CHAPTER ONE
((Preliminaries))

When it comes to testing in general and language testing in particular, our first job is to know what we mean by the term 'test'. Is it different from other related terms such as measurement, assessment, and evaluation? In everyday conversation, we, as not-specialist individuals, tend to use these terms interchangeably, but almost all educators regard them as being different.

1.1. What Is Test?

The term 'test' is often considered the narrowest of these related terms. Test is defined as any procedure used to measure a factor or assess some ability. It refers to an instrument designed to elicit a specific sample of an individual's behavior and often connotes the presentation of a set of questions to be answered, and as a result to obtain a measure (that is, a numerical value) of a characteristic of a person.

A test is a method of measuring a person's abilities, knowledge, or performance in a given domain. As this definition suggests, a test is first a method. It is an instrument - a set of techniques, procedures, or items - that requires performance on the part of the test-taker. Second, a test must measure. Some tests measure general ability, while others focus on very specific competencies or objectives. Next, a test measures an individual's ability, knowledge, or performance. Testers need to understand who the test-takers are, what is their previous experience and background? Is the test appropriately matched to their abilities? How should test-takers interpret their scores? A test measures performance, but the results imply the test-taker's ability, or, to use a concept common in the field of linguistics, competence. Most language tests measure one's ability to perform language, that is, to speak, write, read, or listen to a subset of language. On the other hand, it is not uncommon to find tests designed to tap into a test-taker's knowledge *about* language: defining a vocabulary item, reciting a grammatical rule, or identifying a rhetorical feature in written discourse. Finally, a test measures a given domain (e.g. language, math, etc.).

1.2. What Is Measurement?

Measurement often implies a broader sense than test: we can measure characteristics by means other than tests. Using observations, rating scales, or other devices that allow us to obtain information in a quantitative form is measurement.

Measurement is a process of quantifying the characteristics of a person according to explicit procedures and rules. This definition includes three distinguishing features: quantification, characteristics, and explicit rules and procedures.

A. *Quantification:* Measurement involves the assigning of numbers, and this distinguishes measurement from qualitative description such as a verbal account, or non-verbal, visual representation.

B. *Characteristics:* Measurement involves assigning numbers to both physical and mental characteristics of persons. Physical attributes such as height and weight can be observed directly. Mental attributes and abilities, however, which are sometimes called traits or constructs, can only be observed indirectly. These mental attributes include characteristics such as aptitude, intelligence, motivation, attitude, native language, etc.

C. Rules and procedures: The third distinguishing characteristic of measurement is that quantification must be done according to explicit rules and procedures. That is, the 'blind' or haphazard assignment of numbers to characteristics of individuals cannot be regarded as measurement.

1.3. What Is Evaluation?

Evaluation is the process of delineating, obtaining, and providing useful information for judging decision alternatives. In general, it is the systematic gathering of information for the purpose of decision – making. Therefore, when we deal with evaluation, we are necessarily trying to make a sound decision based on information collected. Evaluation is also interpreted as the determination of the congruence between performance and objectives. Last but not least, evaluation is a process that allows one to make a judgment about the desirability or value of a measure.

Note: It is important to point out that we never measure or evaluate people. We measure or evaluate characteristics or properties of people, e.g. their knowledge of English, their fluency in speaking English, ability to teach, and so forth.

Example 1: **Evaluation is different from testing in that the former is mostly designed for ………** (آزاد ۸۳)

     1) measurement   2) making decision   3) generalization   4) quantification

*Explanation***:** Test, measurement and evaluation are three different terms, so option (1) is out. Evaluation is used for making decisions, so option (2) is the answer. Quantification (option 4) has to do with measurement.

Example 2: **In the process of evaluation, we measure ……… .** (آزاد ۸۴)

     1) people themselves   2) people's characteristics
     3) people's goals   4) people's needs

Explanation: It is important to point out that we never measure or evaluate people. We measure or evaluate characteristics or properties of people: their scholastic potential, knowledge of English, ability to teach and so forth. Thus, option (2) is the answer.

1.4. Relationship among test, measurement, and evaluation

Evaluation is much more comprehensive and inclusive than measurement, and testing is just one type of measurement (see figure below). Evaluation does not necessarily entail testing. By the same token, tests in and of themselves are not evaluative. Tests are often used for pedagogical purposes, either as a means of motivating students to study, or as a means of reviewing material taught, in which case no evaluative decision is made on the basis of the test results. Tests may also be used for purely descriptive purposes. It is only when the results of tests are used as a basis for making a decision that evaluation is involved.

The term measurement is just limited to quantitative description of students, that is, the results of measurement are always expressed in numbers (e.g., Mary correctly scored 35 out of 40 vocabulary items). It does not include qualitative descriptions (e.g., Mary's composition was neat), nor does it imply judgment concerning the worth or value of the obtained results. Evaluation, on the other hand, may include both quantitative descriptions (measurement) and qualitative descriptions (non-measurement) of students. In addition, evaluation always includes 'value judgment' concerning the desirability of the results. Also, it should be noted that measurement has a information-providing function while evaluation has a decision-making function.

The relationships among these three concepts are illustrated in the following figure:
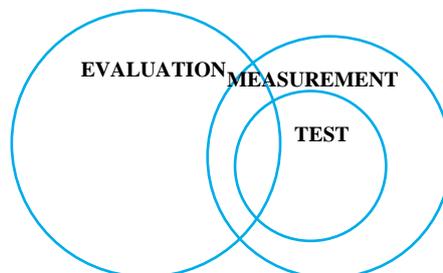
Relationships among measurement, test, and evaluation

Example 3: **Which of the following distinguishes "evaluation" from "testing"?** (سراسری ۸۸)

     1) Decision making   2) Comparison of measures
     3) Reliance on numerical values   4) Quantitative procedures used

Explanation: Evaluation has to do with decision making (so option 1 is correct). Reliance on numerical values (option 3) has to do with test. Quantitative procedures (option 4) relate to measurement.

