Ministry of Higher Education andScientific

Research

University of Tikrit

College of Education for Humanities

English Department

STANDARDIZED TESTING

Prof.dr.Nagham Q.Yahya

nagyahya@tu.edu.iq

A good standardized test is the product of a thorough process of empirical research and development that may extend beyond simply establishing standards or benchmarks. Standardization can also mean the use of systematic procedures for administration and scoring. Further, many standardized tests, especially large-scale tests, are norm-referenced, the goal of which is to place test-takers on a continuum across a range of scores and to differentiate test-takers by their relative rankings.

Characteristics of a standardized test

standards-based

product of research and development

systematic scoring and administration procedures

referenced to norms
relative rankings.

Most elementary and secondary schools around the world use standardized achievement tests to measure students’ mastery of the standards or competencies that have been prescribed for specified grade levels, exit requirements, and entrance to further levels. Secondary schools whose requirements for graduation include English language proficiency sometimes institutionalize countrywide standardized tests to measure such ability (Akiyama, 2004).

A further disadvantage is the potential misunderstanding of the difference between direct testing and indirect testing. Some standardized tests include tasks that do not directly specify performance of the target ability.

Advantages	**Disadvantages**
* readily available product	* possibly inappropriate use of such tests
* easily administered to large groups	* potential test biases
* streamlined scoring and reporting procedures	* indirect testing may not elicit a good sample of performance
* a previously validated product (in many cases)	* multiple-choice formats have the appearance or authority

DEVELOPING A STANDARDIZED TEST

If you are a classroom teacher, you are not likely to be in a position to develop a brand-new, large-scale standardized test with a team of test designers and researchers. However, it is a virtual certainty that someday you will be in a position to (a) revise an existing test, (b) adapt or expand an existing test, and/or (c) create a smaller-scale standardized test for a program you are teaching in. Even if none of these three scenarios should ever apply to you, it is of paramount importance that you understand the process of the development of the standardized tests that have become ingrained in our educational institutions.

Questions to consider

* How are standardized tests developed?

* Where do test tasks and items come from?

* How are they evaluated?

* Who selects items and their arrangement in a test?

* How do such items and tests achieve consequential validity?

* How are different forms of tests designed to be of equal difficulty?

* Who sets norms and cutoff limits?

* Are security and confidentiality an issue?

* Are cultural and racial biases an issue in test development?

Five different standardized tests, are used to exemplify the process of standardized test design. The first four tests (TOEFL, IELTS, PTE, and MELAB) which lists a selection of commercially available tests. They are all tests of general language ability or proficiency. The fifth (CMSPT) is a placement test at a university. In the following sections, we illustrate six steps of development using these five tests.

Step 1: Determine the Purpose and Objectives of the Test

Most standardized tests are expected to provide high practicality in administration and scoring without unduly compromising validity. The initial outlay of time and money for such a test is significant, but the test is usually designed for repeated use. It’s therefore important for its purpose and objectives to be stated specifically. The first four tests are designed to evaluate the general English ability of those whose native language is not English, targeting the four skills of listening, speaking, reading, and writing. They are frequently used to help institutions of higher learning make decisions about the English language proficiency of international applicants for admission. As you can see, the objectives of each of these tests are quite clear. The content of each test must be designed to accomplish those particular ends. This first stage of goal-setting might be seen as one in which the consequential validity of the test is foremost in the mind of the developer: each test has a specific gate-keeping function to perform; therefore, the criteria for entering those gates must be specified accurately.

Step 2 : Design Test Specifications

Now comes the difficult part. Decisions need to be made on how to structure the specifications (or specs, as they are popularly called) of the test. Before specs can be addressed, comprehensive research must identify a set of **constructs** underlying the test itself. To illustrate the design of test specs, we focus on the TOEFL. Construct validation for the TOEFL is carried out by the staff at the Educational Testing Service under the guidance of a policy council that works with a committee of examiners, which comprises appointed external university faculty, linguists, and assessment specialists. Dozens of employees are involved in a complex process of reviewing current TOEFL specifications, commissioning and developing test tasks and items, assembling forms of the test, and performing ongoing exploratory research related to formulating new specs. Because the TOEFL is a “proficiency” test, it should be made clear that many assessment specialists prefer the term ability to proficiency and thus speak of **language ability** as the overarching concept (Bachman, 1990; Bachman & Palmer, 1996, 2010), The latter phrase is more consistent, they argue, with our understanding that the specific components of language ability must be assessed separately. Most current views accept the ability argument and therefore strive to specify and assess the many components of language.

After breaking language competence down into subsets of listening, speaking, reading, and writing, each performance mode can be examined on a continuum of linguistic units: phonology (pronunciation) and orthography (spelling), words (lexicon), sentences (grammar), discourse (beyond the sentence level), and pragmatic (sociolinguistic, contextual, functional, cultural) features of language. How does the TOEFL sample incorporate all these possibilities? Oral production tests can test overall conversational fluency or pronunciation of a particular component of phonology and can take the form of imitation, structured responses, or free responses. Listening comprehension tests can concentrate on a particular feature of language or on overall listening for general meaning. Tests of reading can cover the range of language units and can aim to test comprehension of long or short passages, single sentences, or even phrases and words. Writing tests can take on an open-ended form with free composition or be structured to elicit anything from correct spelling to discourse-level competence. The developer must select, on some systematic basis, a subset from the sea of potential performance modes that could be sampled in a test. To make a very long story short (and leaving out numerous controversies), the TOEFL included for many years three types of performance in its organizational specifications: listening, structure, and reading, all of which tested comprehension through standard multiple-choice tasks.

