

Standardized Testing

Say the word "test" to nearly anyone - student, teacher, administrator -and their faces cloud over. Everyone realizes the "intimidation potential" that tests have, and perhaps there is a good reason for this fear. In fact, in recent years the entire subject of testing and assessment has come under intense scrutiny at all levels of education.

Nevertheless, testing is not about to disappear, and in the ESL field, there already exist some major tests and exams most learners will face at one point or another in their studies. Some of the most widely-accepted are TOEFL, TOEIC, IELTS and Cambridge.

As teachers, it is vital to become aware of these forms of evaluations and help our students as best as possible to be ready for them. These exams can alter lives; a 10% lower than expected grade could mean a missed opportunity for a new job, an acceptance letter for university and so on.

So, let's see what's out there exactly...

TOEFL

Introduction

The TOEFL test refers to the Test of English as a Foreign Language. The TOEFL program is designed to measure the ability of non-native speakers to read, write and understand English as used at college and university campuses in North America. The TOEFL test is also sometimes used to evaluate the English ability of candidates for admission to academic or professional programs outside of English-speaking countries. Since it was first administered in 1965, the TOEFL test has been taken by about 850,000 candidates annually. Currently over 4,300 colleges, universities, professional programs and sponsoring institutions throughout North America require the TOEFL test. The content of the TOEFL test consists of general use of both written and spoken North American English. Although the TOEFL test places special emphasis on English, in both spoken and academic writing contexts, as used on North American university campuses, specialized knowledge of cultural or academic subject matter will not provide an advantage to TOEFL test examinees. Educational Testing Service conducts extensive research and review of TOEFL test content to ensure that there is no bias against or in favor of particular national, racial or linguistic groups.

Test Format

The TOEFL test is administered in two different formats, the traditional paper-based

format and the Computer-based Test (CBT) first introduced in July 1995. The TOEFL test is currently taken in the computer-based format in most parts of the world, however the paper-based TOEFL test continues to be used in areas where the CBT is less feasible. The computer-based TOEFL test uses both question and items similar to the paper-based format and questions that are computer-specific.

The TOEFL CBT Test

Candidates taking the computer-based TOEFL test will first go through a seven-step tutorial designed to teach computer skills necessary for taking the test. The tutorial is un-timed and un-scored. Skills such as using a mouse, scrolling and icon recognition are covered in the tutorial. There is one tutorial for each of the four sections of the TOEFL test; Listening, Structure, Reading and Writing. The Listening and Structure sections of the TOEFL CBT are a computer-adaptive. This means that candidates will first be given a question of average difficulty and will then be given questions of lower or higher difficulty based on whether the question was answered correctly or incorrectly. In other words, computer-adaptive testing tailors difficulty levels based on candidates' abilities. The format of the computer-based TOEFL test is as follows:

Section	Time Limit	No. of Questions
Tutorials	no time limit	
Listening	40-60 minutes	30-50
Structure	15-20 minutes	20-25
Break	5 minutes	
Reading	70-90 minutes	44-55
Writing	30 minutes	1 topic

Listening

The listening section of the TOEFL test is divided into three parts. In the first part, candidates will hear short conversations consisting of two lines followed by a single question. In the second part, candidates will hear a longer conversation between two speakers consisting of seven to lines followed by four to five questions with four answer choices each. Longer lectures, talks or broadcasts comprise the final part of the listening section. Candidates will hear a one to two minute long talk delivered by a single speaker followed by four to five questions. Although most listening section questions use the traditional multiple-choice format, some listening section questions are of a type that can only be offered over a computer. Examples of computer-specific questions involve visual recognition, selecting two of four possible answer choices and items requiring test-takers to order or match objects, phrases or words.

Structure

The Structure section focuses on recognizing vocabulary, grammar and usage proper to standard North American written English. The subject matter of Structure section test items is academic, focusing on science, the arts, literature, culture and history related to North America. However, specific or detailed knowledge of these topics is not necessary for correctly answering the question items. There are two types of questions in the Structure section of the TOEFL test. One question type presents candidates with a sentence containing a blank line. Test-takers must choose a word or phrase that appropriately fills in the blank. The other Structure section question type consists of complete sentences with four separate underlined words. Candidates must choose which of the four underlined answer choices contains an error in grammar or usage.

Reading Section

The Reading Section comprises short passages similar of the type used as academic texts in the US or Canada. The texts focus on the arts, literature, biography, science and history in a North American context. However, specific knowledge in any of the fields above will not provide an advantage for answering the questions correctly. Success on the reading section is achieved solely through understanding the passages and questions. Main ideas, facts, inferences, and vocabulary are the focus of Reading section questions. The Reading section uses the traditional multiple-choice format as well as computer-specific questions requiring candidates to choose a particular word, phrase or paragraph from the text. Other computer-specific questions require candidates to insert a sentence in the text where it is most appropriate.

