

Language Testing

سری کتاب‌های کمک آموزشی کارشناسی ارشد

مجموعه آموزش زبان انگلیسی

Author: Ali Derakhshesh

سرشناسه	Ali Derakhshesh :
عنوان	Language Testing :
مشخصات نشر	تهران : مشاوران صعود ماهان، ۱۴۰۴ :
مشخصات ظاهری	: ۲۹۱ ص
فروست	: سری کتابهای کمک آموزشی کارشناسی ارشد
شابک	: ۹۷۸-۶۰۰-۳۸۹-۰۲۰-۶
وضعیت فهرست نویسی	: فیپای مختصر
یادداشت	: این مدرک در آدرس http://opac.nlai.ir قابل دسترسی است.
شماره کتابشناسی ملی	: ۳۸۷۱۵۸۰



نام کتاب: Language Testing

مؤلف: علی درخشش

ناشر: مشاوران صعود ماهان

نوبت و تاریخ چاپ: دهم / ۱۴۰۴

تیراژ: ۱۰۰۰ نسخه

قیمت: ۵/۲۰/۰۰۰ ریال

شابک: ISBN : ۹۷۸-۶۰۰-۳۸۹-۰۲۰-۶

انتشارات مشاوران صعود ماهان: خیابان ولیعصر، بالاتر از تقاطع مطهری،

روبروی قنادی هتل بزرگ تهران، جنب بانک ملی، پلاک ۲۵۰

تلفن: ۴-۸۸۱۰۰۱۱۳

سخن ناشر

«ن والقلم و ما یسطرون»

کلمه نزد خدا بود و خدا آن را با قلم بر ما نازل کرد.

به پاس تشکر از چنین موهبت الهی، موسسه ماهان درصدد برآمده است تا در راستای انتقال دانش و مفاهیم با کمک اساتید مجرب و مجموعه کتب آموزشی خود برای شما داوطلبان ادامه تحصیل در مقطع کارشناسی ارشد گام موثری بردارد. امید است تلاش‌های خدمتگزاران شما در این موسسه پایه‌گذار گام‌های بلند فردای شما باشد. مجموعه کتاب‌های کمک آموزشی ماهان به منظور استفاده داوطلبان کنکور کارشناسی ارشد سراسری و آزاد تالیف شده‌اند. در این کتاب‌ها سعی کرده‌ایم با بهره‌گیری از تجربه اساتید بزرگ و کتب معتبر داوطلبان را از مطالعه کتاب‌های متعدد در هر درس بی‌نیاز کنیم.

دیگر تالیفات ماهان برای سایر دانشجویان به صورت ذیل می‌باشد.

● **مجموعه کتاب‌های ۸ آزمون:** شامل ۵ مرحله کنکور کارشناسی ارشد ۵ سال اخیر به همراه ۳ مرحله آزمون تالیفی ماهان همراه با پاسخ تشریحی می‌باشد که برای آشنایی با نمونه سوالات کنکور طراحی شده است. این مجموعه کتاب‌ها با توجه به تحلیل ۳ ساله اخیر کنکور و بودجه‌بندی مباحث در هر یک از دروس، اطلاعات مناسبی جهت برنامه‌ریزی درسی در اختیار دانشجو قرار می‌دهد.

● **مجموعه کتاب‌های کوچک:** شامل کلیه نکات کاربردی در گرایش‌های مختلف کنکور کارشناسی ارشد می‌باشد که برای دانشجویان جهت جمع‌بندی مباحث در ۲ ماهه آخر قبل از کنکور مفید می‌باشد. بدین‌وسیله از مجموعه اساتید، مولفان و همکاران محترم خانواده بزرگ ماهان که در تولید و به‌روزرسانی تالیفات ماهان نقش موثری داشته‌اند، صمیمانه تقدیر و تشکر می‌نماییم. دانشجویان عزیز و اساتید محترم می‌توانند هرگونه انتقاد و پیشنهاد درخصوص تالیفات ماهان را از طریق سایت ماهان به آدرس mahan.ac.ir با ما در میان بگذارند.

موسسه آموزش عالی آزاد ماهان

سخن مؤلف

به موازات افزایش تعداد سوالات درس آزمون سازی در کنکور کارشناسی ارشد از سال ۹۳ اهمیت این درس هم افزایش یافته است. در این کتاب سعی شده مطالب مهم کنکوری بطور کامل از کتاب‌های مرجع گردآوری شده و در مجموعه‌ای منسجم در دسترس داوطلبان قرار گیرد. منابع اصلی این درس عبارتند از: FAJAB (فرهادی، جعفرپور و بیرجندی)، Arthur Hughes، James Dean Brown، Douglas Brown و در قسمت مهارت‌ها کتاب‌های Heaton و FAJAB. از آنجایی که کتاب FAJAB مناسب‌ترین چیدمان فصل‌ها را دارد، در تهیه این کتاب نیز، مشابه همان چیدمان در شش فصل اول دنبال شده است. همچنین متن روان کتاب FAJAB باعث شده بیشتر متن کتاب حاضر نیز برگرفته از همین کتاب باشد. با در نظر گرفتن این نکته که در سال‌های اخیر طراحان کنکور توجه بیشتری به کتاب‌های J. Dean Brown، Arthur Hughes و Douglas Brown داشته‌اند، حجم قابل توجهی مطلب هم از این کتاب‌ها به تناسب اهمیت در متن درس گنجانده شده است. مهم‌ترین فصل درس آزمون سازی فصل شش می‌باشد که بیشترین سوالات کنکور هر ساله از این فصل طرح می‌شود. بنابراین، داوطلبان باید توجه ویژه‌ای به این فصل داشته باشند. ویژگی‌هایی که کتاب حاضر را از سایر کتاب‌های موجود در بازار متمایز می‌کند، عبارتند از:

- توضیح روان و کامل مطالبی که پس از تحلیل و بررسی دقیق سوالات کنکور سراسری و آزاد سال‌های اخیر مشخص شده‌اند. با این وجود مطالعه منابع اصلی همچنان می‌تواند درک بهتری به داوطلبان عزیز بدهد. بنابراین توصیه می‌شود، در صورت داشتن زمان کافی، منابع اصلی نیز مطالعه شوند.
- ارائه کامل سوالات کنکورهای سراسری و آزاد از سال ۸۱ به بعد به صورت موضوعی در انتهای هر فصل.
- ارائه مثال‌های کافی جهت فهم بهتر کاربرد فرمول‌ها در فصل‌های چهار، پنج و شش.