✎ *Example* **4:** **The determination of congruence between performance and objective is interpreted as ……… .**

(آزاد ۸۵)

   1) testing   2) evaluation   3) assessment   4) measurement

Explanation: Evaluation has been defined in a variety of ways:
\# The process of delineating, obtaining, and providing useful information for judging decision alternatives;
\# The determination of the congruence between performance and objectives;
\# A process that allows one to make a judgment about the desirability or value of a measure.
Thus, option (2) is the answer.

✎ *Example* **5:** **Which of the following best describes the relationship between measurement and evaluation?**

(دکتری ۹۴)

   1) Both necessarily entail testing.
   2) Both are instruments designed to elicit specific samples of an individual's behavior.
   3) Evaluation only involves qualitative assessment, but measurement involves the assigning of numbers.
   4) Measurement has an information-providing function, while evaluation has a decision- making function.

Explanation: Evaluation does not necessarily entail testing. (So option 1 is ruled out). Option 2 pertains to **test**. Option 3 is ruled out because evaluation involves both **q**ualitative and quantitative assessment. Thus, option (4) is the answer.

✎ *Example* **6:** **……… does not necessarily entail testing; rather, it is involved when the results of a test are used for decision making.**

(سراسری ۹۸)

   1) Evaluation   2) Measurement   3) Assessment   4) Impact

Explanation: Decision making has to do with evaluation. Therefore, option (1) is the answer.

1.5. What Is Assessment?

Another term closely related to testing is assessment. We tend to think of testing and assessing as synonymous terms, but they are not. Tests are prepared administrative procedures that occur at identifiable times in a curriculum when learners muster all their faculties to offer peak performance. **Assessment**, on the other hand, is an **ongoing process** that encompasses a much wider domain. Whenever a student responds to a question, offers a comment, or tries out a new word or structure, the teacher subconsciously makes an assessment of the student's performance. Tests, then, are a subset of assessment; they are certainly not the only form of assessment that a teacher can make. Bachman (2004) defines *assessment* as the process of collecting information about a given object of interest according to procedures that are **systematic** and **substantively grounded**. Assessment can be classified on two continuums: informal / formal and formative / summative.

1.5.1. Informal vs. Formal Assessment

Informal assessment is designed to elicit performance without recording results and making fixed conclusions about a student's competence. It is virtually always nonjudgmental, in that you as a teacher are not making ultimate decisions about the student's performance.

   Informal assessment can take a number of forms, starting with incidental, unplanned comments and responses, along with coaching and other impromptu feedback to the student. Examples include putting a smiley face on homework or saying "Nice job!" or "Good work!" "Did you say *can* or *can't?*" "I think you meant to say you *broke* the glass, not you *break* the glass."

   On the other hand, formal assessments are exercises or procedures specifically designed to tap into a storehouse of skills and knowledge. They are systematic, planned sampling techniques constructed to give teacher and student an appraisal of student achievement.

📖 **Note:** All tests are formal assessments, but *not* all formal assessment is testing.

1.5.2. Formative vs. Summative Assessment

Another useful distinction to bear in mind is the function of an assessment. Two functions are commonly identified in the literature: formative and summative assessments.

🖋 *Example* **10:** **The acronym CRADLE is associated with ………… assessment.** (سراسری ۹۷)

     1) performance   2) traditional   3) portfolio   4) dynamic

Explanation: The term CRADLE pertains to portfolio assessment. Thus, option (3) is the answer.

4.6. Subjective vs. Objective Items

The distinction here refers to the way a test item is scored and has little or nothing to do with the form of a test. If no judgment is required on the part of the scorer, then the scoring is *objective*. Multiple-choice and true-false types of tests are popular kind of the so-called objective measures because in scoring a multiple-choice (or true/false) test, the scoring procedures are almost systematic and there are few to almost no possibilities of scores fluctuating from one scorer to another. If judgment is called for, however, the scoring is said to be *subjective*. Composition-type tests (also called productive tests) and translation tests are usually subjective.

📖 **Note:** It would be a misunderstanding to assume that all composition-type tests are subjective or all multiple choice type tests are objective.

The following examples show what is meant by this.

Ex 1: There are four ……….. in a year.   *a. seasons*   *b. months*   *c. weeks*   *d. days*

Ex 2: There are four ……seasons in a year. a. beautiful   *b. uninteresting*   *c. tiring*   *d. exciting*

The first multiple-choice item is **objective**, because no judgment is required on the part of the scorer. In other words, the answer to this item is such that almost everybody would agree on a particular response (i.e. there is only one possible answer). However, the second multiple choice item is **subjective** because the answer depends on the tastes, attitudes, likes and dislikes of the scorer. This means that such an item has more than one possible response.

Ex 3: What are the four seasons in a year?   **Answer:** *Spring, Summer, Autumn, Winter.*

Ex 4: How many exciting seasons are there in a year?

The first composition-type item (i.e. an item that requires producing language rather than selecting) is **objective**, again because no judgment is required on the part of the scorer. The second composition-type item, however, is subjective because there is no agreement on a particular response to this item.

🖋 *Example* **11:** **The terms subjective and objective refer to the way a test is ……… .** (آزاد ۸۴)

     1) constructed   2) administered   3) scored   4) interpreted

Explanation: These two terms relate to how a test is scored. Thus, option (3) is the answer.

🖋 *Example* **12:** **The so-called 'subjective' items are those in which ……… .** (سراسری ۸۷)

     1) the test taker needs to produce the language

     2) both comprehension and production are necessary

     3) the rater has biases for or against some test takers

     4) there is more than one correct response for each item

Explanation: Option (1) is ruled out because a multiple-choice item which is sometimes subjective does not require producing the language. Option (2) is irrelevant. Option (3) sounds tricky, but it is not necessarily correct. Option (4) is definitely correct because in a *subjective test item,* there is more than one correct response for each item, because judgment is involved.

