به نام خداوند بخشنده مهربان

# Testing

In

# ELT

## Hossein salarian

آمادگی آزمون دکتری

سالاریان، حسین ۱۳۵۱
آزمون‌سازی زبان (Testing in ELT)
مشاوران صعود ماهان ، ۱۴۰۱
۴۰۸ صفحه جدول، نمودار، (آمادگی آزمون دکتری)
شابک ISBN:978-600-458-633 -7
فهرستنویسی بر اساس اطلاعات فیپا.
فارسی و لاتین ـ چاپ اول
۱-آزمون سازی زبان(Testing in ELT) ۲- آزمونها و تمرینها
۲- آزمون دوره‌های تحصیلات تکمیلی ۴-دانشگاه‌ها و مدارس عالی ـ ایران ـ آزمونها
حسین سالاریان
**ج ـ عنوان:**
**شماره کتابشناسی ملی: ۳۷۵۹۷۰۴**

انتشارات مشاوران صعود ماهان



❐ نام کتاب: ..... آزمون‌سازی زبان (Testing in ELT)
❐ مدیران مسئول..... مجید و هادی سیّاری
❐ مولف : ..... حسین سالاریان
❐ مدیر برنامه‌ریزی و تولید محتوا.....سمیه بیگی
❐ ناشر: .....مشاوران صعود ماهان
❐ نوبت و تاریخ چاپ:..... چاپ اول /۱۴۰۱
❐ تیراژ: ..... ۱۰۰۰ نسخه
❐ قیمت:.....۳/۷۹۰/۰۰۰ ریال
❐ شابک: ..... ISBN ۹۷۸-۶۰۰-۴۵۸-۶۳۳ -۷

**انتشارات مشاوران صعود ماهان:** تهران ـ خیابان ولیعصر، بالاتر از تقاطع ولیعصر مطهری، پلاک ۲۰۵۰
تلفن: ۸۸۱۰۰۱۱۳ و ۸۸۴۰۱۳۱۳

## مقدمه ناشر

به نام تو

ایمان داریم که هر تغییر و تحول بزرگی در مسیر زندگی بدون تحول معرفت و نگرش میسر نخواهد بود. پس بیایید با اندیشه توکل، تفکر، تلاش و تحمل در توسعه دنیای فکریمان برای نیل به آرامش و آسایش توأمان اولین گام را برداریم.

چون همگی یقین داریم دانایی، توانایی می‌آورد.

**شاد باشید و دلی را شاد کنید.**

**برادران سیّاری**

**Preface**

Testing and assessment are part of modern life (Fulcher, et. al, 2007). School children around the world are constantly assessed, whether to monitor their educational progress, or for governments to evaluate the quality of school systems. Adults are tested to see if they are suitable for a job they have applied for, or if they have the skills necessary for promotion. Entrance to educational establishments, to professions and even to entire countries is sometimes controlled by tests. According to Dr. Alavi (i.e., oral discussion), testing has become an industry in the world and it will become a major at University of Tehran soon. Tests play a fundamental and controversial role in allowing access to the limited resources and opportunities that our world provides. The importance of understanding what we test, how we test and the impact that the use of tests has on individuals and societies cannot be overstated. Testing is more than a technical activity; it is also an ethical enterprise. The practice of language testing draws upon, and also contributes to, all disciplines within applied linguistics. However, there is something fundamentally different about language testing. Language testing is all about building better tests, researching how to build better tests and, in so doing, understanding better the things that we test. Language testing is about *doing*; it is about *creating* tests (Fulcher, et.al, 2007).

*This Comprehensive Book* contains 40 chapters, all necessary for PhD Entrance Exam in Iran. Each chapter has to some extent a summary with some tests at the end of it. On the whole, it has more than 250 test items with answer keys and explanations, besides the test items in the recent PhD Entrance Exams. It should be mentioned that all the test items of the years 1393 -1398 in PhD entrance exams could be answered using/reading this book. However, this is the edited version of it in which 8 chapters are added. I hope it meets your needs and provides and will be a bridge for the grade A, i.e., the full per cent, in PhD entrance exam. Remember that 'testing' is one of the main courses in this great competition which differentiates and discriminates you from others.

I would like to thank Dr. Alavi, one of the great professors on 'testing and research methodology' in Iran and especially at University of Tehran, for his constant encouragement, help, advice and efficiency to all of the students.

Hossein Salarian
Summer 2019

**Contents:**

# *Chapter 1*

## Measurement, Test and Evaluation

*Psychological Assessment*
*Testing in Contrast to Assessment*
*Informal and Formal Assessment*
*Formative and Summative Assessment*
*Essential Measurement Qualities*
*Properties of Measurement Scales*
*Characteristics that Limit Measurement*
*Steps in Measurement*
*Relevance of Steps to the Development of Language Tests*
*Relevance of Steps to the Interpretation of Test Results*
*Approaches to Language Testing*
*Current Issues in Classroom Testing*
*Tests*

**Measurement**

Measurement in the social sciences is the process of quantifying the characteristics of persons according to explicit procedures and rules. This definition includes three distinguishing features: *quantification*, *characteristics*, and *explicit rules and procedures*.

*Quantification*

Quantification involves the assigning of numbers, and this distinguishes measures from qualitative descriptions such as verbal accounts or nonverbal, visual representation. Non-numerical categories or rankings such as letter grades (A, B, C…), or labels (for example, excellent, good, average…) may have the characteristics of measurement. However, when we actually use categories or rankings such as these, we frequently assign numbers to them in order to analyze and interpret them, and technically, it is not until we do this that they constitute measurement.

*Characteristics* ➔ physical characteristic: observed directly

➘ mental characteristic = trait / construct: observed indirectly

We can assign numbers to both physical and mental characteristics of persons. In testing, we are almost always interested in quantifying mental attributes and abilities, sometimes called traits or constructs, which can only be observed indirectly.

The precise definition of 'ability' is a complex undertaking. In a very general sense, 'ability' refers to being able to do something, but the circularity of this general definition provides little help for measurement.

'Mental ability' refers to performance on a set of mental tasks. We generally assume that there are degrees of ability and that these are associated with tasks or performances of increasing difficulty or complexity. It is important to understand that it is these attributes or abilities and not the persons themselves that we are measuring.

*Rules and procedures*

The third 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. In order to be considered a measure, an observation of an attribute must be replicable, for other observes, in other contexts and with other individuals. Measures are distinguished from such 'pseudo-measures' by the explicit procedures and rules upon which they are based.

**Test**

A test, in simple terms, is a *method* of *measuring* a person's ability, knowledge, or performance *in a given domain*. Let's look at the components of this definition: 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. To qualify as a test, the method must be explicit and structured. 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.*

*A test measures performance*, but the results imply the rest-taker's ability, *or*, to use a concept common in the field of linguistics, *competence*. Performance-based tests sample the test taker's actual use of language, but from those samples the test administrator infers general competence.