Writing Section

The Writing section requires candidates to compose an essay using original thought, analysis, examples, evidence and organization in English. The essays are based on single assigned topics. Candidates have to the choice of composing their essays at a computer or using the traditional hand-written format.

Computer-Based TOEFL Test scoring

The computer-based TOEFL test is scored using a raw score based on the number of questions answered correctly that is converted to a scaled score ranging from 0 to 300. The Writing section is scored on a scale of one to six and comprises about one half of the scaled score on the Structure section. The following chart demonstrates the scoring system for the computer-based TOEFL test.

Listening 0-30
Structure/Writing 0-30
Reading 0-30
Total 0-300
Essay (on a separate scale) 0-6

The TOEFL Paper-based Test

The paper-based TOEFL (PBT) test consists of three sections, similar to the first three sections of the computer-based test. The PBT consists entirely of multiple-choice questions. All answer choices are printed in a test booklet and answers are filled in on a computer-scored answer sheet. The paper-based TOEFL test is now offered only in areas where computer-based testing is not feasible, however the number of candidates taking the paper-based TOEFL test will increase significantly in the near future now that Educational Testing Service has eliminated numerous computer-based testing centers. Since computer-adaptive testing is not possible on the paper-based test, the PBT has three question difficulty levels. Thirty percent of PBT test items are categorized as easy, forty percent are medium and thirty percent are difficult. Easy level questions appear at the beginning of each section, followed by medium-level questions in the middle and difficult questions at the end. There are neither penalties for missing easy questions nor bonuses for answering difficult questions correctly. The format of the PBT is as follows:

Section I: Listening Comprehension

Part A: Dialogues

30 items

In this part, candidates will hear thirty two-line dialogues followed by questions. There are four answer choices printed in the test booklet.

Part B: Extended Conversations

2 conversations

7-8 items

In the extended conversations part, examinees will hear two six to eight line conversations between two speakers followed by three or four spoken questions for each conversation. The four answer choices for each question are printed in the test booklet.

Part C: Mini-talks

3 talks

12-13 items

Examinees will hear three talks, lectures, broadcasts or speeches given by a single speaker followed by three to five questions about each talk. As in the other two parts of the Listening section, the questions for the mini-talks will be heard over audio only.

Time: Approximately 30 minutes

Section II: Structure and Written Expression

Sentence Completion

15 items

The sentence completion part of the TOEFL test is comprised of fifteen sentences with a single blank in place of one or more missing words. Test-takers must choose the word or words that appropriately fill in the blank.

Error Recognition

25 items

Candidates will read twenty-five sentences, each with four answer choices in the form of underlined words or phrases. Candidates must select the answer choice containing an error in grammar or usage.

Time: 25 minutes

Section 3: Reading Comprehension

5-6 passages

50 items

Candidates will read five to six passages followed by seven to twelve answer choices following each passage. Factual information, vocabulary comprehension and inference making are the most important aspects of the Reading Comprehension section.

Time: 55 minutes

Scoring the Paper-Based TOEFL Test

The paper-based TOEFL test is scored based on the number of items answered correctly converted to a scaled score using a weighted system ensuring scores on each TOEFL test are equal regardless of differences of difficulty from test to test. The scaled scores from each of the three sections are added together, multiplied by ten and divided by three leading to score ranges of a minimum of 200 and a maximum of 670. For example if a candidate were to receive scaled scores of a 45 on the Listening section, a 62 on the Structure and Written Expression section and a 55 on the Reading section, his or her final score would be 540. (162×10) divided by 3 = 540.

At Education Canada, we offer a TOEFL preparation course, therefore, feel free to sign out any material that you may like to browse through to help familiarize yourself with this important exam.

TSE

Made by the same company, ETS (Educational Testing Systems), is the recently-added TSE, the Test of Spoken English. It measures the ability of non-native speakers of English to communicate orally in English. The test is approximately 20 minutes long and includes nine questions.

Students' performance on the TSE indicates how their oral language ability might affect their ability to communicate successfully in an academic or professional environment. TSE scores are used by many North American institutions of higher education to select international teaching assistants, sometimes called ITAs. The scores also are used for selecting and certifying health professionals, such as physicians, nurses, pharmacists, physical therapists, and veterinarians.

The test requires demonstration of the ability to communicate in English by responding orally under timed conditions to a variety of printed and recorded information.

On the day of the test, a recorded interviewer asks the student some questions. Some of the questions will be printed in the test book, and the time one has to answer each one is printed in parentheses after the questions.

