

STANDARD SETTING WORKSHOP

The Test of English as a Foreign Language (TOEFL)

Overview

The Educational Testing Service (ETS) conducted a standard setting workshop for the healthcare professions May 18–19, 2005 in Chicago, Illinois. The Federation of State Boards of Physical Therapy (FSBPT) sent three participants and one observer to the workshop. The participants include Martha Clendenin from Florida, Karen Gordon from Texas and Diane Funderburk from South Carolina. Sue Lindeblad, Managing Director of the FCCPT attended as an observer. The final report from this standard setting workshop for the healthcare professionals will soon be available. In this document I have tried to summarize the unofficial findings from the workshop and reflect the comments of the participants.

Background

The iBT (Internet based training)/Next Generation TOEFL has been in the development phase for the past ten years. Field testing took place over the past two years. It is ready to roll out beginning in fall of 2005. The test will be offered in the U.S. starting in September 2005, will expand to Canada, Italy, France and Germany in October 2005. It will be introduced in other countries of the world throughout 2006.

There will still be an overlap of the test availability throughout 2006, thus individuals who did not pass a portion of the current TOEFL, TWE or TSE may take the old forms to complete their testing process rather than having to start over with a whole new test, unless they choose to do so. The other change in the process is that the iBT will be available only on certain days and will be Internet delivered to testing centers which have been deemed secure. The test will be distributed at the same real time, thus eliminating the ability of candidates taking the exam in New Zealand to pass the questions on to candidates taking the exam hours later in California. The projected scoring release time is three weeks from the test date.

Concordance Tables

The Concordance Tables show the comparison recommendations for scoring from the old to new scale scores. The tables reflect the standard setting and field testing done for undergraduate admissions programs at Carnegie Mellon University and University of British Columbia, and graduate admissions programs from Duke, Lehigh and the University of Texas. They are published on the ETS website, http://ftp.ets.org/pub/toefl/ngt_percentile_information.pdf.

Workshop

This was ETS's first healthcare professional standard setting workshop. ETS staff indicated that other professions have not asked for this process, but they were very enthusiastic to have the healthcare professions together. The professional groups that participated in the process included: physical therapy—FSBPT; occupational therapy—NBCOT; nursing—National Council of State Boards of Nursing and CGFNS; pharmacy—National Association of Boards of Pharmacy; and veterinary—American Veterinary Medical Association's Educational Commission for Foreign Veterinary Graduates.

Standard Setting Process

- Reviewed the TOEFL test.
- Reviewed uses for the TOEFL test and interpretation of the scores.
- Introduced new test processes and content areas.
- Reviewed field testing processes and previous standard setting workshops.
- Trained participants in the standard setting process that would be used in the workshop. The Writing and Speaking scores were assessed by a consensus model. The Reading Comprehension and Listening Comprehension Scores were arrived at using a modified Angoff procedure.



Susan K. Lindeblad, PT, PhD, is Managing Director of the Foreign Credentialing Commission on Physical Therapy since 2001. Prior to assuming that role, she was assistant professor at the University of Miami. Sue served on the Florida licensing board for 14 years, chairing it twice. She has served as a member of the Exam Program Oversight Committee; an item writer for the exam; a member of the Analysis of Practice Oversight Committee; delegate to the Federation's Delegate Assembly; chair of the PT Exam Construction and Review Committee and member of the Job Analysis Oversight Committee. Sue obtained her master's degree in education and Ph.D. in Higher Education at the University of Miami. She was an associate professor in the physical therapy program at the University of Miami for the 12 years prior to joining the staff of the FCCPT.

Results of Various Studies

TOEFL iBT

Panel Recommendations for Scaled Cut Scores

Panel:	Duke	Lehigh	UT	CMU	UBC	Healthcare
Level:	Graduate	Graduate	Graduate	Under-Graduate	Under-Graduate	Professional
Writing	20	25	22	25	25	24
Speaking	22	24	23	23	20	26
Reading	21	20	21	21	19	18
Listening	14	17	17	17	14	21
Total	77	86	83	86	78	89

Recommendations from FSBPT Members

The FSBPT members felt that the group consensus from the workshop clearly reflected their own feelings on the recommendations for scoring. They were in agreement within one point on all of the individual areas, and were able to resolve differences to accept the group recommendations. They noted that individual jurisdictions may wish to conduct their own studies and determine cut scores as a states rights issue. The following recommendations were made to the FSBPT Board of Directors for their consideration.

- ✓ Do not redo the study; accept the recommended cut scores for healthcare professionals resulting from this study.
- ✓ Adopt the minimum recommended scores for healthcare professionals in each of the four scoring domains (Listening, Reading, Writing and Speaking), not just the total score, as each area relates to some aspect of professional physical therapy practice, patient care and safety.
- ✓ During the period when there is an overlap of old and new tests, allow applicants to take either test to meet their English language proficiency requirements for immigration or licensure. If they take the old TOEFL—CBT or TOEFL—paper/pencil and the TSE, use current score recommendations. If candidates take the new iBT TOEFL, use the scores set by the healthcare panel.