*Finally a test measures a given domain*.

This domain:

- can be overall proficiency in a language- general competence in all skills, or
- can have more specific criteria, e.g., a test of pronunciation or vocabulary

In the case of a proficiency test, Carroll (1968) provides the following definition of a test:

A *psychological or educational test* is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual.

Thus, a *test is measurement instrument* designed to elicit a specific sample of an individual's behavior. *As one type of measurement*, a test necessarily qualifies characteristics of individuals according to explicit procedures. What distinguishes a test from other types of measurement is that it is designed to obtain a specific sample of behavior.

The inferences and uses we make of language test scores depend upon the sample of language use obtained. Language tests can thus provide the means for more carefully focusing on the specific language abilities that are of interest. As such, they could be viewed as supplemental to other methods of measurement. Given the limitations on measurement, and the potentially large effect of elicitation procedures on test performance, language tests can more appropriately be viewed as the best means of assuring that the sample of language obtained is sufficient for the intended measurement purposes. While measurement is frequently based on the naturalistic observation of behavior over a period of time, such as in teacher rankings or grades, such naturalistic observations might not include samples of behavior that manifest specific abilities or attributes. The value of tests lies in their capability for eliciting the specific kinds of behavior that the test user can interpret as evidence of the attributes or abilities which are of interest.

A test may be defined simply as a measuring device or procedure. When the word test is prefaced with a modifier, it refers to a device or procedure designed to measure a variable related to that modifier. In a like manner, the term psychological test refers to a device or procedure designed to measure variables related to psychology (for example, intelligence, personality, aptitude, interests, attitudes, and values).

**Evaluation**

Evaluation can be defined as the *systematic gathering of information for* the purpose of *making decisions*. The probability of making the correct decision in any given situation is a function not only of *the ability of the decision maker*, but also of *the quality of the information* upon which the decision is based. The more reliable and relevant the information, the better the likelihood of making the correct decision. One aspect of evaluation is the collection of reliable and relevant information. This information need not be, exclusively quantitative. *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. Tests may also be used *for purely descriptive purposes*. When the results of tests are used as a basis for making a decision that evaluation is involved. The majority of tests are used for the purpose of making decisions about individuals, it is important to distinguish the *information-providing function of measurement* from the *decision-making function of evaluation*.

An example of evaluation that does not involve either tests or measures is the use of qualitative descriptions of student performance for diagnosing learning problems.

# *Test:*

*A. pedagogical purposes* as a means of
- *motivating students to study*, or
- *reviewing the materials taught*
*B. purely descriptive purposes:*

*Evaluation purposes:*
*1. No test, no means: the use of qualitative description of students' performances.*
*2. Non-test means: ranking for assigning grades*
*3. Test for evaluation: achievement test determining students` progress.*
*4. Not for evaluation: for research purposes or proficiency test*

Three features of evaluation theory and practice illustrate the complexity of these developments and the difficulties inherent in the task of mapping achievements and directions.

*First, there is the question of definition*; evaluation is a form of enquiry, ranging from research to systematic approaches to decision-making.

*Second, there are two perspectives on evaluation research*.

It is viewed, on the one hand, as a type of study which has both research functions – rolling back the frontiers of knowledge – and evaluation functions – providing information for judgments or decision-making; and, on the other, as research into the processes of evaluation. The former perspective has been significant in language program evaluations, as evidenced by edited collections such as those by Alderson and Beretta (1992) Rea-Dickins and Lwaitama (1995) and Rea-Dickins and Germaine (1998). In the latter perspective, evaluation research can be seen as analogous to the research which has for decades underpinned the validity and reliability of language testing processes.

*Third, many accounts of evaluation do not reach the public domain.*

For a range of reasons, some proper, others less so, evaluation processes and findings remain either insufficiently documented or unpublished. One outcome of this feature of evaluation is the difficulty of mapping theory and practice when some of the terrain is obscured from view.

The *semantic distinction* between psychological testing and psychological assessment is blurred in everyday conversation. In many psychological evaluation contexts, it requires greater education, training, and skill to conduct an assessment than to simply administer a test.

**Psychological assessment** is the gathering and integration of psychology- related data for the purpose of making a psychological evaluation that is accomplished through the use of tools such as tests, interviews, case studies, behavioral observation, and specially designed apparatuses and measurement procedures. **Psychological testing** is the process of measuring psychology-related variables by means of devices or procedures designed to obtain a sample of behavior.

**Note:** Some misconceptions about language testing are:
1) there is no best way to test language ability for any given situation.
2) a test is either good or bad, depending on whether it satisfies one particular quality.
3) language test development depends on highly technical procedures and should be left to experts

**Testing in Contrast to Assessment**

In contrast to the process of administering, scoring, and interpreting psychological tests (psychological testing), psychological *assessment* may be conceived as *a problem-solving process* that can take many different forms. How psychological assessment proceeds depends on many factors, not the least of which is the reason for assessing. Different tools of evaluation—psychological tests among them—might be marshaled in the process of assessment, depending on the particular objectives, people, and circumstances involved as well as on other variables unique to the particular situation.

Admittedly, the line between what constitutes testing and what constitutes assessment is not always as clear as we might like it to be. However, by acknowledging that such ambiguity exists, we can work to sharpen our definition and use of these terms. It seems useful to distinguish the differences between testing and assessment in terms of the objective, process, and outcome of an evaluation and also in terms of the role and skill of the evaluator.

Different assessors may approach the assessment task in different ways. Some assessors approach the assessment with minimal input from assesses themselves. Other assessors view the process of assessment as more of a collaboration between the assessor and the assessee. For example, in one approach to assessment, referred to (logically enough) as ***collaborative psychological assessment***, the assessor and assessee may work as "partners" from initial contact through final feedback (Fischer, 1978, 2004). Another variety of collaborative assessment may include an element of therapy as part of the process. Stephen Finn and his colleagues (Finn, 2003; Finn & Martin,1997; Finn & Tonsager, 2002) have described a collaborative approach to assessment called ***therapeutic psychological assessment***. Here, therapeutic self-discovery and new understandings are encouraged throughout the assessment process.

Another approach to assessment that seems to have picked up momentum in recent years, most notably in educational settings, is referred to as **dynamic assessment**. While the term dynamic may at first glance suggest to some a psychodynamic or psychoanalytic approach to assessment, as used in this context it refers to the interactive, changing, or varying nature of the assessment. In general, dynamic assessment refers to an interactive approach to psychological assessment that usually follows a model of (1) evaluation (2) intervention of some sort, and (3) evaluation. Dynamic assessment is most typically employed in educational settings, although it may be employed in correctional, corporate, neuropsychological, clinical, and most any other setting as well.