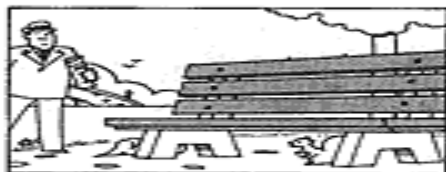
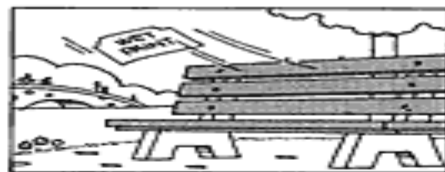
While most of the questions on the test may not appear to be directly related to each individual's academic or professional field, each question is designed to tell the raters how well one can communicate in English.

To help you better understand the nature of this test, here is an example of some questions that the students can expect during their test.

Test of Spoken English Sample Test

Please look at the 6 pictures below. I'd like you to tell me the story that the pictures show, starting with picture number 1 and going through picture number 6. Please take 1 minute to look at the pictures and think about the story. Do not begin the story until I tell you to do so.

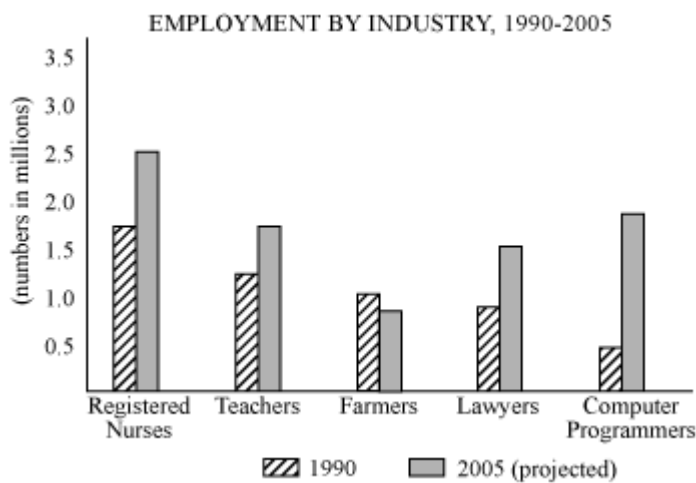
1. Tell me the story that the pictures show. *(60 seconds)*
2. The man in the pictures is reading a newspaper. Both newspapers and television news programs can be good sources of information about current events. What do you think are the advantages and disadvantages of each of these sources? *(60 seconds)*



Now I'd like to hear your ideas about some topics. Be sure to respond to the questions as clearly and completely as you can.

3. Many people enjoy visiting zoos and seeing the animals. Other people believe that animals should not be taken from their natural surroundings and put into zoos. I'd like to know what you think about this issue. *(60 seconds)*
4. If you could visit any place in the world for a month, where would you go and what would you do there? *(60 seconds)*

The graph below shows the number of workers in five different occupations in the United States in 1990 and the projected number for the year 2005. Take 15 seconds to look at the graph.



5. Tell me about the information given in the graph. *(60 seconds)*
6. What do you think might be some of the reasons for the changes represented in the graph above? *(60 seconds)*

In the following questions, you will be asked to imagine yourself in a work situation. The questions are designed to allow you to show how well you can communicate in the workplace. It will be helpful for you to make notes. Remember to make your response appropriate to the situation and to the people you are addressing.

7. Now you will be asked to respond to a co-worker. Imagine that you happen to meet a colleague who has recently received a promotion. Greet your colleague and be sure to
 - mention the recent promotion,
 - express your positive reaction to the promotion, and
 - extend appropriate wishes to the colleague.

You will have 30 seconds to prepare your response. Do not begin speaking until I tell you to do so. *(60 seconds)*

TOEIC

The TOEIC® test provides outstanding benefits to individuals, schools, and companies. It offers an objective assessment of English language proficiency - a quantifiable standard of performance recognized around the world.

The TOEIC test is highly reliable and accurate, available on demand, and features rapid test scoring and reporting.

Individual test-takers find the TOEIC test an excellent means to apply for new positions, to obtain credentials, and to monitor their own improvement in English.

Schools use the TOEIC test to place students into language learning levels, demonstrate the progress of English language students, and evaluate program effectiveness.

Companies, from small businesses to multinationals, rely on the TOEIC test to document progress in English-training programs, to recruit, promote, and hire employees, and to implement a common standard of measurement across multiple corporate sites.

What is the format of the TOEIC® test?

The TOEIC test is a two-hour, paper-and-pencil, multiple-choice test that consists of 200 questions divided into two separately-timed sections:

Section I: Listening: This section consists of 100 questions and is delivered by audiocassette. It is divided into four parts. Examinees listen to a variety of statements, questions, short conversations, and short talks recorded in English, then answer questions based on the listening segments. The Listening section takes approximately 45 minutes.