Chapter 1: Preliminaries of Language Testing	9
WHY TESTING.....	10
BENEFITS/ IMPORTANCE OF TESTING	10
MEASUREMENT, TEST, EVALUATION	11
ASSESSMENT	13
NORM-REFERENCED vs. CRITERION-REFERENCED TESTS	14
TEACHER-MADE Test vs. STANDARDIZED TESTS.....	16
THE CONSEQUENCES OF STANDARDIZED TESTING	18
WASHBACK	19
TEST BIAS	21
ETHICAL ISSUES: CRITICAL LANGUAGE TESTING	22
AUTHENTICITY	23
State University Questions and Answers	25
Azad University Questions and Answers	31
Chapter 2: Language Test Functions.....	35
TWO MAJOR FUNCTIONS OF LANGUAGE TESTS	36
CONTRASTING CATEGORIES OF LANGUAGE TESTS	42
COMPUTER-ADAPTIVE TESTING	44
A GENERAL FRAMEWORK	45
State University Questions and Answers	47
Azad University Questions and Answers	51
Chapter 3: Forms of Language Test	55
STRUCTURE OF AN ITEM	56
CLASSIFICATION OF ITEM FORMS	57
TYPES OF ITEMS	59
ALTERNATIVE VS. TRADITIONAL ASSESSMENT	66
State University Questions and Answers	67
Azad University Questions and Answers	71
Chapter 4: Basic Statistics in Language Testing.....	73
STATISTICS	74
TYPES OF DATA.....	75
TABULATION OF DATA	77
GRAPHIC REPRESENTATION OF DATA	80
DESCRIPTIVE STATISTICS	80
NORMAL DISTRIBUTION.....	87
DERIVED SCORES	92
CORRELATION.....	96
CORRELATIONAL INDEXES	100
CORRELATIONAL FORMULAS	101
State University Questions and Answers	109
Azad University Questions and Answers	115

Chapter 5: Test Construction	121
DETERMINING FUNCTION AND FORM OF THE TEST	122
PLANNING	123
PREPARING ITEMS.....	123
REVIEWING	130
PRETESTING.....	130
VALIDATION	138
ITEM QUALITY ANALYSIS	139
ITEM DEVELOPMENT	141
ADOPTING LANGUAGE TESTS.....	143
DEVELOPING LANGUAGE TESTS.....	143
ADAPTING LANGUAGE TESTS.....	144
State University Questions and Answers	145
Azad University Questions and Answers	157
Chapter 6: Characteristics of a Good Test	161
RELIABILITY: THE GENERAL CONCEPT	162
RELIABILITY IN TESTING	162
CLASSICAL TRUE SCORE THEORY (CTS)	164
APPROACHES TO ESTIMATING RELIABILITY	167
FACTORS INFLUENCING RELIABILITY	173
STANDARD ERROR OF MEASUREMENT	177
OTHER RELIABILITY THEORIES	180
RELIABILITY OF CRITERION-REFERENCED TESTS	180
VALIDITY	182
FACTORS INFLUENCING VALIDITY	187
THE RELATIONSHIP BETWEEN RELIABILITY AND VALIDITY	188
PRACTICALITY	188
EXTRA POINTS TO REMEMBER.....	189
State University Questions and Answers	191
Azad University Questions and Answers	211
Chapter 7: History of Language Testing	221
GRAMMAR-TRANSLATION APPROACH	222
DISCRETE-POINT APPROACH	223
INTEGRATIVE APPROACH	225
FUNCTIONAL-COMMUNICATIVE APPROACH	227
State University Questions and Answers	230
Azad University Questions and Answers	235
Chapter 8: Cloze and Dictation Type Tests.....	237
CLOZE PROCEDURE	238
VARIETIES OF CLOZE TEST	240
CLOZE TASK	242
SCORING A CLOZE TEST	243
DICTATION	244
VARIETIES OF DICTATION	245
SCORING A DICTATION	247
VALIDITY OF CLOZE DICTATION AND CLOZE.....	247

RELIABILITY OF CLOZE AND DICTATION	248
EXTRA POINTS TO REMEMBER	248
State University Questions and Answers	249
Azad University Questions and Answers	254
Chapter 9: Communicative-Functional Testing.....	255
SELECTION OF THE FUNCTION	256
SOCIAL FACTORS.....	256
THE PERFORMANCE CRITERIA	257
DEVELOPING TEST STEM	257
SCORING SYSTEM.....	259
EXTRA POINTS TO REMEMBER	259
State University Questions and Answers	260
Azad University Questions and Answers	260
Chapter 10: A Sketch of Testing The Four Skills	261
TESTING LISTENING COMPREHENSION.....	263
TESTING ORAL PRODUCTION.....	265
TESTING READING COMPREHENSION	269
TESTING WRITING	271
State University Questions and Answers	276
Azad University Questions and Answers	286
MA 1403 Questions	287
MA 1403 Answers.....	289
REFERENCES	291

Chapter 1

Preliminaries of Language Testing

- ◆ **Why Testing**
- ◆ **Benefits/ Importance of Testing**
- ◆ **Measurement, Test, Evaluation**
- ◆ **Assessment**
- ◆ **Norm-Referenced vs. Criterion-Referenced Tests**
- ◆ **Teacher-Made Test vs. Standardized Tests**
- ◆ **The Consequences of Standardized Testing**
- ◆ **Washback (or Backwash)**
- ◆ **Test Bias**
- ◆ **Ethical Issues: Critical Language Testing**
- ◆ **Authenticity**

Preliminaries of Language Testing

If you hear the word test in any classroom setting, your thoughts are not likely to be positive, pleasant, or affirming. The anticipation of a test is almost always accompanied by feelings of anxiety and self-doubt – along with a hope that you will come out of it alive. Tests seem as unavoidable as tomorrow's sunrise in virtually every kind of educational setting. By all the inconvenience and troubles a test brings, why do we test? What are the benefits of testing?

1. WHY TESTING

Education is the most important enterprise in any society. In fact, a considerable amount of budget, time and energy is put into it every year by government. More than one-fourth of the nation's population attends school. Education is truly a giant and an important undertaking and, therefore, it is crucial that its process and products be evaluated. In fact, evaluation is a major consideration in any education setting:

- Students, teachers, administrators and parents all work toward achieving educational goals and it is quite natural that they want to ascertain the degree to which those goals have been realized. In this sense, testing serves as a monitoring device for learning.
- Government and private sectors which pay teachers and who employ the students afterwards are interested in having precise information about students' abilities.
- Most importantly, through testing, accurate information is obtained based on which **educational decisions** are made (from the entrance exam to the universities to placing students in the right level). When a decision is made, whether the decision is great or small, it should be based on as much and as accurate information as possible. The more accurate the information upon which a decision is made the better that decision is likely to be.

2. BENEFITS/ IMPORTANCE OF TESTING

Tests can benefit students, teachers, and even administrators by confirming progress that has been made and showing how we can best redirect our future efforts. Tests can benefit students in the following ways:

- Testing can **create a positive attitude** toward class and will **motivate** them in learning the subject matter. Tests of appropriate difficulty announced well in advance and covering skills scheduled to be evaluated can contribute to a positive tone, and also create a sense of achievement by demonstrating teacher's spirit of fair play and consistency with course objectives.