Intervention between evaluations, sometimes even between individual questions posed or tasks given, might take many different forms, depending upon the purpose of the dynamic assessment (Haywood & Lidz, 2007). For example, an assessor may intervene in the course of an evaluation of an assessee's abilities with increasingly more explicit feedback or hints .

The purpose of the intervention may be to provide assistance with mastering the task at hand. Progress in mastering the same or similar tasks is then measured. In essence, dynamic assessment provides a means for evaluating how the assessee processes or benefits from some type of intervention (feedback, hints, instruction, therapy, etc.) during the course of evaluation. In some educational contexts, dynamic assessment may be viewed as a way of measuring not just learning but so-called learning potential, or "learning how to learn" skills. The interventionist approach is rooted in quantitative interpretation of the ZPD, while the interactionist approach is rooted in qualitative interpretation of the ZPD.

**Note**: One example of *test-management strategy* is going back and forth between a passage and within a given item in order to obtain more information about what we are looking for it.

Psychological tests and other tools of assessment may differ with respect to a number of variables such as content, format, administration procedures, scoring and interpretation, procedures, and technical quality. The content (subject matter) of the test will vary with the focus of the particular test. But even two psychological tests purporting to measure the same thing—for example, personality —may differ widely in item content. This is so because what is deemed important in measuring "personality" for one test developer might be entirely different for another test developer; different test developers employ different definitions of "personality." Additionally, different test developers come to the test development process with different theoretical orientations. For example, items on a psychoanalytically oriented personality test may have little resemblance to those on a behaviorally oriented personality test, yet both are personality tests. A psychoanalytically oriented personality test might be chosen for use by a psychoanalytically oriented assessor, and an existentially oriented personality test might be chosen for use by an existentially oriented assessor.

The term format pertains to the form, plan, structure, arrangement, and layout of test items as well as to related considerations such as time limits. Format is also used to refer to the form in which a test is administered: computerized, pencil-and-paper, or some other form. When making specific reference to a computerized test, format may further refer to the form of the software: PC- or Apple/Mac-compatible. The term format is not confined to tests; it is also used to denote the form or structure of other evaluative tools and processes, such as the specific procedures used in obtaining a particular type of work sample.

Tests differ in their administration procedures. Some tests, particularly those designed for administration on a one-to-one basis, may require an active and knowledgeable test administrator. The test administration may involve demonstration of various kinds of tasks on the part of the assessee as well as trained observation of an assessee's performance. Alternatively, some tests, particularly those designed for administration to groups, may not even require the test administrator to be present while the test takers independently do whatever it is the test requires.

Tests differ in their scoring and interpretation procedures. To better understand how and why, let's define score and scoring.

**Score** is a *code* or *summary statement*, usually but not necessarily numerical in nature, that reflects an evaluation of performance on a test, task, interview, or some other sample of behavior. **Scoring** is the *process* of assigning such evaluative codes or statements to performance on tests, tasks, interviews, or other behavior samples.

Tests differ widely in terms of their guidelines for scoring and interpretation. Some tests are designed to be scored by the test takers themselves, and others are designed to be scored by trained examiners. Still other tests may be scored and fully interpreted within seconds by computer.

Tests differ with respect to their technical quality. More commonly, reference is made to what is called the psychometric soundness of a test. Synonymous with the antiquated term psychometry, **psychometrics** may be defined as the science of psychological measurement. Variants of these words include the adjective psychometric (which refers to measurement that is psychological in nature) and the nouns psychometrics and psychometrician (both referring

to psychological test users). One speaks of the psychometric soundness of a test when referring to how consistently and how accurately a psychological test measures what it purports to measure. Assessment professionals also speak of the psychometric utility of a particular test or assessment method. In this context, *utili*ty refers to the usefulness or practical value that a test or assessment technique has for a particular purpose.

The assessor and the assessee are two parties in any assessment. The third party in an assessment may be an observer who is there for any number of reasons. The third-party observer may be a supervisor of the assessor, a friend or relative of the assessee, a representative of the institution in which the assessment is being conducted, a translator, an attorney, or someone else. The social influence effect that occurs has been referred to in the testing and assessment literature as **social facilitation**, probably because the presence of third parties was initially associated with increments in performance (Aiello & Douthitt, 2001). Proponents of third-party access to psychological assessment argue that it is necessary for purposes such as clinical training.

During test administration, and especially in one-on-one or small-group testing, rapport between the examiner and the examinee can be critically important. In this context, **rapport** may be defined as a working relationship between the examiner and the examinee. Such a working relationship can sometimes be achieved with a few words of small talk when examiner and examinee are introduced.

**Note: Protocol** refers to the form or sheet or booklet on which the test taker's responses are entered.

**Informal and formal assessment**

**Informal assessment** can take a number of forms starting then with incidental unplanned comments and responses, along with coaching and other impromptu feedback to the student .e.g., marginal comments on papers, responding to a draft of an essay advice about home to better pronounce a work., Nice Job, Good work.

Formal assessments:

*1- are procedures for tapping skills and knowledge.*

*2- are planned sampling techniques and systematic.*

They are systematic, planned sampling techniques constructed to give teacher and student an appraisal of students' achievement. All tests are formal assessment.

**Formative and summative assessment**

Two functions are commonly identified in the literature formative and summative assessment. *'Summative assessment'* aims to measure or summarize what a student has grasped, and typically occurs at the end of a course or unit of instruction.

Most of our classroom assessment is **formative:** evaluating students in the process of "forming" their competencies and skills with the goal of helping them to continue that growth process. The key to such formation is the delivery (by the teacher) and internalization (by the student) of appropriate feedback on performance with an eye toward the future continuation (or formation) of learning.

For all practical purposes, virtually all kinds of informal assessment are formative.

Their primary focus is the ongoing development of the learner's language.

A summation of what a student has learned implies looking back and taking stock of how well that student has accomplished objectives, but does not necessarily point the way to future progress. Final exams in a course and general proficiency exams are examples of summative assessment.

One of the problems with prevailing attitudes toward testing is the view that all tests are summative. A challenge to teacher is to change the attitude among the students.

**Note1**: The explicitness of assessment is associated with summative decisions.

**2.** In the case of assessment *for* learning cultures summative assessments can be used for formative purposes.

**Essential Measurement Qualities**

**Reliability**

Reliability as a quality of test scores would be one which is free from errors of measurement. Here are many factors other than the ability being measured that can affect performance on tests, and that constitute sources of measurement error. Individual's performance may be affected by differences in testing conditions, fatigue, and anxiety. Suppose two raters gave widely different ratings to the same writing sample. Reliability has to do with the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context. In any testing situation, there are likely to be several different sources of measurement error, so that the primary concerns in examining the reliability of test scores are first, to identify the different sources of error, and then to use the appropriate empirical procedures for estimating the effect of these sources of error on tests scores. The identification of potential sources of error involves making judgments based on an adequate theory of sources of error. Determining how much these sources of error affect test scores, on the other hand, is a matter of empirical research.

