her T score?
3) What is the similarity and difference between normal-curve distribution and stanine distribution?
4) If in a test the majority of students get low scores, how can we increase the students' scores in a reasonable way?
5) If the results of today's test are below the MPL (minimum passing level), may we ask the students to retake the same test tomorrow?

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 the keys to the exercises and summative tests. You are advised to finish the exercise and summative test first, before looking at the answer keys. Please note that in the keys are the key ideas; therefore, your answers do not necessarily have exactly the same as those in the keys. In some questions you may have different correct answers. If you are not sure of your answers, you may contact the tutors in the Open University.

Exercise 1

1) The strength of objective marking is that it is easy to do, and the weakness is that the mark is difficult to be used for diagnostic purpose.
2) The strength of subjective marking is that it can show the strengths and weaknesses of a student, and the weakness is that the marking should be done by a scorer who is expert in the contents to be tested.
3) For classroom situations, we can use objective test to measure low-level thinking contents (i.e. remembering, understanding, and applying), and we can use subjective test to measure higher-level thinking contents (i.e. analyzing, evaluating, and creating).
4) Speaking skill
5) Because they are not directly related to the learning achievement

Exercise 2

1) Norm-referenced scoring is suitable for proficiency test, because the language abilities of the test-takers usually vary.
2) Criterion-referenced scoring is suitable for achievement test, because it is easy to determine the levels of students' mastery.
3) (A) because feedback is given for the benefit of students
4) (D) because parents have the right to know their children's progress
5) The purpose of the NE is achievement and the university entrance test is prediction; and the NE scores are criterion-reference while the university entrance test is norm-reference.

Formative Test 1

1) Because a single test may now show students' real abilities. At the time of test, many factors may affect, such as sickness, not prepared to take the test, or cheating.
2) One strength of absolute grading is that if two students from different groups/classes get the same grade (e.g. B), it can be concluded that their abilities are more or less the same. One weakness of absolute grading is if the range of scores is low (e.g. 20 – 70), there is a possibility that none of the students gets A (the highest grade).
3) It is not a good statement. It will frighten students. By nature, the term "test" is usually not perceived positively by students; therefore, a teacher should not make the perception of "test" more negative.
4) Because Indonesian students are not yet accustomed to doing self-assessment
5) It is important, especially about the administration of a test. For example, for doing a computer-based test students need to be familiar with the use of computer before doing the test.

Formative Test 2

1) The z score +1.3 comes from Score minus M, divided by SD, namely, Score minus 50, divided by 10. So, the student's score is 63.
2) If a student's z score is +2.35, his/her T score is 10(+2.35) + 50 = 73.5
3) The similarity between normal-curve distribution and stanine distribution is that both belong to norm-referenced interpretation. The difference is that in the normal-curve distribution the range is five, while in the stanine the range is nine.
4) We can increase the students' scores by combining the test scores and the students' scores in their daily/alternative assessment.
5) It is not allowed to ask students to retake the same test in the next day after the first test, because there is a possibility that the students still remember the contents of the test and learn them for the re-test; or, logically between the first test and the second test there is too little time to study again to improve the second test results.

Note:
For further informations about score interpretation, you are advised to read:

- Brown (2005)

- Brown and Abeywickrama (2010)
- Gronlund and Waugh (2009)

(See the list of references below for their details)

References

Bachman. 2005. *Statistical analysis for language assessment.* Cambridge: Cambridge University Press.

Brown, H.D. and Abeywickrama, P. 2010. *Language assessment: Principles and classroom practices (2nd edition).* White Plains, NY: Pearson Education.

Brown, J.D. 2005. *Testing in language programs: A comprehensive guide to English language assessment (new edition)*. New York: McGraw-Hill.

Gronlund, N.E. and Waugh, C.K. 2009. *Assessment of student achievement (9th edition)*. Upper Saddle River, NJ: Pearson Education.

Lougheed, L. 2000. *(Barron's) How to prepare for the TOEIC test (2nd edition)*. Jakarta: Binarupa Aksara.

Sharpe, P.J. 2005. *(Barron's) How to prepare for the TOEFL (11th edition)*. Jakarta: Binarupa Aksara.

Universitas Negeri Malang. 2013. *Pedoman pendidikan UM tahun akademik 2013/2014*. Malang: BAKPIK UM.

www.ets.org (accessed on November 20th, 2015)

www. ibt toefl score (accessed on November 20th, 2015)