- Testing can help students prepare themselves and thus **learn the materials** in three ways. First, learners are helped when they study for exams and again when exams are returned and discussed. Next, where several tests are given, learning can be enhanced by students' growing awareness of the objectives and the areas of emphasis in the course. Finally, tests can foster learning by their diagnostic characteristics; they confirm what each person has mastered, and they point up those language items needing further attention.
- Since tests tend to direct students' learning efforts toward the objectives being measured, they can be used as tools for increasing the **retention and transfer of classroom learning**, if tests are aimed at measuring learning outcomes at the understanding, application, and interpretation levels rather than knowledge level. By including measures of these more complex learning outcomes in our tests, we can direct attention to their importance.
- A major aim of all education is to assist individuals to **understand themselves** better so that they can make more intelligent decisions and can more effectively evaluate their own performance. Periodic testing gives them an insight into the things they can do well and the misconceptions that need correction. Such information provides students with a more objective basis for planning their study program, for selecting future educational experiences, and for developing self-evaluation skills.

Testing can also benefit teachers:

- Testing helps teachers to diagnose their efforts in teaching. It answers the question, "Have I been effective in my instruction?" and therefore testing enables teachers to increase their own effectiveness by making adjustments in their teaching to enable certain groups of students or individuals in the class to benefit more. As we record the test scores, we might well ask the following questions:
Are my lessons on the right level?
Am I aiming my instruction too low or too high?
Am I teaching some skills effectively but others less effectively?
What areas do we need more work on? Which points need reviewing?
- Testing can also help teachers gain insight into ways to improve evaluation process itself:
Were the test instructions clear?
Was everyone able to finish in the allotted time?
Did the test cause unnecessary anxiety or resentment?

3. MEASUREMENT, TEST, EVALUATION

Before we look at tests and test design in second language education, we need to understand three basic interrelated concepts: measurement, test, and evaluation. These terms are sometimes used interchangeably, but some educators make distinctions among them.

MEASUREMENT is the process of quantifying the characteristics of persons according to explicit procedures and rules.

- **Quantification:** Measurement involves the process of assigning numbers, and this distinguishes measures from qualitative descriptions such as a verbal account or visual representation. Non-numerical categories or rankings such as letter grades (A, B, C, ...)

may have the characteristics of measurement because their focus of attention is comparison of testees.

- **Characteristics:** We can assign numbers to both physical and mental characteristics of persons. Physical attributes such as height and weight can be observed directly. In testing, however, we are almost always interested in quantifying mental attributes and abilities, sometimes called *traits* or *constructs*, which can only be observed indirectly. These mental attributes include characteristics such as aptitude, intelligence, motivation, attitude, native language, fluency in speaking, and achievement in reading comprehension.
- **Rules and procedures:** 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 observers, on other contexts and with other individuals. Practically, anyone can rate another person's speaking ability. But while one rater may focus on pronunciation accuracy, another may find vocabulary to be the most salient feature. Such ratings are not considered measurement because the different raters in this case did not follow the same criteria or procedures for arriving at their ratings. Measures are characterized by the explicit procedures and rules upon which they are based. There are many different types of measures in the social sciences, including observations, rankings, rating scales, and tests.

TEST is a measurement instrument or method. Test often connotes the presentation of a set of questions to be answered, to obtain a measure of a characteristic (that is, mental attribute and ability) of a person in a given domain (language, math, etc.). What distinguishes a test from other types of measurement is that it is designed to obtain *a specific sample* of behavior from which one can make inferences about certain characteristics of an individual. Tests are prepared administrative procedures that occur at identifiable times in a curriculum when learners muster all their faculties to offer peak performance. Some tests measure general ability, whereas others focus on very specific competencies of objectives. A multiskill proficiency test determines a general ability level; a quiz on recognizing correct use of definite articles measures specific knowledge. Let's review two examples to illustrate the difference between measurement and test. A qualified interviewer might be able to rate an individual's oral proficiency in a given language according to a rating scale, on the basis of several years' informal contact with that individual, and this could constitute a measure of that individual's oral proficiency. This measure could not be considered a test, however, because the rater did not use an elicitation procedure (e.g., a set of activities or a set of questions) to obtain a specific sample of behavior. Or, the rating of a collection of personal letters based on a rating scale is considered measurement, while asking a person to write an argumentative editorials (to elicit a specific sample of behavior) for a news magazine constitutes a test.

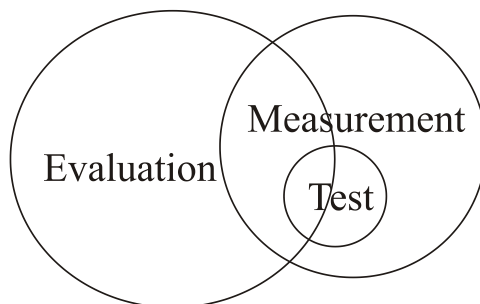
EVALUATION has been defined in a variety of ways:

- 1) The process of delineating, obtaining, and providing useful information for judging decision alternatives.
- 2) The determination of the congruence between performance and objectives.
- 3) A process that allows one to make a judgment about the desirability or value of a measure.

Generally, the purpose of evaluation is to gather information systematically for the purpose of

making decisions. This information need not be exclusively quantitative. Verbal descriptions, ranging from performance profiles to letters of reference, as well as overall impressions, can provide important information for evaluating individuals, as can measures, such as ratings and test scores.

The relationship among measurement, test, and evaluation are illustrated in following figure. As can be seen all tests are measurement but not all tests are evaluation.



Relationship among measurement, test, evaluation

Note: It is important to point out that we never measure or evaluate people. We measure or evaluate characteristics or properties of people such as mental attributes and abilities.

4. ASSESSMENT


ASSESSMENT is appraising or estimating the level or magnitude of some attribute of a person. In educational practice, assessment is an ongoing process that encompasses a wide range of methodological techniques. Whenever a student responds to a question, offers a comment, or tries out a new word or structure, the teacher subconsciously makes an appraisal of the student's performance. Assessment can be classified on two continuums: informal/formal and formative/summative.

4.1. Informal vs. Formal Assessment

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 saying "Nice job!"; "Did you say can or can't?"; or putting a smiley face on some homework.


Informal assessment does not stop there. A good deal of a teacher's informal assessment is embedded in classroom tasks designed to elicit performance without recording results and making fixed conclusions about a student's competence. Informal assessment is virtually always non-judgmental, in that you as a teacher are not making ultimate decisions about the student's performance. Examples at this end of the continuum are marginal comments on papers, responding to a draft of an essay, offering advice about how to better pronounce a word, or suggesting a strategy for compensating for a reading difficulty.

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:** *Is formal assessment the same as a test? We can say that all tests are formal assessments, but not all formal assessment is testing. For example, a systematic set of observations of a student's frequency of oral participation in class is certainly a formal assessment, but it is hardly what anyone would call a test.*

4.2. Formative vs. Summative Assessment

FORMATIVE ASSESSMENTS are given (at the end of a small segment of material) to evaluate 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 of learning. Subsequently, compensatory exercises and activities are provided to help the students fill in the gaps in their learning. Formative tests are either self-graded or no grade is given.

 **Note:** *For all practical purposes, virtually all kinds of informal assessment are (or should be) formative. They have as their primary focus the ongoing development of the learner's language. So when you give a student a comment or a suggestion, or call attention to an error, that feedback is offered to improve the learner's language ability.*

SUMMATIVE ASSESSMENTS are given at the end of a course or unit of instruction and the results are used primarily for assigning course grades, or for certifying student mastery of the instructional objectives. Such tests measure or sum up what the students have learnt from the course.