- ✓ Monitor the rollout process of the iBT TOEFL. If there are no slowdowns or delays, by January 1, 2007 require the use of the iBT TOEFL to test for English language proficiency.
- ✓ Recommend to jurisdictions that they adopt the minimum passing scores in Listening/Reading/Speaking/Writing and overall total cut score for healthcare professionals based upon the outcomes of this standard setting workshop.
- ✓ Along with NBCOT and National Council of State Boards of Nursing, recommend to USCIS that it adopt the minimum passing scores in Listening/Reading/Speaking/Writing and overall total cut score for healthcare immigrants.
- ✓ Offer a presentation by ETS staff at the 2005 annual meeting regarding the new TOEFL exam, exam structure, scoring and the standard setting workshop itself.

Recommendations from ETS Staff

ETS staff recommended that participating groups watch performance on the exam and revisit and review the cut scores for each profession in the next two years. After that, periodically review the cut scores at regular intervals to maintain validity. ETS has developed CDs to conduct standard score setting workshops, which are available for use by any agency/jurisdiction upon request. ■

with a stack of reference books and her computer, she writes from 6 to 20 questions in a morning, she said.

Compensation varies. Jerry Shannon, 27, a community college professor who lives in Minneapolis with his wife and newborn baby, said he was paid \$25 an hour as a freelancer drafting reading passages for language arts tests.

Although the creation of tests might seem to be a statistics-based science, the drafting of questions is much like other writing. Not everyone has talent, for example, and inspiration can strike anywhere.

Mr. Shannon said he dreamed up one passage as he watched some dachshunds roughhousing in a neighbor's yard. He turned it into a humorous sketch of a frustrated man chasing his mischievous dogs.

"Test questions are a little like poems," said Stephanie Alnot, who works with Ms. Oberley as a test developer for the University of Iowa. "Every word matters. Some questions come easy, others come hard."

Those who write the questions turn them over to test assemblers, whose work reminds Tom Ewing, a spokesman for the Educational Testing Service, of gourmet food.

"Item writers just produce parts of the recipe," Mr. Ewing said. "The test assemblers are the chefs who make the magic happen. They put the tests together, making sure that state standards are met, by taking a pinch of this and a sprinkle of that. That's the real wizardry."

But no amount of wizardry can create a good test out of poorly written items, just as no chef can create a tasty meal from rotten food. And quality has emerged as a problem as the country's testing appetite has grown ravenous.

In May, the National Board on Educational Testing and Public Policy, a group affiliated with Boston College, issued a report documenting 50 high-profile testing mistakes that had occurred in 20 states from 1999 through 2002.

In Oregon, there were mislabeled diagrams and other errors on the Oregon Statewide Assessment Tests, and in Virginia, Harcourt acknowledged that it had set its cut score on a writing test too high, causing 5,600 students to fail, the report said.

In a more recent problem, New York Education Commissioner Richard P. Mills last month set aside the results of the Math A Regents exam for juniors and seniors, acknowledging that it was flawed.

Teachers assembled in workshops write questions which Department of Education officials and consultants assemble into the Regents exams, said James A. Kadamus, a deputy commissioner. Other states, like California, hire commercial publishers to write custom tests tailored to local curriculum



standards. Still others, like Georgia, buy off-the-shelf tests from the commercial publishers.

One of the most widely used is the Iowa Test of Basic Skills, written by

Ms. Oberley and her University of Iowa colleagues and offered at cost to all Iowa school districts. Riverside Publishing of Itasca, Ill., markets it to districts in the other 49 states.

Ms. Oberley, who earns \$58,000 a year, is a former teacher who switched to hand-scoring exams and moved up to item writing. She can crank out as many as 30 items a day, colleagues said, but has been known to ponder a question for months, seeking to phrase it perfectly. Later her questions pass through a gantlet of reviews and experimental trials, which eliminate about half of them.

One hurdle is the bias and sensitivity review, in which representatives of various groups—women, blacks, Muslims, people with disabilities, others—critique the questions.

In "The Language Police," a book published this spring, the author Diane Ravitch described how reviewers at Riverside Publishing deleted from a national assessment test a question that mentioned Mount Rushmore because they considered the monument upsetting to Indians, and rejected an essay on peanuts because some students might be allergic to them. Dr. Ravitch said the bias reviewers exercise a "regime of censorship."

But others defend the system. Ms. Oberley said the reviewers did point out legitimate problems. An example, she said, was a question she wrote to measure kindergarten students' comprehension of the word "driveway." It included sketches of a driveway leading to a suburban garage, of cars on an urban boulevard, and of others on a freeway.

"We have many gravel roads and few paved highways," an American Indian reviewer wrote. "Our children may think these are all driveways."

Yet most questions founder for reasons that have nothing to do with bias, Ms. Oberley said.

The question about the American colonies' boycott of British goods got all the way to Phase 9 in a 14-phase process, passing under the eyes of several editors down the hall from her office, under the pencils of 450 children in middle schools across Iowa in a field test, and onto the desks of Riverside editors in Itasca. Then statistics from a national field trial showed that two-thirds of high-performing students had chosen the absurd tea bag answer, sealing its doom.

"Believe me, writing tests is not just pulling a few facts out of the air," Ms. Oberley said. ■