4.7. Recognition vs. Suppletion Items

Recognition-form items (the same as Brown's receptive response items) refer to items in which the examinees recognize the correct response form among the alternatives provided for each stem. Therefore, **multiple-choice** items and **true/false** items are **recognition** items. On the other hand, **suppletion** or **completion** form items (the same as Brown's productive response items) require the examinee to supply the missing part(s) of the stem or complete an incomplete stem (i.e. these items require production). Thus, essay-type items, fill-in-the-blank or short response items fall into this category.

✎ *Example* **13:** **Sentence completion-type items are sometimes preferred to MC-type items because they** ……….....… .

(آزاد ۸۲)

   1) can tap production to some extent   2) are quite valid

   3) are scored very easily   4) are used by most teachers

Explanation: Unlike Multiple choice (MC) items, which require no language production on the part of examinees, completion-type items / suppletion / essay-type items tap production. Thus, option (1) is the answer.

✎ *Example* **14:** **Which of the following are types of personal-response assessment?**

(سراسری ۹۸)

   1) Conferences, portfolios, and performance
   3) Conferences, portfolios, and self-assessments
   2) Portfolios, self-assessments, and performance
   4) Portfolios, self-assessments, and short-answer items

Explanation: Personal response items encourage the students to produce responses that hold personal meaning. In other words, the responses allow students to communicate in ways and about things that are interesting to them personally. Personal response item formats include self-assessments, conferences, and portfolios. Thus, option (3) is the answer.

4.8. Productive (or Constructed) Response Items

As already discussed, *productive (or constructed) response items* are those in which the testees are required to actually produce language by writing, speaking, or acting in some way, rather than simply selecting answers receptively. This category includes fill-in, short-response, and task types of items. Here are some advantages and disadvantages of *productive (or constructed) response items:*

Advantages:
o There is virtually no guessing factor
o They allow for testing productive language use
o They allow for testing the interaction of receptive and productive skills

Disadvantages:
o They are difficult and time consuming to score
o Scoring is subjective
o Bluffing is possible

✎ *Example* **15:** **All of the following are advantages of constructed-response items EXCEPT** ………. .

(سراسری ۹۸)

   1) there is virtually no guessing factor
   2) they allow for testing productive language use
   3) they allow for testing the interaction of receptive and productive skills
   4) they are directly related to and integrated into curriculum

Explanation: Choice (4) is unrelated.
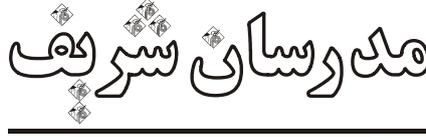
CHAPTER EIGHT
((Reliability))

8.1. A General Definition

Apart from being practical, authentic, and interactive, a good test must be **reliable**. Reliability, in as non-technical terms as possible, means **consistency or stability** of one's score on its various administrations. This means if a test is administered to a group of examinees on Tuesday and Wednesday, the results must be not fluctuate. In reality, however, this is not possible. Even if we assume that the test is excellent, that the conditions of administration are almost identical, that the scoring calls for no judgment on the part of the scorers and is carried out with perfect care, and that no learning or forgetting has taken place during the one-day interval, nevertheless we would not expect every individual to get precisely the same score on the Wednesday as they got on the Tuesday. Human beings are not like that; they simply do not behave in exactly the same way on every occasion, even when the circumstances seem identical. A reliable test is usually contrasted with an **unreliable test**, in which one's score might fluctuate from one administration to the other. That is, one's score on its various administrations will be inconsistent. The concept of reliability can be investigated within several theoretical frameworks, the easiest one being the **Classical True Score (CTS**, hereafter) theory.

8.2. Classical True Score (CTS)

According to this theory, changes in one's score on its various administrations are inevitable, i.e. if one takes two measures of the same attribute, e.g., height, or verbal knowledge, the two measures will not resemble each other exactly. These fluctuations can be due to different reasons: headache, some disturbing factors, not getting enough sleep the night before the exam, some sort of learning, testee's mental alertness or emotional state, changes in the test environment from one day to the next, etc. The changes that are predictable are called **systematic variations** (or **systematic errors**). The changes, however, which are not predictable are called **unsystematic variations** (or **random errors**). Thus, whenever several observations of a person are made and several scores are recorded on the same ability, those scores are likely to differ from one another. This variation is due, in part, to systematic variation and, in part, to unsystematic variation. The systematic variation contributes to the reliability and the unsystematic variation contributes to the unreliability of a test.

8.2.1. Assumptions of the CTS

One important assumption of the CTS is that the score one obtains on a test (called **observed score**) cannot be true manifestation of one's ability in that particular trait because this score comprises two factors or components: **a true score** (i.e. an errorless score) that is due to an individual's level of ability and an **error score** (or a measurement error), that is due to factors other than the ability being tested (factors such as poor health, fatigue, lack of interest or motivation). This assumption can be represented in the following formula:

$$\boxed{X = T + E}$$

Where, X is the *observed score*, T is the *true score*, and E is the *error score*. Since the observed score includes the error score, it can be greater than, equal to, or smaller than the true score. If there is absolutely no error of measurement (i.e. error score = 0), the observed score will equal the true score $(X = T + 0 \Rightarrow X = T)$. However, when there is a measurement error, which is often the case, it can lead to an overestimation or an underestimation of the true score (i.e., X< T and X > T, respectively). From the above formula, it can easily be understood that there is an inverse relationship between the error score and the true score. The greater the error score, the smaller the true score. By the same token, the smaller the error score, the greater the true score.

These scores can be changed into their corresponding **variance** terms as follows:

$$V_x = V_t + V_e$$

Where, V_x is the observed score variance, V_t is the true score variance component, and V_e is the error score variance component. In this case, the variance of the true scores does not change because the true scores are constant. The variance of the observed scores, nonetheless, fluctuates because of the extent of the error score. Since error variance is included in the observed variance, the variance of the observed scores is always greater than the variance of the true scores. That is, the magnitude of the observed variance equals the magnitude of the true variance plus the magnitude of the error variance.