 **Note:** *Formative test is ongoing and implies the observation of the "process" of learning, while summative test is concerned with the "product" of learning.*

5. NORM-REFERENCED TEST vs. CRITERION-REFERENCED TEST

This distinction refers to different *interpretation of scores*. After a test has been administered and the scores have been computed, the basic issue is how we derive meaning from the scores. To attain interpretive results, two ways of interpretation are identified: "norm-referenced" and "criterion-referenced". When test scores are interpreted in relation to the performance of other testees or a particular group of testees, we speak of a norm-referenced interpretation. If, on the other hand, they are interpreted with respect to a specific level or domain of ability, we speak of a criterion-referenced interpretation.

In **NORM-REFERENCED TESTS (NRT)**, test results may be interpreted with reference to the performance of a given group, or norm. The 'norm group' is typically a large group of individuals who are similar to the individuals for whom the test is designed. In the development of NRTs the norm group is given the test, and then the characteristics, or norms, of this group's performance are used as reference points for interpreting the performance of other students who take the test. In other cases, NR test results are interpreted and reported solely with reference to the actual group taking the test, rather than to a separate norm group. Perhaps the most familiar example of this is what is sometimes called *grading on the curve*, where, say, the top ten percent of the students receive an 'A' on the test and the bottom ten percent fail, irrespective of the absolute magnitude of their scores.

Evaluation may, at times, be simply carried out to determine whether testees have achieved

certain objectives; it is not intended to differentiate testees. **CRITERION REFERENCING**, as this approach is called, focuses on what the testees can do with what they know. Some confusion has developed over the years about what the criterion refers to. This confusion is understandable, because two definitions have evolved for criterion. For some authors, the material that the student is supposed to learn in a course is the criterion against which he or she is being measured. Hence, a basic concern in developing a CRT is that it adequately represents the criterion *ability level* or *content domain* to be evaluated. For other authors, the term criterion refers to the predetermined standard, called a criterion level, against which each student's performance is judged (for instance, if the cut-point for passing a CRT is set at 70%, that is the criterion level). Often but not always, the application of CR test involves the use of cutoff scores that separate competent from incompetent examinees. Usually, a testee passes the test only when he has given the right answer to all or a specific number of test items. Since the purpose of testing is to see if the testee has arrived at a certain mastery, a higher score would make no difference.

Characteristics	Norm-referenced	Criterion-referenced
Type of Interpretation	Relative (A student's performance is compared to those of all other students in <i>percentile</i> terms ¹)	Absolute (A student's performance is compared only to the amount, or <i>percentage</i> ² , of material learned)
Type of Scores Reported	A percentile rank or a standard scores such as z-score, T-score, stanine score, etc.	A statement of whether or not a student has achieved a predetermined percentage or number correct.
Type of Measurement	To measure general language abilities or proficiencies such as reading ability in French, listening comprehension in Chinese, TOEFL	To measure specific objectives-based language points such as mid-term and final exams in schools and language institutes
Purpose of Testing	Spread students out along a continuum of general abilities or proficiencies so that any existing differences among the individuals can be distinguished	Assess the amount of material known or learned by each student, so the focus is on the individuals' knowledge or skills
Distribution of Scores	Normal distribution of scores around the mean	Varies; often non-normal. Students who know the material should score 100%
Test Structure	A few relatively long subtests with a variety of content	A series of short, well-defined subtests with similar item contents
Knowledge of Questions	Students have little or no ideas of what content to expect in test items	Students know exactly what content to expect in test items

¹ Teachers are interested in the student's percentile score, which tells them the proportion of students who scored above and below the student in question.

² Teachers care about the percentage of questions the students answered correctly (or percentage of tasks the students correctly completed) in connection with the material at hand and perhaps in relationship to a previously established criterion level.

Missed Items	When a great number of testees miss an item, it is eliminated from the test	When a test item is missed by a great number of testees, the instructional materials are revised or additional work is given
Example of Test Interpretation	You performed better on this test than approximately 75% of the students in the group against which you are being compared.	You have answered 60% of the items for this unit correctly so you may move on to the next unit.

Distribution of scores. NRTs must be constructed to spread students out along a continuum or distribution of scores, leading to a normal distribution. Such a distribution is desirable so that any existing differences among the students will be clearly revealed. In other words, if there is variation within the group with regard to the knowledge or skill being tested, any differences among students should be reflected in their scores.

In contrast, on a criterion-referenced final examination, all students who have learned all the course material should be able to score 100% on the final examination. Thus, very similar scores can occur on a CRT. As a corollary, in the first week of class, those students who do not know the material (because they have not learned it yet) should all score very low. Again, very similar scores might be produced in such a situation. In short, very similar scores among students on a CRT may be perfectly logical, acceptable, and even desirable if the test is administered at the beginning or end of a course.

Test structure. Typically, an NRT is relatively long and contains a wide variety of different types of question content. Indeed, the content can be so diverse that students find it difficult to know exactly what will be tested. Such a test is usually made up of a few subtests on rather general language skills like reading comprehension, grammar, writing, and so forth.


In contrast, CRTs usually consist of numerous, shorter subtests. Each subtest will typically represent a different instructional objective, and often, each objective will have its own subtest. If a course has twelve instructional objectives, the associated CRT will usually have twelve subtests, although sometimes only a subsample of the objectives will be tested. Because the subtests are often numerous, they must remain short for practical reasons. Sometimes for economy of time and effort, subtests on a CRT will be collapsed together, which makes it difficult for an outsider to identify the subtests. For example, on a reading comprehension test, the students might be required to read five passages and answer four multiple-choice questions on each passage. If on each passage there is one fact question and one inference question, the teachers will most likely consider the two fact questions together as one subtest, and the two inference questions together as another subtest. In other words, the teachers will be focusing on the question types as subtests, not the passages, and this fact might not be obvious to an outside observer.

Knowledge of questions. Students might know what question formats to expect (for example, multiple-choice, true-false), but seldom would the actual language points be predictable. This unpredictability of the question content results from the general nature of what NRTs are testing and the wide variety of question contents that are typically used.

On a CRT, good teaching practice is more likely to lead to a situation in which the students can predict not only the question formats on the test but also the language points that will be tested.

If the instructional objectives for a course are clearly stated, addressed by the teacher, and adequately practiced and learned, then the students should know exactly what to expect on the test. Such statements often lead to complaints that the development of CRTs will cause teachers to “teach to the test,” to the exclusion of other more important ways of spending classroom time. James Dean Brown argues that teaching to the test should nevertheless be a major part of what teachers do. If the objectives of a language course are worthwhile and have been properly constructed to reflect the needs of the students, then the tests that are based on those objectives should reflect the important language points that are being taught. Teaching to such a test should help teachers and students stay on track, and the test results should provide useful feedback on the effectiveness of the teaching and learning processes.

A very useful side effect of teaching to the test is the fact that the information gained can have what Oller termed *instructional value* – that is, to enhance the delivery of instruction in student populations. In other words, such CRT scores can provide useful information for evaluating the effectiveness of the needs analysis, the objectives, the tests themselves, the materials, the teaching, the students’ study habits, and so forth. In short, CRTs can prove enlightening in the never ending evaluation process.