**Validity**

The most important quality of test interpretation or use is validity, or the extent to which the inferences or decisions we make on the basis of test scores are *meaningful*, *appropriate*, and *useful* so that a test score would be a meaningful indicator of a particular individual's ability. In examining the meaningfulness of test scores, we are concerned with demonstrating that they are unduly affected by factors other than the ability being tested. If test scores are strongly affected by errors of measurement, they will not be meaningful, and cannot provide the basis for valid interpretation or use. A test score that is not reliable, therefore, cannot be valid.

In examining validity, we must also be concerned with the appropriateness and usefulness of the test score for a given purpose. While reliability is a quality of test scores themselves, validity is a quality of test interpretation and use. As with reliability, the investigation of validity is both a matter of judgment and of empirical research, and involves gathering evidence and appraising the values and social consequences that justify specific interpretations or uses of test scores. There are many types of evidence that can be presented to support the validity of a given test interpretation or use, and hence many ways of investigating validity. Neither, is a quality of tests themselves. Neither is an absolute. Determining what degree of relative reliability or validity is required for a particular test context thus involves a value judgment on the part of the test user.

**Properties of Measurement Scales**

Unlike physical attributes, we cannot directly observe intrinsic attributes or abilities. The scales we define can be distinguished in terms of four properties: *distinctiveness, ordered in magnitude, equal intervals, and absolute zero point.* That is, the way we interpret and use scores from our measures is determined through them.

**Nominal scale**

A nominal scale comprises numbers that are used to 'name' the classes or categories of a given attribute. The distinguishing characteristic of a nominal scale is that while the categories to which we assign numbers are distinct, they are not ordered with respect to each other. Nominal scales thus *possess the property of distinctiveness,* because they quantify categories, nominal scales are also sometimes referred to as 'categorical' scales. A special case of a nominal scale is a dichotomous scale, in which the attribute has only two categories, such as 'sex'.

**Ordinal scale**

An ordinal scale comprises the numbering of different level of an attribute that are ordered with respect to each other. The points, or levels, on an ordinal scale can be characterized as 'greater than' or 'less than' each other, and ordinal scales thus possess, in addition to the property of distinctiveness, the property of ordering. The use of subjective ratings in language tests is an example of ordinal scales.

**Interval scale**

An interval scale is a numbering of different levels in which the distances, or intervals, between the levels are equal. Interval scales thus possess the properties of distinctive, ordering, and equal intervals.

The test scores indicate that these individuals are not equally distant from each other on the ability measured. This additional information is not provided by the rankings, which might be interpreted as indicating that the intervals between these five individuals' ability levels are all the same.

**Ratio scale**

None of the scales discussed thus far *has an absolute zero point*, which is the distinguishing characteristic of ratio scale. Each of the four properties provides a different type of information, and the four measurement scales are thus ordered, with respect to each other, in terms of the amount of information they can provide. These different scales are also sometimes referred to as *levels of measurement*. The nominal scale is thus the lowest type of scale, or level of measurement, since it is only capable of distinguishing among different categories, while the ratio scale is the highest level, possessing all four properties and thus capable of providing the greatest amount of information.

**Characteristics that limit measurement**

The most valuable basis for keeping this clearly in mind can be found, in an understanding of the characteristics of measures of mental abilities and the limitations these characteristics place on our interpretation of test scores. These limitations are of two kinds: limitations in specification and limitations in observation and quantification.

**Limitations in specification**

The performance of an individual will be affected by a large number of factors, such as the testing context, the type of test tasks required, and the time of day, as well as her mental alertness at the time of the test, and her cognitive and personality characteristics. The most important factor that affects test performance, with respect to language testing is the individual's language ability, since it is language ability in which we are interested. In order to measure a given language ability, we must be able to specify what it is, and this specification generally is at two levels. *At the theoretical level*, we can consider the ability as a type, and need to define it so as to clearly distinguish it *both* from other language abilities and from other factors in which we are not interested. At the theoretical level we need to specify the ability in relation to, or in contrast to, other language abilities and other factors that may affect test performance. *At the operational level*, we need to specify the instances of language performance that we are willing to interpret as indicators. This level of specification defines the relationship between the ability and the test score, between type and token. When we design a test, we cannot incorporate all the possible factors that affect performance. From a practical points of view, it means there are virtually always more constructs or abilities involved in a given test performance than we are capable of observing or interpreting. We are simplifying, or underspecifying the factors that affect the observations we make. Whether the indeterminacy is at the theoretical level of types, and language abilities are not adequately delimited or distinguished from each other, or whether at the operational level of tokens, where the relationship between abilities and their behavioral manifestations is miss pacified, the result will be the same: our interpretations and uses of test scores will be of limited validity. For language testing research, this indeterminacy implies that any theory of language test performance we develop is likely to be underspecified. Measurement theory has developed, to a large extent, as a methodology for dealing with the problem of under specification, or the controlled effects of factors other than the abilities in which we are interested.

**Limitations in observation and quantification**

These derive from the fact that all measures of mental ability are necessarily *indirect, incomplete, imprecise, subjective*, and *relative*.

***Indirectness***

In the majority of situations where language tests are used, we are interested in measuring the test taker's underlying competence, or ability, rather than his performance on a particular occasion. This is particularly critical since the term 'direct test' is often used to refer to a test in which performance resembles 'actual' or 'real-life' language performance.

Bachman and Palmer explain the features of the *relationship between input and response* in the following terms: *Reactivity and Reciprocal*. So, the relationship between input and response in a test of speaking in which the candidate gives his or her opinion of a recent event is broad and indirect.

***Incompleteness***: ➔1. Measuring students' observation of a sample of total performance.
➘ 2. Interpreting the results with reference & a group performance.

A different approach would be to identify critical features, or components of language ability and then design test tasks that include these. This is the approach that underlies so-called 'discrete-point' language tests.

The approach we choose in specifying criteria for sampling language use on tests will be determined, to a great extent, by how we choose to define what it is we are testing. Therefore, in interpreting individual test scores we must recognize that they are but estimates of ability based on incomplete samples of performance, and that both their reliability and the validity of their interpretation and use will be limited accordingly.

***Imprecision***

The accuracy or precision of our measurements is a function of both the representativeness and the number of tasks or units with which we define our scales.