- Part 1: Photographs 20 items (4-choice)
- Part 2: Question-Response 30 items (3-choice)
- Part 3: Short Conversations 30 items (4-choice)
- Part 4: Short Talks 20 items (4-choice)

Section II: Reading. The Reading section consists of 100 questions presented in written format in the test booklet. Examinees read a variety of materials and respond at their own pace to questions based on the item content. The Reading section lasts approximately 75 minutes.

- Part 5: Incomplete Sentences 40 items (4-choice)
- Part 6: Error Recognition 20 items (4-choice)
- Part 7: Reading Comprehension 40 items (4-choice)

Examinees respond to test questions by marking one of the letters (A), (B), (C), or

(D) with a pencil on a separate answer sheet. Although the actual testing time is approximately two hours, additional time is needed to allow examinees to complete the biographical questions on the answer sheet and to respond to a brief questionnaire about their educational and work history. Therefore, you should allow approximately 2 1/2 hours to take the test.

Clients and examinees require rapid, affordable, and convenient service, as well as high reliability. The decision was therefore made to measure only listening and reading skills directly. These skills can be tested objectively, cost-effectively, and efficiently. Testing speaking and writing directly requires considerable time and expense, both for administering the test and for scoring. Furthermore, direct tests of speaking and writing are generally not only less objective, but are also less reliable.

However, the TOEIC test does provide an indirect measure of speaking and writing. Studies with large samples of non-native speakers of English from around the world have confirmed a strong link between TOEIC results and oral proficiency. Smaller studies have shown a similar link with writing skills. Please see the TOEIC Technical Manual for further details about these studies. [Click here](#) to learn more about the manual and other TOEIC publications. The TOEIC and the TOEFL tests were developed to serve different purposes. Therefore, the design, content, context, and ranges of proficiency that each test measures are also different. The TOEFL test was created by [Educational Testing Service](#) for foreign students seeking admission to colleges and universities in North America. Students planning to pursue undergraduate or graduate degrees in North America will wish to take the TOEFL test. Organizations that document employees' English proficiency and individuals who want to demonstrate their ability to use English in the global workplace will prefer to use the TOEIC test.



[Previous Section](#) | [Chapter 2 Table of Contents](#) | [Next Section](#)

Standardized Tests: Their Use and Misuse

In April 1988 Congress enacted legislation which for the first time calls for using standardized tests to evaluate ABE and ESL programs funded under the Adult Education Act. The Adult Education Amendments of 1988 (Public Law 100-97) and the implementing regulations of the U.S. Department of Education (August 1989) require that the results of standardized tests be used as one indicator of program effectiveness.

For the adult education and literacy community this new mandate brings special urgency to what was already a matter of growing concern: the use and misuse of standardized tests.

From the sheer volume of standardized test-giving, it would appear that we are a nation obsessed. For example, a study by the National Center for Fair and Open Testing estimates that U.S. public schools administered 105 million standardized tests during the 1986-87 school year alone. This included more than 55 million tests of achievement, competency, and basic skills which were administered to fulfill local and state mandates, some 30-40 million tests in compensatory and special education programs, two million tests to screen kindergarten and pre-kindergarten students, and 6-7 million additional tests for the GED program, the National Assessment of Educational Progress, and the admissions requirements of various colleges and secondary schools.

A major reason that standardized tests have come into such pervasive use is that they are relatively easy to administer on a wide scale, no small matter when dealing with a large population. Moreover, they are viewed by their advocates as scientific measuring instruments that yield reliable and objective quantitative data on the achievement, abilities, and skills of students, data that are free from the vagaries of judgment by individual teachers. Because the test and the conditions under which they are administered are (theoretically) constant, except for the skill being tested, they are thought to be useful for comparing a person's ability from one time to another, as in pre- and post-testing. By the same token they are viewed as useful for evaluation program effectiveness and by extension as a tool for improving educational quality.

However, as standardized tests have come into sweeping use throughout education and employment, so have complaints about them and challenges to their validity. They have been the subject of criticism in congressional hearings and state legislatures, and are increasingly the subject of lawsuits in state and federal courts.

Not surprisingly, when the new federal requirements for standardized testing in ABE and ESL were set forth this past August, it was over the objections and protest of many members of the adult basic education community. [Note: See the Federal Register, August 18, 1989.]

The reasons are compelling. Assessment in adult literacy is a central issue with high stakes. The authority vested in these tests can

determine the way programs are developed, what is taught, and the climate of teaching and learning. It shapes legislation and the funding policies of public and private agencies. It is tied to welfare eligibility for young parents. It drives government job training programs. It can deny entry into the military, or crucial access to a diploma or a job.