 **Note:** *NRT helps administrators and teachers make program level decisions, such as admission, proficiency and placement decisions, and the other family helps teachers make classroom level decisions (that is, assessing what the students have learned through diagnostic or achievement testing).*

6. TEACHER-MADE TEST vs. STANDARDIZED TEST

In any consideration of educational testing, a distinction must be drawn between “teacher-made” and “standardized” instruments. A **TEACHER-MADE TEST** is a small scale, classroom test which is generally prepared, administered and scored by one teacher. In this situation, test objectives are based directly on course objectives, and test content derived from specific course content. Such tests have the following advantages:

- They measure students’ progress based on the classroom activities.
- They provide an opportunity for the teacher to diagnose students’ weaknesses concerning a given subject matter.
- They help the teacher make plans for remedial instruction, if needed.
- They motivate students.

Characteristics	Teacher-Made Test	Standardized Test
Type of Interpretation	Criterion-referencing	Norm-referencing
Direction for Administration and Scoring	Usually no uniform directions specified	Specific, culture-free direction for every testee to understand; standardized administration and scoring procedures
Sampling of Content	Both content and sampling are determined by classroom teacher	Content determined by curriculum and subject-matter experts; involves extensive investigations of existing

		syllabi, textbooks, and programs; sampling of content done systematically
Construction	May be hurried and haphazard; often no test blueprints, item tryouts, item analysis or revision; quality of test may be quite poor	Uses meticulous construction procedures that include constructing objectives and test blueprints, employing item tryouts, item analysis, and item revisions
Norms	Only local classroom norms are available, i.e. they are determined by the school or a department	In addition to local norms, standardized tests typically make available national schools district norms
Purpose and Use	Best suited for measuring particular objectives set by teacher and for intra-class comparisons	Best suited for measuring broad curriculum objectives and for inter-class, school and national comparisons
Quality of Items	Unknown; usually lower than standardized tests due to limited time and skill of teacher	High; written by specialists, pretested and selected on the basis of effectiveness
Reliability	Unknown; usually high if carefully constructed	High

On the other hand, **STANDARDIZED TESTS** are commercially prepared by skilled test-makers and measurement experts. They provide methods of obtaining samples of behavior under uniform procedures. By a uniform procedure it is meant that the same fixed set of questions are administered with the same set of directions, time restrictions, and scoring procedures. Scoring is usually based on an objective procedure. Such tests have a wide range of coverage that is, they cover more material. They are used to assess either one year's learning or more than one year's learning. Most elementary and secondary schools in the US have standardized achievement tests to measure children's mastery of the standards or competencies that have been prescribed for specified grade levels. College entrance exams such the Scholastic Aptitude Test (SAT®) are part of the educational experience of many high school seniors seeking further education in the US. Examples of standardized language proficiency tests are TOEFL and IELTS.

7. THE CONSEQUENCES OF STANDARDIZED TESTING

Since testing takes place in an educational or social context, we must also consider the educational and social consequences of the uses we make of tests (Messick's *unitary concept* of validity). In fact, the widespread global acceptance of standardized tests as valid procedures for assessing individuals in many walks of life brings with it a set of consequences that fall under the category of consequential validity. **CONSEQUENTIAL VALIDITY** encompasses all the consequences of a test, including such considerations as its accuracy in measuring intended criteria, its impact on the preparation of test-takers, its effect on the learner, and the (intended and unintended) social consequences of a test's interpretation and use. One of the aspects of consequential validity which has drawn special attention is the effect of test preparation courses and manuals on performance. McNamara cautions against test results that may reflect socioeconomic conditions such as

opportunities for coaching, that are “differentially available to the students being assessed (for example, because only some families can afford coaching, or because children with more highly educated parents get help from their parents).”

8. WASHBACK

A facet of consequential validity is washback. Consider the following scenario: you are working in an institution that gets more funding if the number of students reaching a certain standard on the standardized test at the end of the year increases. As a result, at the end of the year, your director will be keeping tabs on how many of your students make the standard for funding. Do you think that would affect your teaching? How much would your teaching change? Would you be more likely to teach material that is related to the test? Material that you know will actually be found on the test? This cluster of issues is about washback. **WASHBACK** (also called *backwash*, *measurement driven instruction*, *curriculum alignment*, *bogwash*) generally refers to the effects the tests have on instruction/pedagogy/learning/education in terms of how students prepare for the test. ‘Cram courses’ and ‘teaching to the test’ are examples of such washback. Another form of washback that occurs more in classroom assessment is the information that washes back to students in the form of useful diagnoses of strengths and weaknesses. Students’ incorrect responses can become windows of insight into further work. Their correct responses need to be praised, especially when they represent accomplishments in a student’s interlanguage. Teachers can suggest strategies for success as part of their coaching role. Washback enhances a number of basic principles of language acquisition: intrinsic motivation, autonomy, self-confidence, language ego, interlanguage, and strategic investment, among others.


Washback can vary along two dimensions: in terms of *degree* (from strong to weak) and in terms of *kind* (positive or negative). The degree and kind of washback depend on: the degree to which the test counters to current teaching practices, what teachers and textbook writers think are appropriate test preparation methods, how much teachers and textbook writers are willing and able to innovate, and the status of the test (and the level of *stakes* involved). The issues of stakes is divided into low stakes versus high stakes situations. Low stakes situations typically involve classroom testing, which is being used for learning purposes or research. For students, high stakes situations usually involve more important decisions like admissions, promotion, placement, or graduation decisions that are directly dependent on test scores. The washback effect is obviously much stronger in high stakes situations than in low stakes situations.

In terms of kind, we have the following definitions:

- **Negative (or harmful) washback:** It is said to occur when test items are based on an outdated view of language which bears little relationship to the teaching curriculum, i.e. when the test content and testing techniques are at variance with the objectives of the course. An instance of this would be where students are following an English course which is meant to train them in the language skills necessary for university study in an English-speaking country, but where the language test which they have to take in order to be admitted to a university does not test those skills directly. If the skill of writing, for example, is tested only by multiple-choice items, then there is great pressure to practice such items rather than practice the skill of writing itself.

- **Positive (or beneficial) washback:** It is said to result when a testing procedure encourages good teaching practice. Take an English medium university in a non-English speaking country where multiple-choice items are administered at the end of an intensive year of English study and the results should be used to determine which students would be allowed to go on to their undergraduate course (taught in English) and which would have to leave the university. If this test is replaced by a test which is based directly on an analysis of the English language needs of first year undergraduate students and which included tasks as similar as possible to those which they would have to perform as undergraduates, the syllabus would be redesigned, new books are chosen, and classes are conducted differently. The result of these changes is that students would reach a much higher standard in English. This is a case of beneficial backwash. As another example, the use of an oral interview in a final examination may encourage teachers to practice conversational language use with their students.

It was once said that the good test is an obedient servant since it follows and apes the teaching. However, it is difficult to agree with this situation. The proper relationship between teaching and testing is surely that of partnership. It is true that there may be occasions when teaching is poor or inappropriate and when testing is able to exert a beneficial influence. We cannot expect testing only to follow teaching. Rather, we should demand of it that it is supportive of good teaching and, where necessary, exerts a corrective influence on bad teaching.