The precision of scales defined by the number of tasks successfully completed will depend upon the number of tasks or items that constitute the units of the scale, with large numbers of items generally yielding more representative samples of performance. Equally important in the precision such a scale is the comparability of these tasks. We can determine the comparability of tasks, in terms of difficulty, from empirical data from trial administrations of the test.

***Subjectivity***

It consists of

- *Test developer*: in the design of test and selecting the specific ability to test.
- *Test writer*: in producing the test
- *Test taker*: in taking the test
- *Scorers*: in scoring the test
- *The procedures in scoring*:
    A. the decision on the type of the test
    B. the interpretation on the level of the ability

All of them affect reliability and validity

Pilliner (1968) noted that languages are subjective in nearly all aspects. More info: Hossein _salarian@yahoo.com.

***Relativeness***

The last limitation on measures of language ability is the potential relativeness of the levels of performance or ability we wish to measure. The concept of ‘zero’ language ability is a complex one, since in attempting to define it we must inevitably consider language ability as a cognitive ability.

At the other end of the spectrum, the individual with absolutely complete language ability does not exist. In addition to differing norms across varieties of a given language, test developers must consider differences in norms of usage across registers. Finally, test developers must consider differences between ‘prescriptive’ norms and the norms of actual usage. The other approach to defining language test content, that of identifying components, or abilities, provides a means for developing measurement scales that are not dependent upon the particular domain of performance or language users. Such scales can be defined in terms of abilities, rather than in terms of actual performance or individuals, and thus provide the potential for defining absolute ‘zero’ and ‘perfect’ points.

**Steps in Measurement**

A major concern of language test development is to minimize the effects of these limitations. To accomplish this, the development of language tests needs to be based on a logical sequence of procedures linking the putative ability, or construct, to the observed performance.

***Defining constructs theoretically***

- *Physical characteristics*: experienced directly & defined by direct comparison.
- *Mental characteristics*: obtained by inferring abilities from observing behavior.

**Note**:

Approaches for defining language ability/ performance:

1. *real-life approach*: - a domain of actual/ real life language use
   - Language proficiency is defined & distinct scale points are defined in terms of this domain.
2. *Interactional/ability approach:* defined in terms of components of language ability.

1. Whichever approach is followed, domains of real-life or components abilities, definitions must be clear and unambiguous.
2. The definitions upon which the tests are based must also be acceptable to test users.

***Defining constructs operationally***

The second step in measurement, defining constructs operationally, enables us to relate the constructs we have defined theoretically to our observations of behavior. This step involves, determining how to isolate the construct and make it observable. We must therefore decide what specific procedures, or operations, we will elicit the kind of performance that will indicate the degree to which the given construct is present in the individual. The theoretical definitions itself will suggest relevant operations.

The context in which the language testing takes place also influences the operations we would follow.

*Quantifying observations*

The third step in measurement is to establish procedures for quantifying or scaling our observations of performance. The primary concern in establishing scales for measuring mental abilities, therefore, is defining the units of measurement.

**Relevance of steps to the development of language tests**

These general steps in measurement provide a framework both for the development of language tests and for the interpretation of language tests results, in that they provide the essential linkage between the unobservable language ability or construct we are interested in measuring and the observation of performance.

In a different context, the theoretical definition might be made operational in a different way. These steps in measurement are relevant to the development of achievement tests, where the learning objectives of the syllabus constitute the theoretical definitions of the abilities to be tested. In determining operational procedures for testing, both the context of learning and the teaching / learning activities employed need to be considered. By using testing techniques that are similar to activities used for learning, the test developer will minimize the possible negative bias of test method, since students will be expected to perform familiar tasks on the

test. In developing a language proficiency test, the test developers does not have a specific syllabus and must rely on a theory of language proficiency for providing the theoretical definitions of the abilities to be measured. One could use a 'far out' approach simply to assure that familiarity with the testing procedure does not favor some test takers over others. This would raise a different problem, however, in that the 'far our' approach may seriously disadvantage all test takers.

**Relevance of steps to the interpretation of test results**

The first step, defining constructs theoretically, provides the basis for evaluating the validity of the uses of test scores. The definition of the content domain thus provides a means for examining the content relevance of the test.

The second steps, defining constructs operationally, is also related to test validity, in that the observed relationships among different measures of the same theoretical construct provide the basis for investigating concurrent relatedness. The appropriateness of our operational definitions, or testing methods, will also affect the authenticity of the test tasks, and the way test is perceived by test takers and test users.

Finally, the third step, how we quantify our observations, is directly related to reliability.

**Notes:**

1. The distinction is sometimes made between 'examinations' and 'tests'. As Pilliner (1968) Pointed out, there is no consensus on what the distinction is.

Sometimes the distinction is made in terms of time allowed – a typical' examination' lasts two, three, or more hours; a typical 'tests' one half to one hour…

Or the distinction may be hierarchical. A university professor 'examines' his students …; a primary school teacher 'tests' her nine-year olds. Finally, the distinction may depend on whether assessment is; subjective' or 'objective'.

2. The inclusion of subjective measurement procedures such as the oral interview and the composition in the category of tests is different from Raatz's (1981) argument that oral interviews are no tests, primarily because they are not objective

3. It should be noted that the decision-making view of evaluation is not universally held.

**Approaches to Language Testing: A Brief History**

Historically, language- testing trends and practices have followed the shifting sands of teaching methodology. The first stage/ heresy is called 'The Garden of Eden', 'the pre-scientific era' and the examination was based on the traditional, essay-based, native-speaker language syllabus including an English literature paper. The second is psychometric/structuralist era and the third, is integrative era.

**Discrete-Point and Integrative Testing**

They were debated in the 1970s and early1980. These approaches still prevail today, even if in mutated forms.

**Discrete point tests** are constructed on the assumption that language can be broken down into its component parts and that those parts can be tested successfully. These components are the skills of listening, and various units of language. Such an approach demanded a decontextualization that often confused the test taker. Oller (1979) argued that language

competence is a unified set of interacting abilities that cannot be tested separately. His claim was that communicative competence is so global and requires such integration.

Two types of tests have historically been climbed to be examples of integrative tests: cloze test and dictations. A **cloze test** is a reading passage in which roughly every sixth or seventh word has been deleted, the test –taker is required to supply words that fit into those blanks. It is claimed that cloze test results are good measures of overall proficiency. According to theoretical constructs underlying this claim, the ability to supply appropriate words in blanks requires a number of abilities that lie at the heart of competence in a language .knowledge of vocabulary, grammatical structure, discourse structure, reading skills and strategies, and an internalized 'expectancy grammar' (enabling one to predict an item that will come next in a sequence ).