The growing concern of literacy service practitioners, theorists, and test designers, among others in the field, is sparking much debate and a hard look at just what standardized tests actually test and for what purposes, and whether the results tell us anything of real value, indeed whether they are not harmful. It is also beginning to result in a search for alternative assessment approaches.

The complexities of the testing controversy are vast and beyond the scope of this general article, but opponents of standardized basic skills tests fault them for a host of reasons, some of which are discussed below. Objections tend to fall into two broad categories: their intrinsic defects, and their misuse.

MAKING GRADE LEVEL COMPARISONS

The most commonly used general literacy tests are off-the-shelf commercially-produced tests of reading achievement. Virtually all are "normed" on children. That is, their scores are based on the average performance of children at various grade levels. Because adults bring years of prior knowledge and experience to the acquisition of literacy skills, comparisons with the performance of children are considered by most experts to be inappropriate.

Test scores are usually in the form of grade-level equivalents. A person may score at a 4.2 grade level, say, meaning that he or she reads on the level of a child in the second month of the fourth grade. Not only is this humiliation to people already the victims of past school failure, charge the critics, but it is meaningless to tell adults of any age that they read like a nine-year-old. More importantly, it is not a useful measure of what adults can do in terms that are contextually meaningful and it does not point to an appropriate instructional program.

In fairness, it must be noted that the Test of Adult Basic Education

(TABE) which appears to be the most widely used of all general literacy tests and which has been mandated for use throughout New York State has recently been improved. Analysts indicate that while TABE is still strongly tied to childhood norms, the newer version does make it possible to interpret test scores in relation to other adult in certain ABE programs, rather than to children. It also produces scaled scores rather than grade-level equivalents (though many administrators are apparently falling back on the grade-level scoring system they know because they find the scaling system hard to interpret).

TESTING TRIVIAL SUB-SKILLS

"TABE and other standardized general literacy tests are not a true representation of how people read." says Clifford Hill, Professor of Applied Linguistics at Columbia University Teachers College. "They force the reader to recycle very low level trivial details and don't really represent the reading process with all its complexity." The questions they pose deal with isolated, decontextualized bits and pieces of reading sub-skills such as word recognition spelling, or paragraph comprehension. Questions are framed in a multiple choice format, and they dictate one right answer. There is no applied use of reading or math, no writing component, no higher order thinking or problem solving. "The way the tests are set up, the research shows that even people who read well often don't perform well on most reading tests."

Tom Sticht, one of the nation's pre-eminent test designers, agrees. Sticht notes, for example, that "people who wish to join the armed forces are excluded if they test anywhere from the 10th to 30th percentile in the Armed Forces Qualification Test. But the research shows that eight out of ten people in this category, when they were allowed in, completed their three years with satisfactory performance."

KNOWLEDGE THEORY IGNORED

Recent advances in knowledge theory point to the central role of prior knowledge in understanding or interpreting new information. But most tests exclude prior knowledge; in fact, they assert it as a virtue that they measure comprehension in a manner unaffected by a student's background knowledge. Yet, according to *What The Reading Tests Neglect*, a 1987 study by Anne Bussis and Edward Chittenden of the

Educational Testing Service, "the best a person can do is merely repeat or slightly paraphrase the author's words... The up-shot is that tests...tend to focus attention on the surface structure of text rather than on its underlying meaning..."

LITERACY IN A VACUUM

While it is well established that what constitutes literacy differs from one context to another, the tests treat literacy as a neutral mechanical skill unrelated to different communities and cultural and linguistic traditions. They assume that all individuals perceive information and solve problems the same way. Test results may therefore reflect differing styles, not differing abilities. By the same token, they tend to place superior value on one set of cultural assumptions over another.

Just recently, the National Academy of Sciences conducted a study (*Fairness in Employment Testing: Validity Generalization. Minority Issues, and the General Aptitude Test Battery*) for the U.S. Department of Labor on the use of the General Aptitude Test Battery (GATB). They concluded that the test does not give equally valid responses for blacks, whites, and Hispanics and recommended the use of "within-group" norms. In the end, they declared that no job seeker should be obliged to take the GATB because its negative aspects might outweigh its usefulness.

TESTING OF WHAT, FOR WHAT?

While this central question should guide every test given anywhere, failure to honor it creates special mischief in the workplace. There, lawsuits claiming test misuse have become commonplace; in particular from general basic skills tests given to employees or job applicants that are unrelated to specific job requirements.

According to the experts, there is usually a high correlation between the ability to perform generalized skills and job-related skills, but this correlation is far from perfect and not an adequate basis for predicting a person's performance on a given job. "One of the things you've got to do whenever you're building a test to see if a person can or cannot perform the literacy requirements of a specific job is to design a specific test derived from the analysis of the job of the job field." notes

Tom Sticht. "That way you can show that the test has content validity, or task validity. Only if you test the kinds of tasks that will have to be performed on a job, can you meet the legal requirements of being content of task-related. General literacy tests won't do that."