Now that we have come to a clear understanding of the concepts of true score variance, error score variance, and observed scored variance, we can define reliability in its technical form:

Reliability, which is represented by r, is the **ratio/ proportion of true score variance to observed score variance**. As a formula, this means:

$$r = \frac{V_t}{V_x}$$

Where, r is reliability, V_t is the true score variance, and V_x is the observed score variance. Since the true score is not directly measurable, the value of V_t is never known. Therefore, the above formula can be rewritten as follows:

$$r = \frac{V_x - V_e}{V_x}$$

If $V_e = V_x$ (i.e., the error score variance is so large to equal the observed score variance), then $r = 0$:

$$r = \frac{V_x - V_e}{V_x} = \frac{V_x - V_x}{V_x} = \mathbf{0}$$

If $V_e = 0$ (i.e., the error score variance equals zero), then $r = 1$:

$$r = \frac{V_x - V_e}{V_x} = \frac{V_x - 0}{V_x} = \frac{V_x}{V_x} = \mathbf{1}$$

Thus, the magnitude of reliability (i.e. **reliability coefficient**) can range from **zero to one**. The reliability of zero, which is the minimum, means that all observed variation is due to error. That is, the test is completely unreliable. On the other hand, the reliability of 1 indicates that there is no error in measurement and the test is perfectly reliable. Of course, this does not happen in reality. All tests show a certain degree of unreliability. But the closer the magnitude of reliability to unity, the more reliable the test will be.

Reliability coefficients can be interpreted as the percent of systematic, or consistent, or reliable variance in the scores on a test. For instance, if the scores on a test have a reliability coefficient of $r = 1.00$, by moving the decimal two places to the right, the tester can say that the scores are 100% consistent, or reliable. Similarly, if $r = 0.91$, the tester can say that the scores are 91% consistent, or reliable, with 9% measurement error $(100\% - 91\% = 9\%)$. Similarly, if $r = 0.40$, then the variance on the test is only 40% systematic and 60% measurement error.

8.2.2. Problems with the CTS
There are two problems with the CTS:
❖ It treats error variance (or error score variance) as homogeneous in origin.
❖ It considers all error to be random (or unsystematic), and consequently fails to distinguish systematic error from random error.

✎ *Example* **1: The possible ranges of reliability, item discrimination, and item facility are respectively**

……… . (سراسری ۸۷)

   1) −1 …+1; 0 … +1; −1 … +1   2) −1 … +1; 0…+1; 0.37…0.63

   3) 0…+1; −1 … +1; 0…+1   4) 0…+1; −1…+1; 0.37…0.63

Explanation: Reliability ranges from zero to unity; i.e. $0 \leq r \leq +1$.
Item discrimination ranges from -1 to +1; i.e. $-1 \leq ID \leq +1$.
Item facility ranges from 0 to 1; i.e., $0 \leq IF \leq 1$. Thus, option (3) is the answer.

🐟 *Example* 2: **Which of the following is a problem with the classical true score theory (CTS)?** (سراسری ۹۵)

　　　1) It considers all errors to be systematic.　　　2) It treats error variance as homogeneous in origin.
　　　3) It distinguishes systematic errors from random errors.　　4) It defines reliability in terms of true score variance.

Explanation: One problem with the CTS model is that it treats error variance as homogeneous in origin. Thus, option (2) is the answer.

———

🐟 *Example* 3: **The magnitude of reliability will equal +1 when ……… .** (سراسری ۸۹)

　　　1) the true score is greater than the observed score
　　　2) the estimate of the true score approximates its real value
　　　3) systematic variation is greater than error variance
　　　4) there is no unsystematic variation in measurement

Explanation: When there is no unsystematic variation in measurement, it means V_e (error variance) = 0, in which case r = 1 according to the following formula: (thus, option (4)) is the answer)

$$r = \frac{V_x - V_e}{V_x}$$

———

As mentioned, **error variance** (or *error score variance*) pertains to the effect of those variables generating variance due to other extraneous sources (i.e., sources not directly related to the purpose of the test, such as headache, fatigue, etc.). If we minimize the effects of these extraneous sources, we minimize measurement error, in which case reliability is maximized. Thus, the investigation of reliability is concerned with answering the question, 'How much of an individual's test performance is due to measurement error, or to factors other than the language ability we want to measure?' and with minimizing the effects of these potential factors on test scores. Here is a list of these extraneous sources: (To increase reliability, attempts should be made to control for these variables).

☐ **Variance due to environment**
　　☐ location
　　☐ space
　　☐ ventilation
　　☐ noise
　　☐ lighting
　　☐ weather
☐ **Variance due to administration procedures**
　　☐ directions
　　☐ equipment
　　☐ timing
　　☐ mechanics of testing
☐ **Variance due to scoring procedures**
　　☐ errors in scoring
　　☐ subjectivity
　　☐ evaluator biases
　　☐ evaluator idiosyncrasies
☐ **Variance attributable to the test and test items**
　　☐ test booklet clarity
　　☐ answer sheet format
　　☐ particular sample of items
　　☐ item types
　　☐ number of items
　　☐ item quality
　　☐ test security

☐ **Variance attributable to examinees**
　　☐ health
　　☐ fatigue
　　☐ physical characteristics
　　☐ motivation
　　☐ emotion
　　☐ memory
　　☐ concentration
　　☐ forgetfulness
　　☐ carelessness
　　☐ test-wiseness*
　　☐ comprehension of directions
　　☐ guessing
　　☐ task performance speed
　　☐ chance knowledge of item content