**Dictation** is a familiar language- teaching technique that evolved into a testing technique. The listening portion usually has three stages: an oral reading without pauses, an oral reading with long pauses between every phrase (to give the learner time to write and a third reading at normal speed to give test-takers a chance to check what they wrote.

Dictation is an integrative test because it taps into grammatical and discourse competencies required for other modes of performance in a language. Dictation testing is usually classroom-centered since large-scale administration of dictations is quite impractical from a scoring standpoint. Reliability of scoring criteria for dictation tests can be improved by designing multiple-choice or exact-word cloze test scoring.

Proponents of integrative test methods centered their arguments on **unitary trait hypothesis,** which suggested an indivisible, view of language proficiency: that vocabulary, grammar, phonology, and the four skills. It is contended that there is a general factor of language proficiency such that all the discrete points do not add up to that whole. Others argued strongly against the unitary trait position.

Farhady (1982) found significant and widely varying differences in performance on an ESL proficiency test, depending on subjects, native country, major field of study, and graduate versus undergraduate status. Farhady's contentions were supported in other research that seriously questioned the unitary trait hypothesis. Finally, in the face of the evidence Oller retreated from his earlier stand and admitted that the unitary trait hypothesis was wrong (1983, p. 352).

**Communicative Language Testing (CLT)**

By the mid-1980s the language- testing field had abandoned arguments about the unitary trait hypothesis and had begun to focus on designing communicative language-testing tasks. Bachman and Palmer include among "fundamental principles" of language testing the need for a correspondence between language test performance and language use:

Weir (1990, p.6) noted, Integrative tests such as cloze only tell us about a candidate's linguistic competence. They do not tell us anything directly about a student's performance ability. And so a quest for authenticity was launched.

Bachman and palmer (1996) also emphasized the importance of **strategic competence** (the ability to employ communicative strategies to compensate for breakdowns as well as to enhance the rhetorical effect of utterances) in the process of communication. Communicative testing presented challenges to test designers. Test constructors began to identify the kinds of

real-world tasks that language learners were called upon to perform. It was clear that the contexts for those tasks were extraordinarily wieldy varied and that the sampling of tasks for any one assessment procedure needed to be validated by what language users actually do with language. And, the assessment field became more and more concerned with the authenticity of the tasks and the genuineness of the texts.

An outgrowth of the communicative language movement of the 1970s, language for specific (LSP) testing arose out of the practical need to assess individuals' abilities to perform specific tasks in academic and professional settings. This historical review traces the evolution of LSP testing in the language testing literature, focusing specifically on theory and research in two key areas: (a) authenticity, (b) the interaction between language knowledge and background knowledge, and (c) specificity of content**.**

**Performance- Based Assessment**

It involves oral production, written production, open-ended responses, integrated performance group performance and other interactive tasks. Such assessment is time-consuming and therefore expensive. In technical terms, higher content validity is achieved because learners are measured in the process, e.g., performing the targeted linguistic acts. In an English language-teaching context, performance-based assessment means that you may have a difficult time distinguishing between formal and informal assessment.

A characteristic of many (but not all) performance-based language assessments is the presence of interactive tasks. In such cases the assessment involves learners in actually performing the behavior that we want to measure. In interactive tasks, test-takers are measured in the act of speaking, requesting, responding or in combining listening and speaking, and in integrating reading and writing. Paper-and –pencil tests certainly do not elicit such communicative performance.

A prime example of an interactive language assessment procedure is an oral interview in which language elicited and volunteered by the student can be personalized and meaningful and the tasks can be the authenticity of real- life language use.

**Current Issues in Classroom Testing**

***New views on intelligence***

Intelligence was once viewed strictly as the ability to perform (a) linguistic and (b) logical-mathematical problem solving. Smartness in general was measured by timed discrete- point test. However, research on intelligence by psychologists has begun to turn the psychologists.

Gardner 1983_1999 for example, extended the traditional view of intelligence to seven different components. But he included five other frames of mind in his theory of multiple intelligences:

…Spatial intelligence (the ability to find your way around an environment to form mental images of reality)

.. Musical intelligence (the ability to perceive and create pitch and rhythmic patterns)

.. Bodily kinesthetic (fine motor movement, athletic prowess)

.. Interpersonal intelligence (the ability to understand others and how they feel and interact with them)

…Intrapersonal intelligence (the ability to understand oneself and to develop a sense of self-

identity)

Robert Sternberg (1988, 1997) also charted creative thinking and manipulative strategies as part of intelligence. All smart people aren't necessarily adept at first, reactive thinking. They may be very innovative in being able to think beyond the normal limits imposed by existing tests.

Daniel Goleman's (1995) concept of EQ (emotional quotient) has spurred us to underscore the importance of the emotions in our cognitive processing. Those who manage their emotions that can be detrimental – tend to be more capable of fully intelligent processing.

These new conceptualization of intelligence have not been universally accepted by the academic community. Couple with parallel educational reforms at the time they helped to free us from relying exclusively on timed, discrete-point, analytical tests in measuring language.

**Traditional and 'Alternative' Assessment**

Traditional assessments are one-shot, standardized exams, timed, MC-format, decontextualized, norm-referenced, non-interactive, summative, product-oriented, and foster extinctive motivation. Alternatives are more authentic in their elicitation of meaningful communication. They are continuous long-term assessment, untimed, and have individualized feedback and washback, formative, process-oriented, interactive and foster intrinsic motivation. It is difficult to draw a clear line of distinction between what Armstrong (1994) and Baily (1998) have called traditional and alternative assessment. As Brow and Hudson (1998) aptly pointed out, the assessment traditions available to us should be valued and utilized for the functions that they provide.

**Computer-Based Testing**

It is a burgeoning of assessment in which the test- taker performs responses on a computer. Some computer-based tests (also known as adaptive test "computer-assisted "or" web-based " tests) are *small-scale "home-grown"* tests available on websites. Others are standardized, large-scale tests in which thousands or even tens of thousands of test-takers are involved. Students receive prompts or probes in the form of spoken or written stimuli from the computerized test and are required to type their responses.

Almost all computer-based test items have fixed, closed-ended responses, however, test like the Test of English as a Foreign Language offer a written assay section that must be scored by humans (as opposed to automatic , electronic , or machine scoring ).

A specific type of computer-based test is a **computer – adaptive test (CAT).** Each test-taker receives a set of questions that meet the test specifications and that are generally appropriate for his or her performance level. The CAT starts with questions of moderate difficulty. As test-takers answer each question, the computer scores the question. As long as examinee responds correctly, the computer typically brings questions of lesser or equal difficulty. The computer is programmed to fulfill the test design as it continuously adjusts to find questions of appropriate.

Computer-based testing, with or without CAT technology, offers these advantages:

..Classroom -based testing

..Self-directed testing on various aspects of a language.