CONFUSING LEARNER AND PROGRAM EVALUATION

Standardized tests which examine what an adult has learned over a period of time are often used, or misused, as a substitute for program evaluation. When someone wants to know how effective a program is, they look at the test scores.

The trouble is that tests scores alone are not a reliable indicator of what a program has actually accomplished. For one thing, the tests usually are not linked to any particular curriculum; as a consequence there is apt to be a disconnection between what is taught and what is tested. For another, because little is known about the prior knowledge of learners or the learning they may have achieved elsewhere, the test scores may reflect information on skills not in fact taught by the program being evaluated. Furthermore, many elements that are critical to judging program effectiveness internal management, quality of curriculum and teaching, retention rates -- may well be passed over or down graded in favor of the test scores.

In short, program evaluation is more than an aggregation of test results, and multiple interments are needed to measure the effectiveness of discrete program components. "ABE is largely a field devoid of theory," notes Judith Alamprese, Director of Education and Training at the Cosmos Corporation, "so we don't really understand the relation between what we do and what we get. We need research to develop models that can do that."

STANDARDIZATION: WHAT IT MEANS

At best, testing an evaluation is a highly complex enterprise with confusion even among the experts as to the meaning an appropriate use of different testing instruments.

Standardized tests, for example, are often confused with "norm-referenced" and "criterion-referenced" tests. This is a serious matter

because a standardized test by definition is a test designed to be given under specified, standard conditions, whether or not it is norm- or criterion-referenced. (Norm-referenced tests are used to compare the performance of some other group with the "normal" performance of some other group, or for comparing an individual's ability from one time to another, as in pre- and post-tests. Criterion-referenced tests assess a learner's gains according to some criterion or particular learning goal.)

A standardized test may be either norm-referenced or criterion-referenced, but if it is administered under non-standard conditions the results are next to meaningless. For instance, standardized tests are designed to be timed but sometimes are not, or at least not uniformly. An untimed test cannot usefully be compared to one that is timed. Sometimes tests are even taken apart and only certain sections used. Variations in the psychological state of the test-takers can also create non-standard condition. Some people may be under stress because they are unprepared in test-taking strategies while others with more experience are more relaxed., Because of such differences, the point in a program at which a test should be administered is an important matter. (In New York City, where students are required to be tested within the first 12 hours of entering a program, savvy teachers give the tests at the 12th hour.)

Tests and measurements are a complex stew to begin with, but the problem is made worse by the fact that adult literacy programs are staffed in the main by part-time people and volunteers, and by people running programs who are not trained in assessment or have little professional preparation. "When put in the hands of novices, the test can actually amount to malpractice," observes Tom Sticht. "If you went to a physician who tested your blood for cholesterol but didn't use the test instrument the way it was designed to be used, ignored the time required for analysis of the blood or maybe combined the wrong chemicals in the analysis and then gave you a false number, you could sue the physician for malpractice. Because then you might walk out thinking you have no problem and indulge yourself in all kinds of things that wind you up in a heart attack. That may sound like a blatant example, but it's similar in education. When you misuse a test instrument you're representing information falsely to the learner and to

the program sponsor, and eventually you open yourself to lawsuits."

CASAS And NAEP: An Advance

It is the opinion of some that all standardized tests are tarred by the same brush. But there is much agreement that two standardized testing systems represent a very strong forward movement: the Comprehensive Adult Student Assessment System (CASAS) and the National Assessment of Educational Progress (NAEP).

CASAS is keyed to life skills criteria established in the groundbreaking Adult Performance Level Study carried out by the University of Texas in the 1970's. It tests basic skills from a bank of some 4,000 test items, all meaningful in the context of everyday adult life. It serves as a diagnostic tool that places the learner at an appropriate level of instruction and contains pre- and post-test components for a systematic way of monitoring progress and moving the learner on the next level. It tests the student for achievement independently of comparison with others, but has been normed on adult groups and thus can be used for more valid comparison across programs. The competencies tested are on a continuum that range from beginning through advanced levels of ABE and ESL. Teaching materials are used that teach what is going to be tested, coordinating the assessment with instruction. So far the teaching materials are comprised of commercially-available publications identified by CASAS as meeting the curriculum, though CASAS is presently developing some of its own materials.

Developed in 1982 for use by the state of California, the major contribution of CASAS is its focus on adult life skills, accurate placement, ongoing assessment for movement across levels, and linking curriculum to assessment. People who adopt the method are trained in how to administer the tests and how to interpret the scores which are based on a system of scales rather than grade-level equivalents. A number of states including California, Connecticut, Maryland, Washington, and Oregon have adopted CASAS for statewide assessment of their ABE programs, in large part to determine employment eligibility in JTPA and in welfare reform programs.