Table: Potential sources of error variance

* *Test-wiseness is a test taker's capacity to utilize the characteristics and formats of the test and the test taking situation to guess the correct answer and hence receive a high score. It includes the ability to comprehend easily almost any test directions, or knowledge of guessing strategies, or strategies for maximizing the speed of task performance.*
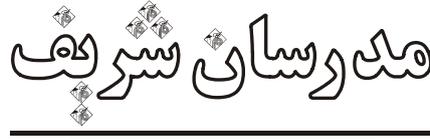
CHAPTER TWELVE

((Testing Vocabulary & Grammar))

12.1. Testing Vocabulary

The goal of testing vocabulary is to assess the subjects' knowledge of lexical items. To do so, we can utilize one of the following techniques:

1. Limited Response

This technique, which is based on physical responses and visuals, is particularly useful with testing *children* and *beginning-level adults*, who have not yet mastered language skills. Here are examples of how this test format looks:

The tester holding a book: *What color is this book?*   **Testee (or testees) responds**: *green*

The tester pointing to a chair: *What is this?*   **Testee (or testees) responds**: *a chair*

Advantages of Limited Response:

1. It causes less stress or nervousness than other types of tests.

2. It avoids skills such as reading and writing that have not yet been developed.

3. It can be scored easily and objectively.

Limitations of Limited Response:

1. It requires individual testing, which takes longer than group testing.

2. It is usually difficult to test abstract words with this technique.

3. Sketches are sometimes ambiguous (e.g., an orange may look like a ball; running may look like dancing or jumping)

2. Multiple-choice Paraphrase

In this technique, a sentence with one word underlined or bolded is given. The testee's task is to choose the best *synonym* or *paraphrase* of the underlined word. Here is a typical item:

➢ *He was <u>irate</u> when he heard about the new plans.*

 a. interested   *b. surprised*   *c. angry**   *d. sad*

This is a good way of testing the subjects' understanding of specific words and expressions. This technique, nonetheless, has three disadvantages. In the first place, it limits the testing to only one word in each test item. In the second place, sometimes lexical items do not lend themselves to four sensible paraphrases. In the third place, it allows the testees to ignore the whole context and get to the meaning of the word being tested. A variation of this technique looks like this:

➢ **"Irate" means ………..**

 a. interested   *b. surprised*   *c. angry**   *d. sad*

Advantages of Multiple-choice Paraphrase:

1. Context preparation is rather easy.

2. Scoring is easy and consistent.

3. It is a sensitive measure of achievement.

Limitations of Multiple-choice Paraphrase:

1. It is difficult to find good synonyms.

2. It is easy for students to cheat

3. Multiple-choice Completion

You are already familiar with this format, so an example of how this test form looks suffices:

➢ *She quickly …… her lunch.*

 a. drank   b. ate*   c. drove   d. slept

Advantages of Multiple-choice Completion:

1. It helps students see the full meaning of words by providing natural contexts. Also, it is a good influence on instruction: It discourages word-list memorization.

2. Scoring is easy and consistent.

3. It is a sensitive measure of achievement.

Limitations of Multiple-choice Completion:

1. It is rather difficult to prepare good sentence contexts that clearly show the meaning of the word being tested.

2. It is easy for students to cheat.

✎ *Example* **1: Which of the following is an advantage of the multiple-choice paraphrase technique in testing vocabulary?** (سراسری ۹۷)

    1) Context preparation is rather easy.

    2) It causes less stress compared to other types of tests.

    3) It resembles more real-life teaching approaches.

    4) It avoids skills such as reading and writing that have not been developed yet.

Explanation: Options (2) and (4) are advantages of *limited response* technique. Option (3) is irrelevant. Option (1) is an advantage of *multiple-choice paraphrase*. Thus, option (1) is correct.

12.2. Guidelines for Constructing Multiple-choice Vocabulary Items

In constructing vocabulary items, the following guidelines should be kept in mind, as well as those presented in chapter six:

1. The context should be clear enough to provide the testees with a clear meaning. Here is an example of a vocabulary item with poor and unclear context:

➢ **Poor item**. *Reza ……. Ali's car.*

 a. bought   b. washed   c. borrowed   d. returned

2. A vocabulary item should not contain more difficult semantic features in the stem than the area being tested.

Poor item. *Being unfortunate to have been bereaved of his belongings, John........ lucky David's book.*

 a. sold   b. borrowed   c. lent   d. returned

This item is poor because words such as 'unfortunate', 'bereaved' and 'belongings' are more difficult than the word being tested.

3. Avoid items that require the examinees to possess a certain body of knowledge which is beyond the mastery of lexical items.

➢ **Poor item.** *We can raise sugar cane in Khuzestan because cane needs a .... climate.*

 a. dry   b. cold   c. warm   d. humid

4. Make sure you don't give away the right answer through grammatical cues.

Poor item. *Mary was ……..… in the job of helping the poor.*

 a. interested*   b. tired   c. angry   d. sad

In this item, the preposition *in* reveals that *interested* is the correct response.

5. Try to avoid pairing a word of opposite meaning with the right answer, as in options b and c in item below:

➢ **Poor item:** *He plans to <u>purchase</u> some candy for his mother.*

 a. make   b. buy*   c. sell   d. steal

6. Try to avoid distractors with the same meaning, as in options c & d in item below:

➢ **Poor item.** *His <u>remorse</u> was great indeed.*

 a. wealth   b. sadness*   c. strength   d. power

Similarly, item below is poor because *glad* and *pleased* are synonyms:

➤ ***The old woman was always <u>courteous</u> when anyone spoke to her.***

 a. polite*   b. glad   c. kind   d. pleased

In this item, if the testees recognize the synonym, they may realize immediately that neither is the correct option, since there can be only one correct answer.

7. If the item being written is a multiple-choice paraphrase type, the choices should be easier than the word being tested. Moreover, they should be of the same grammatical form as the underlined word.

➤ **Poor.** ***The child was <u>frightened</u> of being left alone in home.***

 a. annoyed   b. terrified*   c. ashamed   d. dismayed

This item is poor because choices b & d are more difficult than *frightened*.