.. Practice for upcoming high- stakes standardized tests.

.. Some individualization in the case of CATs

.. Large-scale standardized

*Some disadvantages* are
- Lack of security and possibility of cheating are inherent in classroom-based. "Home – grown" quizzes that may be mistaken for validated assessments.
- The multiple-choice (MC) format preferred for most computer-based test contains the usual penitential for flawed item design.
- Open-ended responses are less likely to appear.
- The human interactive element is absent.

Anyway, by using technological innovations creatively, testers will be able to enhance authenticity, to increase interactive exchange, and promote autonomy.

## *Tests:*

**1. All of the followings are the distinguishing features of measurement, EXCEPT …… .**

*a. quantification* *b. indirectness* *c. characteristics* *d. explicit procedures*

**2. Proponents of …… test methods centered their arguments on Unitary Trait Hypothesis.**

*a. communicative* *b. integrative* *c. structuralist* *d. pre-scientific*

**3. The accuracy or precision of measurements is a function of …….. .**

*a. number of tasks or units with which we define our scales*
*b. explicit rules and procedures*
*c. defining the construct operationally*
*d. both the representativeness and the number of tasks or units with which we define our scales*

**4. Which scale is only capable of distinguishing among different categories**?

*a. nominal* *b. ratio* *c. ordinal* *d. interval*

**5. Which of the following features is Not related to the complexity of evaluation theory and practice?**

*a. the question of definition*
*b. the perspectives on evaluation research*
*c. degrees of ability associated with tasks*
*d. many accounts of evaluation do not reach the public domain*

**6. …… provides a means for evaluating how the assessee processes or benefits from some type of intervention during the course of evaluation.**

*a. Dynamic assessment* *b. Collaborative psychological assessment*
*c. Therapeutic psychological assessment* *d. Psychological test*

**7. For …….. , reference is made to what is called the psychometric soundness of a test.**

*a. format*
*b. administration procedures*
*c. interpretation procedures*
*d. technical quality*

**8. All of the followings are the heresies of language testing research, EXCEPT…….. .**

*a. language heresy* *b. the testing heresy*

*c. the test delivery heresy* *d. the research and development heresy*

**9. The research and development heresy considers the following factor(s):**

*a. test analysis* *b. wash back and impact* *c. test delivery* *d. both 'a' and 'c'*

## *Answer Key:*

**1. b**

Measurement definition includes three distinguishing features: quantification, characteristics, and explicit rules and procedures.

**2. b**

Proponents of integrative test methods centered their arguments on **Unitary trait hypothesis,** which suggested an indivisible, view of language proficiency:

**3. d**

The accuracy or precision of our measurements is a function of both the representativeness and the number of tasks or units with which we define our scales.

**4. a**

The nominal scale is thus the lowest type of scale, or level of measurement, since it is only capable of distinguishing among different categories,

**5. c**

Three features of evaluation theory and practice illustrate the complexity of these developments and the difficulties inherent in the task of mapping achievements and directions.

**6. a**

Dynamic assessment provides a means for evaluating how the assessee processes or benefits from some type of intervention (feedback, hints, instruction, therapy, etc.) during the course of evaluation.

**7.d**

Psychological tests and other tools of assessment may differ with respect to a number of variables such as content, format, administration procedures, scoring and interpretation procedures, and technical quality. Tests differ with respect to their technical quality. More commonly, reference is made to what is called the psychometric soundness of a test.

**8. c**

The THREE heresies of language testing research are: The language heresy, the testing heresy and the research and development heresy.

**9. d**

Here are the two issues of test analysis, and of test delivery.

# *Chapter 2*

## Uses of Language Tests:

*Uses of language tests in educational programs*
*Types of decisions*
*Research uses of language tests*
*Features for classifying different types of language test*
*Tests*

**Uses of Language Tests**

The fundamental use of testing in an educational program is to provide information for making decisions, that is, for evaluation. Evaluation comprises essentially two components: (1) information, and (2) value judgments, or decisions. The information relevant to evaluation can be either qualitative (non-measurement) or quantitative (measurement). Qualitative information can be obtained from observations in a wide variety of ways, including performance checklists and observation schedules, as well as from narrative accounts of class performance of student self-reports.

Quantitative information can include measures such as class rank, teacher ratings, and self-ratings, as well as tests.

In short:

* In educational program:

Most importantly: evaluation ➔ Information ➔ quantitative (measurement)
↘ ↘ Qualitative (non-measurement)
Making ↘
Decision obtained from:
• Observation
• Performance Checklist
&
• Narrative accounts

* In research:
• Basic
•applied

"Quantitative" information obtained from:
• Class rank
• Teacher ratings
• Self-ratings
• Tests

**Assumptions and considerations**

1. Effectiveness: accountability & feedback
2. Improving learning & teaching
3. Educational outcomes are measurable.

**Types of decisions**

• About individuals (micro-evaluation) ➔ students: selection/ placement/ diagnosis/ progress ↘ teachers
• About programs (macro-evaluation)

**Decisions about teachers**

We also need to make about teachers in most educational programs. The decision to hire a given individual as a teacher, will depend on a wide range of information, some of which may be obtained from tests. It is particularly important, in this regard, to recognize that the proficiency required of a language teacher may be both quantitatively and qualitatively different from that of the students. A language teacher must be proficient in not only the complete range of language abilities (grammatical, textual, illocutionary, and sociolinguistic competence), but must also command a wide variety of pedagogically oriented communication strategies and metalinguistic competencies.

**Decisions about programs**

Information from language tests can also be useful in making decisions about programs. In developing a new program, we will be concerned with evaluating specific components in terms of their appropriateness, effectiveness, and efficiency, so as to make improvements that will maximize these characteristics. For purposes of the formative evaluation of programs, where the focus is on providing information that will be useful for making decisions about a program while it is under development, achievement tests that are based on the content of the syllabus are appropriate.

If we are interested in summative evaluation of program, in which the focus is on whether our program is better than other comparable programs, or whether it is the 'best' program currently available, achievement tests by themselves will be of limited use. This is because such tests will provide no information about whether the students have learned skills and abilities beyond those stated as program objectives. For purposes of summative evaluation, therefore, it is often necessary to obtain measures of language proficiency in addition to information on student achievement of syllabus objectives.