NAEP, in a new four-year multi-million dollar project funded by the U.S. Department of Education, may bring even further advances to the art of standardized testing. The project came about because Congress decided in the Adult Education Act of 1987 that it wanted a definition of adult literacy and an estimate of its prevalence. Building on its 1985-86 literacy assessment of persons aged 21-25, NAEP is developing a set of survey instruments to measure and estimate the literacy abilities of Americans aged 16-64, according to race, ethnic background, levels of education, gender, and the like. The new information should provide a valid base for making comparisons among regions of the country and also provide policy makers with data they need to make informed decisions.

NAEP's survey instrument will differentiate among three types of literacy: prose (newspapers, magazines, books), documents (charts, graphs, forms), and applied numerical activities (computing the cost of a meal or interest on a loan). Instead of multiple choice questions it will use open-ended exercises that require the test-taker to respond by actually using language and writing out the answers. The test will be designed to cover a gamut of ability from the most basic to the most advanced levels of graduate education. Thus, the data collected will be representative of the entire population (as compared to the CASAS which deals with populations at the adult basic education level). The long range goal of NAEP is to produce tests that program planners, in both general literacy and workplace settings, can use not only to diagnose individual's skills problems but also to design suitable education programs.

ALTERNATIVE ASSESSMENT APPROACHES

While both CASAS and NAEP are hailed as "better" psychometric instruments" than we have had in the past, there are some who remain unimpressed.

"Better tests or not, they represent the psychometric mentality and some of us don't buy that," says Clifford Hill. "Even if you're using better test techniques, what you're measuring is still limited pieces of reading mechanics, and that's misleading. In the real world reading is a complex inter-related holistic process."

A growing number of practitioners around the country agree and have begun to explore alternative approaches to assessment. The perspective that guides these efforts is that the paramount purpose of assessment should be to help the learner achieve his or her goals; that what is assessed must reflect what the learner wishes or needs to accomplish; that the process must build on the learner's experience and strengths rather than deficits; that assessment is not something done to the learner; that it should not be externally imposed nor shrouded in mystery, nor separated from what goes on in the regular course of learning activity. Rather, it is postulated, assessment should be an organic part of the learning experience an ongoing collaboration between the teacher, the learner, and the text, to review and refocus what should take place in the light of progress being made. It should not depend on a single procedure but a variety of procedures. And one of its major functions should be to produce feedback that will make programs more effective. Most of all, testing instruments should convey respect for learners.

The basic point in this line of thinking is that assessment is much more than testing. There are a host of measures that can serve as indicators of achievement e.g. interviews on the use of literacy in contexts other than the program, interactive readings selected by participants for discussion, portfolios of student writing, observation by teachers and peers, simulations of tasks involving life skills, and performance demonstrations. Information derived from an array of indicators, collected over time and assembled into a descriptive package, can provide a rich view of learning and accomplishment.

A YOUNG MOVEMENT

At present these ideas are more a set of principles than a systematic set of applications. In fact, a major task confronting the field is to systematize alternative assessment approaches into strategies that can be used in a wide range of contexts. The challenge is especially difficult because by definition, "learner-centered" assessment is non-standardized. It varies with the context, from learner to learner and from program to program.

It is not known either whether all service providers, regardless of their organizational type and differing clienteles, need to gather the same

kind of information, or whether funding agents can accept diversity in the reporting and be educated to understand and accept different ways of looking at program and student achievement. Relatedly, because evaluation is ordinarily for purposes of accountability or for admittance into jobs or other education, it is not clear how assessment data should be analyzed and reported out to various parties with often-incompatible purposes i.e. the learner, the general or workplace literacy program, funders, and other groups. Two other problems also loom large: Descriptive assessment approaches are very labor intensive and ways need to be found to make the process more time- and cost-efficient. Moreover, the capacity of literacy practitioners to construct their own assessment procedures is presently limited, pointing to a tremendous staff development and teacher training need.

These and other issues are currently being probed in a variety of promising projects around the country. One of these is the Adult Literacy Evaluation Project (ALEP), a venture of the University of Pennsylvania's Literacy Research Center and the Philadelphia Center for Literacy. The ALEP effort, directed by Susan Lytle, is developing and examining evaluation procedures in some 70 adult basic education programs in the Philadelphia area.