8. Try to get distractors that are related to the area or topic covered in the stem.

➤ **Poor item.** ***He just hit his <u>shin</u>.***

 a. leg*   b. cousin   c. fender   d. fruit

In reading this item, students may recognize *shin* as some part of the body. If so, they can get the question right by the process of elimination. More challenging distractors would include *back*, *foot*, and *arm*.

9. Tests of vocabulary should avoid grammatical structures that the students may find difficult to comprehend. Similarly, tests of grammar should contain only those lexical items which present no difficulty to the testees.

10. Avoid items that require taking into account cultural differences. For example, in the following item, options b and d would be correct in certain societies since it is impolite to accept a gift without first vehemently refusing it.

➤ **Poor item.** ***Emma cried out with ……….... at the beautiful present Mrs. White gave her.***

 a. delight   b. horror   c. dismay   d. anger

✎ *Example 2:* **The item below has all of the following shortcomings EXCEPT ……… .**   (سراسری ۸۶)

It is ……… to pass the exam to get the job.

a) imperative   *a) unnecessary*   *c) polite*   *d) courteous*

    1) there is one pair of opposite words
    2) the choices widely vary in terms of the level of difficulty
    3) the stem is conductive to more than one correct choice
    4) the item measures knowledge of both inflection and derivation

Explanation: Choice (1) is a deficiency of the item because *imperative* and *unnecessary* are opposite words.

Choice (2) is a deficiency of the item because *polite* and *imperative* are of different difficulty level.

Choice (3) is a deficiency of the item because both *imperative* and *unnecessary* could be the answer. Thus, choice (4) is the answer.

✎ *Example 3:* **The following item is bad due to all of the following EXCEPT ……… .**   (سراسری ۸۷)

When people ……… pictures of the atrocities on TV, there was a spontaneous reaction against the war, which seemed quite logical

a) saw   *b) heard*   *c) listened*   *d) analyzed*

    1) lexical difficulty of the stem   2) options with different lengths
    3) options related to different concepts   4) unnecessary phrase/phrases in the stem

Explanation: The words chosen for the options show that the testees should not have been high level students. Meanwhile, the stem contains such difficult words (such as *atrocities* and *spontaneous*) that don't match proficiency level of testees. As a result, choice (1) is not the answer. Options a, b and c are sensory verb, but option d is a performative verb. Thus, options are not related to the same general topic or area. Therefore, choice (3) is not the answer either. The clause at the end of the sentence could be reduced, so choice (4) is also wrong. Therefore, only option (2) is the answer because in this item, choices are of the same length.

✍ *Example* **4: What is the problem with the following vocabulary item in which the candidates should choose the best definition for the underlined word?** (سراسری ۹۶)

The old man was always <u>courteous</u> when people approached him.
a) polite   *b) happy*   *c) kind*   *d) pleased*

    1) There is a pair of synonyms used as distractors.
    2) The stem does not provide sufficient contextual clues.
    3) The stem provides a grammatical clue as to what the correct answer is.
    4) The correct option and the distractors are not at the same level of difficulty.

Explanation: It is advisable to avoid using a pair of synonyms as distractors (Options b & d are synonymous). If the testees recognize the synonym, they may realize immediately that neither is the correct option, since there can be only one correct answer. Thus, option (1) is the answer.

✍ *Example* **5: Why is the following test item poorly constructed?** (سراسری ۹۳)

Everyone was ………….. with John's performance on the test.
a) amazed   *b) satisfied*   *c) discouraged*   *d) ashamed*

    1) There is grammatical clue which gives away the correct answer.
    2) The options are not at the same level of difficulty.
    3) There is more than one correct answer.
    4) The distractors are very tricky.

Explanation: The preposition *with* following the blank is a grammatical clue that gives away the correct response: b (*satisfied*). Thus, option (1) is the answer. Amazed **at** / discouraged **by/at** / ashamed **of**.

✍ *Example* **6: What is wrong with the following multiple-choice item?** (سراسری ۹۴)

I can't believe him, because he's always ……… lies.
 a) saying   *b) telling*   *c) keeping*   *d) talking*

    1) The options are of more or less the same length.
    2) The options do not belong to the same area of meaning.
    3) There is a pair of synonyms in the options.
    4) The stem is not long and informative enough.

Explanation: Choices (1) and (4) are certainly wrong. The other two choices, however, require in-depth analysis. Option (3) is wrong because *say* and *tell* are not necessarily synonyms. If we consider them synonyms, then *say lies must also be correct. This is while there is only one correct response (choice b) to this item. According to Heaton, words like these are **false synonyms**, i.e. words with equivalent meanings but not interchangeable, that are in fact highly recommended as distractors. Moreover, this item deals with knowledge of **collocations**. Thus, only choice (2) is correct because in vocabulary tests, some test writers argue that the options should be related to the same general topic or area of meaning. In this item, the verb *keeping* needs improvement. A more challenging distractor would be *speaking*.

12.3. Testing Grammar (or structure)
The goal of testing grammar is to assess the testees' knowledge of grammatical structures. To do so, we can utilize one of the following techniques: (The techniques with symbol (*) will be presented in the chapter dealing with testing writing)

1	Limited response	*2*	Simple completion
3	Multiple-choice completion	*4*	Error-recognition items *
5	Rearrangement (or ordering) items *	*6*	Transformation (or paraphrasing) items *
7	Pairing and matching items (already discussed)	*8*	Combination items *
9	Addition items *		

CHAPTER SIXTEEN
((Testing Writing))

16.1. Composition

It would seem obvious that the most **direct** way of measuring students' writing ability would be to have them write. This is known as **free writing** or **composition** (or essay) writing, which involves, in essence, a topic for the examinees to write a composition of a certain length. The content of the required topic should be familiar to the testees who should as well be precisely guided as to what is expected of them.