Step 3: Design, Select, and Arrange Test Tasks/Items

Once specifications for a standardized test have been stipulated, the often endless task of designing, selecting, and arranging items begins. The specs act much like a blueprint in determining the number and types of items to be created. In the case of a test like the JELTS (or any validated standardized test), before any such items are released in published form, they are pretested on sample audiences and scientifically selected to meet difficulty specifications within each subsection and section, and on the test overall. Further, those items are also selected to meet a desired discrimination index.

Step 4 : Make Appropriate Evaluation of Different Kinds of Items

The concepts of item facility (IF), item discrimination (ID), and distractor. As the discussion there showed, such calculations provide useful information for classroom tests, but sometimes the time and effort involved may not be practical, especially if the classroom-based test will be administered only once. These indices are, however, essential for a standardized multiple-choice test that is designed for a commercial market, and/or administered a number of times, and/or administered in a different form. For other types of response formats—namely, oral and written responses— different forms of evaluation become important. The principles of practicality and reliability are prominent, along with the concept of facility. Practicality issues in such items include the clarity of directions, timing of the test, ease of administration, and how much time is required to score responses. Reliability is a major player in instances in which more than one scorer is employed and to a lesser extent when a single scorer must evaluate tests over long spans of time, which could lead to deterioration of standards. Facility is also a key to the validity and success of an item type: unclear directions, complex language, obscure topics, fuzzy data, and culturally biased information may all lead to a higher level of difficulty than one desires.

Step 5: Specify Scoring Procedures and Reporting Formats

A systematic assembly of test items in preselected arrangements and sequences, all of which are validated to conform to an expected difficulty level, should yield a test that can then be scored accurately and reported back to test-takers and institutions efficiently. For an example of scoring procedures, let's take a look at the University of Michigan’s MELAB. The standard form of the test is divided into three parts: (1) written composition; (2) listening comprehension; and (3) four different subsections on grammar, cloze, vocabulary, and reading comprehension. Score reports provide separate results for each of the three parts, plus a final score that is essentially a mean of the three section scores.

To calculate these scores Parts 2 and 3 are multiple-choice items and are machine-scored. Part 1 is a composition that involves two (and sometimes three) human scorers who use a rubric to achieve a final result. This rubric, found on the MELAB Web site.

Step 6: Perform Ongoing Construct Validation Studies

From the above discussion , it should be clear that no standardized instruments is expected to be used repeatedly without a rigorous program of ongoing construct validation. Any standardized test, once developed, must be accompanied by systematic periodic corroboration of its effectiveness and by steps toward its improvement from administration to administration. This rigor is especially true of tests that are produced in **equated forms,** that is, forms that are reliable across several administrations (a score on a subsequent form of a test has the same validity and interpretability as a score on the original).

Pearson’s PTE Academic requires test-takers to respond to number of different three sections: (1) speaking and writing; (2) reading; and (3) listening. The tasks range from selected and limited responses (e.g., fill in the blank, multiple choice) to extensive and extended production (e.g. summarize a spoken / written text, give a personal introduction, describe an image , retell a lecture, write an essay), and the test therefore uses both right/wrong and partial-credit scoring. Because all scoring is done by machine , PTE is able to provide efficient turnaround and states that results are typically available in five business days. Experts in the field have expressed concern that relying exclusively on automated scoring as the sole scoring procedure may require more rigorous criteria, however pae’s study validated the stability of PTE academic as a useful measurement, tool for assessing language learners, academic English.

STANDARDIZED LANGUAGE PROFICIENCY TESTING

Tests of language ability presuppose a comprehensive definition of the specific competencies that comprise overall language ability This is not the only way to conceptualize the concept. Swain (1990) offered a multidimensional view of proficiency assessment by referring to three linguistic traits (grammar, discourse, and sociolinguistics), which can be assessed by means of oral, multiple-choice, and written responses. Another definition and conceptualization of ability is suggested by the ACTFL association mentioned earlier. ACTFL takes a holistic and more unitary view of proficiency in describing four levels: superior, advanced, intermediate, and novice.

Superior-level speakers are characterized by the ability to:

- participate fully and effectively in conversations in formal and informal settings on topics related to practical needs and areas of professional and/or scholarly interests
- provide a structured argument to explain and defend opinions and develop effective hypotheses within extended discourse
- discuss topics concretely and abstractly
- deal with a linguistically unfamiliar situation
- maintain a high degree of linguistic accuracy
- satisfy the linguistic demands of professional and/or scholarly life.

The construction of a valid standardized test is no minor accomplishment, whether the instrument is large- or small-scale.

1. First, a standardized test should be founded on soundly constructed standards, free of bias This is a tall order and requires careful gathering and analysis of performance data and institutional goals.

2. Second, the designing of specifications alone

3. Third, the construction of items and scoring/interpretation procedures may require a lengthy period of trial and error with prototypes of the final form of the test