 **Note:** *Backwash can be viewed as part of something more general – the impact of assessment. The term ‘impact’, as it is used in educational measurement, is not limited to the effects of assessment on learning and teaching but extends to the way in which assessment affects society as a whole (see consequential validity).*

A number of suggestions have been made over the years for ways to promote positive washback. The following list is adopted from Brown (2005, p. 254).

Test design

1. Sample widely and unpredictably.
2. Design tests to be criterion-referenced.
3. Design the test to measure what the programs intend to teach.
4. Base the test on sound theoretical principles.
5. Base achievement tests on objectives.
6. Use direct testing.
7. Foster learner autonomy and self-assessment.

Test content

1. Test the abilities whose development you want to encourage.
2. Use more open-ended items.
3. Make examinations reflect the full curriculum, not merely a limited aspect of it.
4. Assess higher-order cognitive skills to ensure they are taught.
5. Use a variety of examination formats, including written, oral, and practical.
6. Do not limit skills to be tested to academic areas.
7. Use authentic tasks and texts.

Logistics

1. Insure that test-takers, teachers, administrators, curriculum designers understand the purpose of the Test.
2. Make sure language-learning goals are clear.
3. Where necessary, provide assistance to teachers to help them understand the tests.
4. Provide feedback to teachers and others so meaningful change can be effected.
5. Provide detailed and timely feedback to schools on levels of pupils performance and areas of difficulty in public examinations.
6. Make sure teachers and administrators are involved in different phases of the testing process because they are the people who will have to make changes.
7. Provide detailed score reporting.

Interpretation/ Analysis

1. Make sure exam results are believable, credible, and fair to test takers and score user.
2. Consider factors other than teaching effort in evaluating published examination results and national rankings.
3. Conduct predictive validity studies of public examinations.
4. Improve the professional competence of examination authorities, especially in test design.
5. Insure that each examination board has a research capacity.
6. Have testing authorities work closely with curriculum organizations and with educational administrators.
7. Develop regional professional networks to initiate exchange programs and to share common interests and concerns.

9. TEST BIAS

It is no secret that standardized tests involve a number of types of test bias. Some of the sources of bias are background knowledge, native language, cultural background, race, gender, age, cognitive characteristics, and learning styles. A test or item can be considered to be **BIASED** if one particular section of the candidate population is advantaged or disadvantaged by some feature of the test or item which is not relevant to what is being measured. An item that is biased against one group of people is testing something in addition to what it was originally designed to test, and such an item cannot provide clear and easily interpretable information. For instance, consider an IQ item where the answer hinges on understanding the differences between the terms rain, snow, sleet, and hail. Such an item might naturally be biased against students who grew up in a tropical area because many of them have never seen anything resembling snow, sleet, or hail. An obvious example of bias is shown by the item below, which appeared in the State Examination of English Language for Elementary Level:

Mia: What should I do with this *martabak*?

Mom: Just put them on a (a) drawer, (b) plate, (c) stove, (d) mug

Examinees that come from Java or are somehow familiar to Indian culture will find the item above easy to answer. Yet, for some others coming from different areas, the word *martabak* may be entirely new and therefore they do not know whether a *martabak* refers to a kind of stationery, food, a cooking device, or a kind of beverage. Despite their excellent English proficiency, there is no way

they can get at the right answer. The item, in other words, is culturally biased against these examinees.

Let examine another example adopted from a listening comprehension item taken from TOEFL Test Preparation Kit Workbook:

Man: I'm taking up a collection for the jazz band. Would you like to give?

Woman: Just a minute while I get my wallet.

(narrator) What will the woman probably do next?

- a. Put some money in her wallet
- b. Buy a band-concert ticket
- c. Make a donation
- d. Lend the man some money

The right answer is c. However, it is very unlikely that examinees of non-Western culture are familiar with the habit of collecting money for a band in the US culture. Therefore, being largely unfamiliar with the meaning of “taking a collection” and looking at the word “give” from the man, they may be misled into thinking that the answer is d. alternatively, if they have no idea whatsoever that a band in the US may need to collect some money, they may choose b.

- **Fairness:** It can be defined as the degree to which a test treats every student the same, or the degree to which it is impartial. Teachers would generally like to ensure that their personal feelings do not interfere with fair assessment of the students or bias the assignment of scores. The aim in maximizing objectivity is to give each student an equal chance to do well. Equitable treatment in terms of testing conditions, access to practice materials, performance feedback, retest opportunities, and other features of test administration, including providing reasonable accommodation for test takers with disabilities when appropriate, are important aspects of fairness under this perspective. This tendency to seek objectivity has led to the proliferation of ‘objective’ tests which minimize the possibility of varying treatment for different students.

10. ETHICAL ISSUES: CRITICAL LANGUAGE TESTING

Shohamy sees the ethics of testing as an extension of what educators call critical pedagogy, or more precisely in this case, **CRITICAL LANGUAGE TESTING**. For a better understanding of critical language testing, we need to know what critical pedagogy is. As language teachers we have to remember that we are all driven by convictions about what this world should look like, how its people should behave, how its governments should control that behavior, and how its inhabitants should be partners in the stewardship of the planet. We embody in our teaching a vision of a better and more humane life. However, critical pedagogy brings with it the reminder that our learners must be free to be themselves, to think for themselves, to behave intellectually without coercion from a powerful elite, to cherish their beliefs and traditions and cultures without the threat of forced change. In our classrooms, where the dynamics of power and domination permeate the fabric of classroom life, we are alerted to a possible covert political agenda beneath our overt technical agenda.

One of the byproducts of a rapidly growing testing industry is the danger of an abuse of

power. As Shohamy claims “Tests represent a social technology deeply embedded in education, government and business; as such they provide the mechanism for enforcing power and control. Tests are most powerful as they are often the single indicators for determining the future of individuals”. Proponents of a critical approach to language testing claim that large-scale standardized testing is not an unbiased process, but rather is the “agent of cultural, social, political, educational, and ideological agendas that shape the lives of individual participants, teachers and teachers”. The issues of critical language testing are numerous:

- Psychometric traditions are challenged by interpretive, individualized procedures for predicting success and evaluating ability.
- Test designers have a responsibility to offer multiple modes of performance to account for varying styles and abilities among test-takers.
- Tests are deeply embedded in culture and ideology.
- Test-takers are political subjects in a political context.

One of the problems of critical language testing surrounds the widespread conviction that standardized tests designed by reputable test manufacturers are infallible in their predictive validity. Universities, for example, will deny admission to a student whose TOEFL score falls one point below the requisite score (usually around 500), even though that student, if offered other measures of language ability, might demonstrate abilities necessary for success in university program. One standardized test is deemed to be sufficient, follow-up measures are considered to be too costly.