Another is the Adult Educators Development Project, a program of the Lehman College Institute for Literacy Studies which is directed by Marcie Wolfe. Under a three-year research grant from the Fund for Improvement of Post secondary Education, the project is bringing together practitioners from a mix of New York City literacy programs to examine alternative approaches across different settings with different populations. Still another initiative is the California Adult Learner Progress Evaluation Process (CALPEP), developed by the Educational Testing Service (ETS) for the California Literacy Campaign. CALPEP is presently operating in more than 80 local libraries up and down the state, where some 15,000 adult students are taught by volunteer tutors. The system was commissioned by the California State Library (CSL) after surveying adult assessment practices nationwide. "We had such grave doubts about the standardized tests available that we felt them to be useless, if not worse," observes Al Bennett of CSL. The state's literacy clientele is comprised heavily of adults with low skills levels, people for whom the tests were felt to be the most threatening and

inappropriate. So an alternate approach was needed.

Basically, CALPEP is a joint perceptual activity involving both students and tutors. Together they judge progress according to the student's personal literacy goals and the uses of literacy in their daily lives. A statewide computerized data base allows local library programs to enter student assessment data which is then stored at a central location. This permits program administrators to monitor and quantify learner progress, to better match tutors with students, and to coordinate reporting formats for funders. With the first year of field testing now complete, plans are in process to develop a system to train volunteer tutors in how to implement the new procedures. Ron Solorzano of ETS, among others, stresses that a most significant aspect of CALPEP is that it was initiated at the state level and launched with a research and development plan for making the process systematic.

Finally, the workplace is another setting where alternative evaluation methods are in use or under study. A prime example is the Massachusetts Work-place Education Initiative, a state-funded program that helps local partnerships of employers, unions, and education providers deliver workplace basic skills programs. The Initiative has recently concluded a pilot study based on open-ended interviews with management, supervisors, and union officials. In essence, the question asked was "What are the changes you have seen on the job (as a result of your literacy program to date) and what are you looking for?" The aim was to identify critical factors in evaluating the outcomes of workplace education. The findings, which include anecdotal information about what really matters to employers, will be used to shape a structured questionnaire for more formal evaluation. The plan is to extend the results of the pilot to all 25 programs of the Initiative in 40 workplace sites across the state.

"We have taken the attitude that employers are looking for hard, bottom-line dollar measurement," says Sondra Stein, Director of the Massachusetts Commonwealth Literacy Campaign, "but employers are smarter than that. What they're seeing are workers with better skills and morale, people who are more self-confident and able to work independently. They're seeing changed behavior on the work floor and they're saying that's what they're looking for, not test results on paper."

What we're learning is that companies are understanding quality of work-life issues."

"Ironically," notes Susan Lytle, "the workplace may well lead the way in the development of alternative assessment procedures. It is there that literacy assessment is most closely tied to the functions and purposes of the setting. Assessment is about the meaningful use of literacy in a context; it's not an abstract matter."

Promising alternative assessment work is being done in other workplace settings too numerous to include here. They range from community colleges in partnership with one or more local businesses (e.g. Gateway Community College working with Honeywell in Phoenix), to such industry-wide efforts as that involving the UAW, Ford Motor, and Eastern Michigan University, to the work of Cox Educational Services with several major corporations and public-sector employer around the country.

PERHAPS A BLESSING IN DISGUISE

While the federal call for standardized assessment in ABE and ESL is objectionable to many, others take it as good news, as a sign that adult basic education may be coming of age. Marginal affairs can get by without much scrutiny, they say, but demands for accountability always go with significant resource investments.

That the field of assessment is in ferment is also good news. A decade ago there was little attention to the subject. Today there is not only interest, but considerable searching, experimentation, and variety in actual practice. The notion has taken root that service providers should be showing evidence of program effectiveness. We are certainly more attuned to the diverse purposes of assessment and the need for different testing instruments for different purposes. We have begun to understand that assessment is more than testing, that what can be learned from giving a standardized test is but part of the story. We have made progress in the development of better standardized tests, but we have also grown more sensitive to their limitations (especially to those normed with children and used on adults). At the same time we have a growing movement toward alternative assessment,

characterized by the fact that it is non-standardized.

So the trend is definitely on a positive track. At the same time, however, it is daunting to consider the formidable challenges that face us. To highlight just three:

- There is clearly a tremendous need for research and demonstration to develop a deeper professional knowledge about assessment and the role and use of standardized tests.
- Alternative assessment is a labor intensive activity requiring sophisticated training not presently available to people in the field. This suggests the need for advocacy, and for the development of training structures and programs that move toward professionalizing the entire adult literacy field.
- In literally hundreds of local general and workplace programs around the country assessment is being carried on quietly and out of the public eye, much of it growing informally out of day-to-day practice. No one knows what the accumulated experience adds up to and how it can be used to guide the field. We need mechanisms for collecting and distilling this information.

[Back to Top](#) | [Next Section](#)

www.literacynet.org/icans/chapter02/tests.html