**Some programs for decisions**

ENTER ➔ INSTRUCTION ➔ EXIT

FIGURE (1) PROGRAM 1

Characteristics:
* No decision
* No test

Problems for the above characteristics:
* No indication of the appropriacy of the program for all learners
* No feedback about students' learning

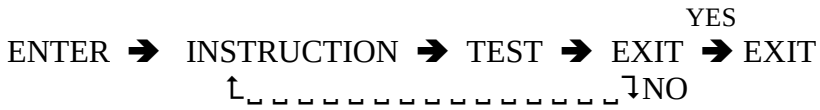



Figure (2) program 2

*Characteristics:*
* Providing feedback
* Solving the problem of program 1
* E.g. achievement test

*Problems:*
1. There is nothing for those who pass the test
2. Failing in addressing program appropriacy

YES

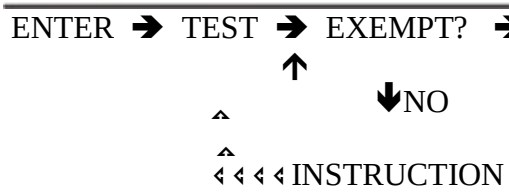



FIGURE (3) PROGRAM 3

- *A solution to program appropriateness*

One approach to developing tests for this program would be to develop achievement tests. The placement test could be a multi-level achievement test based on the objectives of all three levels, while the tests at the end of each level would focus on the objectives of that level.

This program is not intended as a 'model' program to be emulated. This program could be altered in a number of ways to meet different needs and situations.

Programs such as those described in these examples, then, illustrate the fundamental consideration regarding the use of tests in educational programs: the amount and type of testing we do depends on the number any kinds of decisions to be made.

**Research uses of language tests**

1. *Into the nature of language proficiency*
2. *Into the nature of language processing*
3. *Into the nature of language acquisition*
4. *Into the language attrition*
5. *The investigation of the effect of different instructional setting and techniques on language acquisition.*

As operational definitions of theoretical constructs, language tests have a potentially important role in virtually all research, both basic and applied, that is related to the nature of language proficiency, language processing, language acquisition, language attrition, and language teaching. The question of whether language proficiency is a single unitary competence or whether it is composed of distinct component traits researchers for several years, and which also has implications for the theory of language acquisition and for language teaching.

Much current *research into the nature of language proficiency* has now come to focus on identifying and empirically verifying its various components. Of particular interest in this regard are models of communicative competence, which have provided the theoretical definitions for the development of tests of constructs such as sensitivity to cohesive relationships, discourse organization, and differences in register. Such tests in turn provide the basis for verifying these theoretical models. This research involves the construct validation of language tests.

Language tests can also be used in *research into the nature of language processing*. Responses to language tests can provide a rich body of data for the identification of processing errors and their explanation, while language testing techniques can serve as elicitation procedures for collecting information on language processing. In the investigation of how individuals process information in a reading passage, for example, the cloze would seem to have a great deal of potential. Through careful observation and analysis of subjects' response patterns, such as the order in which they complete the blanks and the changes they make in their answers as they work through the passage, we may begin to be able to test some of the hypotheses that are suggested by various theories of reading.

A third research use of language tests is *in the examination of the nature of language acquisition*. Studies of language acquisition often require indicators of the amount of language acquired for use as criterion or dependent variables, and these indicators frequently include language tests. Several studies have used tests of different components of communicative language ability as criteria for examining the effect of learner variables such as length of residence in country, age of first exposure to the target language, and motivational orientation on language acquisition. Language tests are also sometimes used as indicators of factors related to second language acquisition, such as language aptitude and level of proficiency in the native language. Gardner et al. (1983, 1985b), for example, used measures of attitudes, motivational intensity, and prior language achievement to examine a model of language acquisition.

Although *language attrition, or loss*, is not simply the reverse of language acquisition, many of the same factors that have been examined with respect to language acquisition are also hypothesized to affect language attrition, and language tests also have a role to play in this area of research. Oxford (1982) and Clark (1982), for example, both discuss the role of language tests in research on language attrition, as well as considerations for their use in such research. Furthermore, it is clear from both Gardner's (1982) review of the research on social factors in language retention and his own research on attrition that language tests play a vital role in such research.

A fifth area of research in which language tests play an important role is *in the investigation of effects of different instructional settings and techniques* on language acquisition. As well as the more recent large-scale study of bilingual proficiency conducted by the Modern language centre of the Ontario Institute for Studies in Education (Allen et al. 1982, 1983; Harley et al. 1987). Language tests have also provided criterion indicators of language ability for studies in classroom-centered second language acquisition, and for research into the relationship between different language teaching strategies and aspects of second language competence.

**Features for classifying different types of language test**

1. For educational program ➔ according to the type of decisions
2. In research:
   - For comparing performance of individuals, based on:
     - » different characteristic
     - » in different conditions
   - For testing hypotheses about the nature of language.

*Content of language tests can be based on:*

- A theory of language proficiency = proficiency test, language aptitude
- Course syllabus (specific of domain content) = achievement test

The 'content' of language tests can be based on either a theory of language proficiency or a specific domain of content, generally as provided in a course syllabus. We can refer to theory-based tests as proficiency tests, while syllabus-based tests are generally referred to as achievement tests. Whether or not the specific abilities measured by a given proficiency test will depend, of course, on the extent to which the theory upon which the syllabus is based.

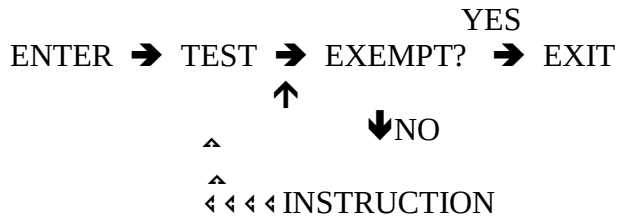
Language aptitude tests are also distinguished according to content. Like language proficiency tests, language aptitude tests are theory-based, but the theory upon which they are based includes abilities that are related to the acquisition, rather than the use of language.

# Tests:

**1. Which of the followings is better for purposes of summative evaluation?**

*a. discrete point test* *b. collaborative test*
*c. proficiency test* *d. placement test*

**2. The figure below ……. .**




*a. is used as a solution to program appropriateness*
*b. shows no indication of the appropriacy of the program for all learners*
*c. is used for demonstrating no decision in a program*
*d. shows coaching*

**3. Which statement is wrong?**

*a. We can refer to theory-based tests as proficiency tests.*
*b. Unlike language proficiency tests, language aptitude tests are theory-based.*
*c. Syllabus-based tests are generally referred to as achievement tests.*
*d. The content of language test can be either theory-based or based on course syllabus.*

**4. All of the followings are correct about features for classifying different types of language test EXCEPT …….. .**

*a. For educational program, it is according to the type of decisions*
*b. In research, it is used for comparing performance of individuals*
*c. For educational program, it is based on the nature of language*
*d. In research, it is used for testing hypotheses about the nature of language*

**5. In developing a new program, we will be concerned with evaluating specific components in terms of their …….. .**

*a. appropriateness* *b. effectiveness* *c. efficiency* *d. all of the above*