A further problem with our test-oriented culture lies in the agendas of those who design and those who utilize the tests. Tests are used in some countries to deny citizenship. Tests are by nature culture-biased and therefore may disenfranchise members of a non-mainstream value system. Test givers are always in a position of power over test-takers and therefore can impose social and political ideologies on test-takers through standards of acceptable and unacceptable items. Tests promote the notion that answers to real-world problems have unambiguous right and wrong answers with no shades of gray. A corollary to the latter is that tests presume to reflect an appropriate core of common knowledge and acceptable behavior; therefore the test-taker must buy into such a system of beliefs in order to make the cut.

Shohamy (1998) pointed out that politicians had capitalized on language tests for tackling thorny political issues that they failed to address by other policy-making process. They could set the benchmark for passing a language test for immigration purposes without any justification, thereby allowing them the flexibility to create immigration quotas. For example, the government of Australia drew on language tests to manipulate the number of immigrants and to determine if refugees could be accepted or rejected. Similarly, Latvia used strict language tests to prevent Russians from obtaining citizenship in the wake of its independence.

11. AUTHENTICITY

AUTHENTICITY is the degree of correspondence of the characteristics of a given language test task to the features of target language use (TLU) tasks³. Essentially, when you make a claim for authenticity in a test task, you are saying that this task is likely to be enacted in the real world. This concept is shown in the following figure.



Many test item types fail to simulate real-world tasks. The sequencing of items that bear no relationship to one another lacks authenticity. One does not have to look very long to find reading comprehension passages in proficiency tests that do not reflect a real-world passage.

In a test, authenticity may be present in the following ways:

- The language in the test is as natural as possible.
- Items are contextualized rather than isolated.
- Topics are meaningful for the learner.
- Some thematic organization to items is provided, such as through a story line or episode.
- Tasks represent, or closely approximate, real-world tasks.

The authenticity of test tasks in recent years has increased noticeably. Reading passages are selected from real-world sources that test-takers are likely to have encountered or will encounter. Listening comprehension sections feature natural language with hesitations, white noise, and interruptions. More and more tests offer items that are “episodic” in that they are sequenced to form meaningful units, paragraphs, or stories.

³ TLU tasks are those tasks which the learner is likely to face in real-life context

State University Questions

- 1- To answer the question ‘Why have a test at all?’ which one of the following do you find irrelevant?** (State University, 81)
- 1) Why does this learner fit in our teaching program?
 - 2) What is the learner’s general level of language ability?
 - 3) How much has the learner learned from a particular course?
 - 4) What are the learner’s particular strengths and weaknesses?
- 2- Backwash effect can be defined as -----.** (State University, 83)
- 1) the influence of testing on teaching
 - 2) the importance of analyzing test results
 - 3) the importance of contrastive analysis to test development
 - 4) the impact of language sub-skills on communication skills
- 3- The purpose of norm-referenced tests is to -----.** (State University, 84)
- 1) measure communicative competence
 - 2) relate one testee’s performance to that of others
 - 3) use objective linguistic norms to measure proficiency
 - 4) classify people in terms of their ability to perform a set of tasks
- 4- Achieving beneficial backwash requires -----.** (State University, 84)
- 1) sampling widely
 - 2) ensuring that the test format is unknown to testees
 - 3) using indirect testing
 - 4) developing proficiency rather than achievement test
- 5- ----- tests are prepared on the basis of instructional objectives to determine to what degree the students have learned the material presented in class.** (State University, 85)
- 1) Criterion-referenced
 - 2) Norm-referenced
 - 3) Proficiency
 - 4) Standardized
- 6- “Backwash effect” refers to the effect of -----.** (State University, 85)
- 1) unsystematic sources of variance on the observed score
 - 2) face validity considerations on test form selection
 - 3) item difficulty on the true score
 - 4) testing on pedagogy
- 7- Formative evaluation -----.** (State University, 85)
- 1) refers to the need for testing students to elicit information
 - 2) is the ongoing evaluation involved in all phases of teaching programs
 - 3) refers to the formal exams administered at the end of teaching programs
 - 4) is intended to check students’ progress in regard to their mastery of linguistic forms
- 8- In a norm-referenced test, -----.** (State University, 86)
- 1) a higher score would make no difference
 - 2) the goal is to select the examinees with the complete mastery of a skill
 - 3) standard scores and percentile ranks show a testee’s relative position
 - 4) the focus is on assuring that testees have achieved certain objectives

- 9- Norm-reference measurement helps us -----.** (State University, 87)
- 1) evaluate the success of an educational program
 - 2) determine the extent to which students have met educational objectives
 - 3) choose the best students to receive a particular type of education
 - 4) determine whether we need to revise our current teaching activities
- 10- Which of the following distinguishes “evaluation” from “testing”?** (State University, 88)
- 1) Decision making
 - 2) Comparison of measures
 - 3) Reliance on numerical values
 - 4) Quantitative procedures used
- 11- A change in testing leading to a change in teaching is known as -----.** (State University, 88)
- 1) washback
 - 2) test facet
 - 3) curricular validity
 - 4) communicative interaction
- 12- “Cram courses” and “teaching to the test” are examples of -----.** (State University, 89)
- 1) test washback
 - 2) test tasks
 - 3) authenticity
 - 4) directed response
- 13- When one designs a test with an eye to its impact on the teaching enterprise, one has technically concerned oneself with -----.** (State University, 90)
- 1) washback
 - 2) proficiency
 - 3) aptitude
 - 4) knowledge
- 14- Norm-referenced tests rely on -----.** (State University, 91)
- 1) course objectives
 - 2) teacher-made items
 - 3) a continuum in rank order
 - 4) giving test-takers feedback on specific lesson objectives
- 15- The claim that “tests are deeply rooted in culture and ideology” is most likely made in -----.** (State University, 92)
- 1) communicative language testing
 - 2) critical language testing
 - 3) integrative language testing
 - 4) task-based assessment
- 16- In ----- tests, each candidate’s score is interpreted relative to the scores of all other candidates who take the test.** (State University, 93)
- 1) criterion-referenced
 - 2) placement
 - 3) norm-referenced
 - 4) aptitude
- 17- Which of the following does NOT represent authenticity in a given test?** (State University, 93)
- 1) Contextualization of test items
 - 2) Ease of scoring the test items
 - 3) Naturalness of the language used in the test
 - 4) Resemblance of test items to real-world tasks
- 18- It is NOT true that a norm-referenced test -----.** (State University, 94)
- 1) measures general language abilities
 - 2) includes a variety of test content
 - 3) is based on what students exactly expect of test question
 - 4) relies on the normal distribution of scores around a mean
- 19- Washback in language testing -----.** (State University, 94)
- 1) can be either summative or formative
 - 2) refers to the effect of testing in large-scale assessment
 - 3) is limited to formative assessment
 - 4) is a feature of consequential validity

20- In interpreting a student's score on a criterion-referenced test, -----. (State University, 95)

- 1) there is no need for any reference to the actual number of test questions the student has answered correctly
- 2) the primary focus is on how much of the material the student has learned in relative terms
- 3) there is no need for any reference to the performances of other students
- 4) the focus must be on the student's percentile rank

21- All of the following statements are TRUE regarding testing, assessment and evaluation EXCEPT -----. (State University, 96)