Composition is the most **face-valid** test of writing. Over the past century, it has been both praised and criticized. Those who have championed composition writing have generally included the following points in their defense:

❖ Writing compositions provides the testees with an opportunity to show their ability to organize and communicate their own ideas, using their own vocabulary, register, and style.

❖ Since the testees actually write, composition tests motivate them to improve their writing. On this account, composition tests have a **positive backwash effect** on writing instruction.

❖ Composition tests are easy and quick to prepare, an important advantage to the busy classroom teacher.

❖ Composition writing is an important, sound measure of overall writing ability.

The critics of composition testing have usually answered along the following lines:

❖ Composition tests are *unreliable* measures because (1) students perform differently on different topics and on different occasions; and (2) the scoring of compositions is by nature highly **subjective.**

❖ In writing compositions, students can cover up weaknesses by avoiding problems (e.g., the use of certain grammatical patterns and lexical items) they find difficult. Such evasion is impossible with other well-prepared tests.

❖ Composition tests require much scoring time; for this reason, compositions add greatly to the expense and administrative problems of large-scale testing.

16.2. Indirect Ways of Measuring Writing

In addition to composition, which is a direct way of measuring writing, there are some indirect ways that you can employ. These indirect tasks depend on considering a variety of factors, such as the proficiency level of the testees, their age, purpose of writing and so on.

1. Copying

In a copying test, the test-taker should reproduce a written model as closely as possible. The testees' performance is then gauged on the perfectness of copying. This technique is particularly useful because it makes the subjects more conscious of the *mechanics* of writing and the discourse of the text. Certainly, it is more serviceable for subjects whose native language uses a writing system different from that of the language being tested.

2. Picture-Cued Tasks

Familiar pictures are displayed, and test-takers are told to write a description of what the picture represents. Assuming no ambiguity in identifying the picture, successful completion of the task requires no reliance on aural comprehension.

3. Form Completion Tasks

A variation on pictures is the use of a simple form (registration, application, etc.) that asks for name, address, phone number, and other data. Assuming, of course, that prior classroom instruction has focused on filling out such forms, this task becomes an appropriate assessment of simple tasks such as writing one's name and address.

4. Converting Numbers and Abbreviations to Words

Some tests have a section in which numbers—for example, hours of the day, dates, or schedules—are shown and test-takers are directed to write out the numbers. This task can serve as a reasonably reliable method to stimulate

handwritten English. It lacks **authenticity**, however, in that people rarely write out such numbers (except in writing checks), and it is more of a reading task (recognizing numbers) than a writing task. If you plan to use such a method, be sure to specify exactly what the objective is and then proceed with some caution.

5. Spelling Tests

In a traditional spelling test, the teacher dictates a simple list of words, one word at a time; then uses the word in a sentence and repeats the sentence; then pauses for test-takers to write the word. Scoring emphasizes correct spelling. You can help to control for listening errors by choosing words that the students have encountered before—words they have spoken or heard in their class.

6. Listening Cloze Selection Tasks

These tasks combine dictation with a written script that has a relatively frequent deletion ratio (every fourth or fifth word perhaps). The test sheet provides a list of missing words from which the test-taker must select. The purpose at this stage is not to test spelling but to give practice writing. To increase the difficulty, the list of words can be deleted, but then spelling might become an obstacle.

7. Multiple-Choice Techniques

Presenting words and phrases in the form of a multiple-choice task risks crossing over into the domain of assessing reading, but if the items have a follow-up writing component, they can serve as formative reinforcement of spelling conventions. They might be more challenging with the addition of homonyms. Here is an example.

Choose the word with the correct spelling to fit the sentence, then write the word in the space provided.

➤ *He washed his hands with …..*

 a. soap　　　　　　*b. sope*　　　　　　*c. sop*　　　　　　*d. soup*

8. Grammatical Transformation Tasks

The practice of making grammatical transformations—orally or in writing—was very popular in the heyday of structural paradigms of language teaching with slot-filler techniques and slot-substitution drills. To this day, language teachers have also used this technique as an assessment task, ostensibly to measure grammatical competence. Numerous versions of the task are possible:

❖ Change the tenses in a paragraph.
❖ Change full forms of verbs to reduced forms (contractions).
❖ Change statements to yes/no or wh- questions.
❖ Change questions into statements.

9. Ordering Tasks

One task at the sentence level may appeal to those who are fond of word games and puzzles: ordering (or reordering) a scrambled set of words into a correct sentence. Here is the way the item format appears:

Put the words below into a possible order to make a grammatical sentence:

1. cold / winter / is / weather / the / in / the

2. studying / what/you / are

3. next / clock / the / the / is / picture / to

10. Paraphrasing (also called transformation)

These assessment tasks require the student to write a sentence equivalent in meaning to the one that is given. It is helpful to give part of the paraphrase in order to restrict the students to the grammatical structure being tested. Here are two typical examples:

1. It is six years since I last saw him.

I ……… six years. {The paraphrased form will be: *I haven't seen him for six years*}

2. It was impossible to work under those conditions.

Working ………….{The paraphrased form will be: *Working under those conditions was impossible*}

11. Short-Answer and Sentence-Completion Tasks

Here is the way this item format appears:

Alicia: Where's she from?

Tony: ….. Italy.

12. Combination

As it name implies, in this assessment task test takers are required to connect sentences. Combination is done in two ways: combination by adding a connective and combination by putting one sentence inside the other. Examples:

1. She didn't feel well today …. she didn't go to work. {She didn't feel well today, **so** she didn't go to work}

2. Some people come late. They will not get good seats. {People that come late will not get good seats}

13. Sentence expansion

This assessment task involves simply adding words such as adjectives and adverbs. Or it can require adding phrases and clauses. Example:

The ( ) man hurried ( ) to the ( ) horse. ⇨ [The **old** man hurried **out** to the **frightened** horse.]