- 1) all tests are formal assessments, but not all formal assessment is testing
- 2) assessment is usually time-constrained and draws on a limited sample of behavior
- 3) evaluation is a process that allows us to judge the value or desirability of a measure
- 4) a test is a prepared administrative procedure that occurs at an identifiable time in a curriculum

22- Norm-referenced and criterion-referenced tests differ in all of the following characteristics EXCEPT the -----. (State University, 96)

- | | |
|------------------------|---------------------------|
| 1) purposes of testing | 2) type of measurement |
| 3) length of the test | 4) type of interpretation |

23- Which of the following tips does NOT foster beneficial washback? (State University, 96)

- 1) Test the abilities whose development you wish to promote
- 2) Base achievement tests on objective
- 3) Sample widely and unpredictably
- 4) Use indirect testing

24- Percentage and percentile are the terms used to capture the difference between ----- tests, respectively. (State University, 97)

- | | |
|---|---|
| 1) criterion-referenced and norm-referenced | 2) norm-referenced and criterion-referenced |
| 3) direct and indirect | 4) indirect and direct |

25- Authenticity in a test may be present in all of the following ways EXCEPT when -----.

(State University, 97)

- 1) some thematic organization to items is provided
- 2) the items are contextualized rather than isolated
- 3) the difficulty level of the test presents a reasonable level of challenge
- 4) topics are meaningful to test-taker

26- Which of the following are threats to achieving beneficial backwash?

(State University, 98)

- | | |
|---|--|
| 1) Direct and norm-referenced testing | 2) Direct and criterion-referenced testing |
| 3) Indirect and norm-referenced testing | 4) Indirect and criterion-referenced testing |

27- What type of assessment do impromptu student responses to teacher's questions represent?

(State University, 98)

- | | |
|------------------------|----------------------|
| 1) Informal, summative | 2) Formal, formative |
| 3) Informal, formative | 4) Formal, summative |

28- ----- does not necessarily entail testing; rather, it is involved when the results of a test are used for decision making.

(State University, 98)

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

29- Which of the following statements is difficult to agree with?

(State University, 99)

- 1) The proper relationship between teaching and testing is that of partnership.
- 2) A good test is an obedient servant since it follows and apes the teaching.
- 3) A good test is likely to bring about changes in the syllabus and methodology.
- 4) The impact of a test goes beyond learning and teaching and affect society as a whole.

30- What type of assessment is involved in simple observation of students' performance on learning tasks and the study of the portfolios that they have made on their work?

(State University, 99)

- 1) Formative-traditional
- 2) Summative-alternative
- 3) Formative-alternative
- 4) Summative-traditional

31- Which of the following statements is NOT true about criterion-referenced tests?

(State University, 99)

- 1) They are likely to result in beneficial backwash.
- 2) They help students measure their progress based on meaningful standards.
- 3) They are intended to assess the amount of material known, or learned, by each individual student.
- 4) They give the students the feeling that they are less able than most of their fellows and hence are destined to fail.

32- Which of the following is NOT a feature of an authentic test?

(State University, 99)

- 1) It offers tasks that replicate real-world tasks.
- 2) It includes meaningful, relevant, interesting topics.
- 3) It includes contextualized rather than isolated items.
- 4) It includes tasks that can be accomplished within an allotted time limit.

33- All of the following are issues related to critical language testing EXCEPT that -----.

(State University, 99)

- 1) tests are deeply embedded in culture and ideology
- 2) communicating with students, families, and other audiences about student progress is essential
- 3) psychometric traditions are challenged by interpretive procedures, for predicting and evaluating abilities
- 4) test designers have a responsibility to provide various modes of performance to account for different styles and abilities among test takers

34- Which statement is NOT TRUE regarding the notion of "criterion" in CRT? (State University, 1402)

- 1) The notion of criterion in CRT has been well-defined in the relevant literature.
- 2) The material that the student is supposed to learn in a course is the criterion against which he or she is being measured.
- 3) The term criterion refers to the standard, called a criterion level, against which each student's performance is judged.
- 4) If the cut-point for passing a CRT is set at 70%, that is the criterion level.

35- In general, ----- is associated with CRTs and ----- is very much a part of NRT decisions.

(State University, 1402)

- 1) percentage – percentile
- 2) percentile – percentage
- 3) percentage – percentage
- 4) percentile – percentile

36- What is the purpose of "instructional value" according to Oller (1979)?

(State University, 1403)

- 1) To balance the delivery of instruction for student learning
- 2) To enhance the delivery of instruction for student learning
- 3) To balance the delivery of instruction in student populations
- 4) To enhance the delivery of instruction in student populations

State University Answers

1- Choice 1

2- Choice 1

Refer to Section 8.

3- Choice 2

The type of interpretation in a NRT is relative, i.e. a student's performance is compared to those of all other students in percentile terms.

4- Choice 1

Refer to Section 8.

5- Choice 1

The type of interpretation in a CRT is absolute, i.e. a student's performance is compared only to the amount, or percentage, of material learned

6- Choice 4

Refer to Section 8.

7- Choice 2

Refer to Section 4.2.

8- Choice 3

In an NRT the relative position of examinees is reported in terms of percentile rank.
Choice 1, 2 and 4 are all characteristics of CR tests.

9- Choice 3

Choice 3 describes the function of placement test (see chapter 3) which is an NRT.
Choice 1, 2 and 4 are all characteristics of CR tests.

10- Choice 1

Refer to Section 3.

11- Choice 1

Refer to Section 8.

12- Choice 1

Refer to Section 8.

13- Choice 1

Refer to Section 8.

14- Choice 3

Refer to Section 5.

15- Choice 2

Refer to Section 10.

16- Choice 3

Refer to Section 5.

17- Choice 2

Refer to Section 11.

18- Choice 3

In a CRT students know exactly what content to expect in test items.

19- Choice 4

Refer to Section 8.

20- Choice 3

We make a reference to the performances of other students in case of NRTs.

21- Choice 2

Choice two is the definition of test.

22- Choice 3

Refer to the table in Section 5.

23- Choice 4

Refer to the table in Section 8.

24- Choice 1

Refer to the table in Section 5.

25- Choice 3

Refer to Section 11.

26- Choice 3

Refer to the table in Section 8.

27- Choice 3

Refer to Section 4.

28- Choice 1

Refer to Section 3.

29- Choice 2

Refer to Section 8.

30- Choice 3

As an informal assessment, “observation of students’ performance” is formative assessment and all personal response items including portfolio, self-assessment, journal and conference are alternative assessments.

31- Choice 4

The purpose of CRT is to classify people according to whether or not they are able to perform some task or set of tasks satisfactorily. The tasks are set, and the performances are evaluated. It does not matter in principle whether all the candidates are successful, or none of the candidates is successful. The tasks are set, and those who perform them satisfactorily pass; those who don’t, fail. This means that students are encouraged to measure their progress in relation to meaningful criteria, without feeling that, because they are less able than most of their fellows, they are destined to failed.

32- Choice 4

Refer to Section 11.

33- Choice 2

Refer to Section 10.

34- Choice 1

Refer to Section 5.

35- Choice 1

Refer to the table in Section 5.

36- Choice 4

Refer to Section 5.