14. Reduction

This involves simply reducing longer sentences to shorter ones. Example:

He told us about a man who had a wooden leg (with) ⇨ [He told us about a man with a wooden leg]

15. Error recognition

This type of assessment task includes a sentence with some words/phrases underlined. The test-takers are to choose the one underlined portion that needs correction. Example:

The position <u>taken</u> in his <u>most recent</u> speeches <u>seem</u> to indicate a willingness <u>to compromise</u>.
<div style="text-align:center">A   B   C*   D</div>

This type of item is very popular in writing tests, yet it has to be used with reservation. It is not pedagogically wise to expose the subjects to language that contains errors; it is the correct form of the language that they should be in constant contact with. In addition, although this type of item may tell the examiner something about the editing ability of the subjects, it does not provide any information about their own writing capability when they are left to themselves.

✍ *Example 1:* **In measuring writing ability, the identification of the erroneous elements in a sentence is an example of ....** (سراسری ۹۱)

    1) subjective testing   2) recognition items   3) multiple-choice tests   4) transformation items

Explanation: *Recognition items* (or more specifically *error recognition items*) include sentences with some words/phrases underlined. The test-takers are to choose the one underlined portion that needs correction. Thus, option (2) is the answer.

✍ *Example 2:* **The following item is an example of a .... item.** (سراسری ۹۳)

Although he was too tired, he didn't stop work.
Despite ....
    1) transformation   2) rearrangement   3) combination   4) pairing

Explanation: In this item examinees should re-write the original sentence without changing the meaning, i.e. transform (or paraphrase) it. *[Despite being too tired, he didn't stop work.]* Thus, option (1) is correct.

✍ *Example 3:* **What kind of item is the following?** (سراسری ۹۴)

Reza can swim better than you.
You cannot swim ....
    1) Combination   2) Addition   3) Rearrangement   4) Transformation

Explanation: In this item examinees should re-write the original sentence without changing the meaning, i.e. transform (or paraphrase) it. [*You cannot swim better than Reza*]. Thus, option (4) is the answer.

16.3. Scoring Methods for Writing Tasks

Test designers commonly use three major approaches to scoring writing performance in general and composition writing in particular: **holistic**, **primary trait**, and **analytical**. In the first method, a single score is assigned to an essay, which represents a reader's general overall assessment. Primary-trait scoring is a variation of the holistic method in that the achievement of the primary purpose, or trait, of an essay is the only factor rated. Analytical scoring breaks a test-taker's written text into a number of subcategories (organization, grammar, etc.) and gives a separate rating for each. Each of these scoring types is discussed below in more detail.

16.3.1. Holistic Scoring

Holistic scoring (sometimes referred to as '**impressionistic**' scoring) involves the assignment of a single score to a piece of writing on the basis of an overall impression of it. In this method, the examiner glances through the writing quickly and assigns a rating such as excellent/good/fair, pass/fail, or even a score on a scale system. Inasmuch as scoring is based on the raters' impression, it is fast and time-preserving.

Advantages of holistic scoring include:
• fast evaluation
• relatively high inter-rater reliability
• scores represent "standards" that are easily interpreted by laypersons
• scores tend to emphasize the writer's strengths
• applicability to writing across many different disciplines
Its **disadvantages** includes:
• One score masks differences across the subskills within each score.
• No diagnostic information is available (no washback potential).
• The scale may not apply equally well to all genres of writing.
• Raters need to be extensively trained to use the scale accurately.

16.3.2. Analytic Scoring
For classroom instruction, holistic scoring provides little washback into the writer's further stages of learning. Primary-trait scoring focuses on the principal function of the text and therefore offers some feedback potential but no washback for any aspects of the written production that enhance the ultimate accomplishment of the purpose. Classroom evaluation of learning is best served through **analytic scoring,** in which as many as six major elements of writing are scored, thus enabling learners to home in on weaknesses and capitalize on strengths (i.e., analytic marking is **diagnostic**).

There are a number of advantages to analytic scoring. First, it disposes of the problem of uneven development of subskills in individuals. Secondly, scorers are compelled to consider aspects of performance which they might otherwise ignore. And thirdly, the very fact that the scorer has to give a number of scores will tend to make the scoring more reliable.

This method of scoring is, however, criticized for three reasons: first, it is time-consuming. Even with practice, scoring will take longer than with the holistic scoring. A second disadvantage is that concentration on the different aspects may divert attention from the overall effect of the piece of writing. A third problem is that one never knows just how to weigh each area being analyzed.

16.3.3. Primary-Trait Scoring
Yet another method of scoring, **primary trait,** focuses on how well students can write within a narrowly defined range of discourse. This type of scoring emphasizes the task at hand and assigns a score based on the effectiveness of the text's achieving that one goal. For example, if the purpose or function of an essay is to persuade the reader to do something, the score for the writing would rise or fall on the accomplishment of that function. If a learner is asked to exploit the imaginative function of language by expressing personal feelings, then the response would be evaluated on that feature alone. The advantage of this method is that it allows both writer and evaluator to focus on function.

✎ *Example 4:* **The method of scoring which requires a separate score for each of a number of aspects of a task is said to be ……… .**   (سراسری ۹۳)

   1) holistic   2) analytic   3) independent   4) impressionistic

Explanation: In *analytic* scoring, the rater breaks a test-taker's written text into a number of subcategories (organization, grammar, etc.) and gives a separate rating for each. Thus, option (2) is the answer.

✎ *Example 5:* **All of the following are disadvantages of holistic scoring of the writing skill EXCEPT it ....**   (سراسری ۹۶)

   1) provides no diagnostic information   2) may not apply equally well to all genres
   3) is a relatively slow process of evaluation   4) masks differences across various micro skills

Explanation: Holistic scoring is advantageous for fast evaluation. Thus, option (3) is the answer.

✎ *Example 6:* **An impressionistic scoring of writing ……… .**   (سراسری ۹۱)

   1) is holistic by nature
   2) is related to the process of writing
   3) is only viable at the advanced level
   4) involves the assignment of a set of tentative sores to a single paper