

Developments in English Language Assessment APEC Strategic Plan for English and Other Languages

I. Introduction

English has become a global language (Crystal, 1997). As a consequence, in APEC economies where English is not the native, majority or official language, it has become a priority foreign language. In a background paper presented at the APEC EDNET symposium convened in Xi'an, China, Chen and his colleagues reported the results of the APEC-EDNET survey they conducted on the status of foreign language standards and assessment among APEC member economies. The authors noted that English was the primary foreign language for 80% of the APEC members (Chen, Sinclair, Huang & Eyerman, 2008). Given the significance placed on the English language in many APEC economies, it is important to monitor global trends and important developments in the assessment of English language ability and to consider their implications for APEC members.

This paper supports Activity 6 in the Strategic Plan for English and Other Languages and complements research conducted by Chen et al. (2008) on language standards and assessment. In this paper, I (1) review some notable developments related to high-stakes assessments of English language ability used in selected APEC economies, (2) highlight key issues in the assessment of English language ability and discuss their implications for developers of high-stakes second language (L2) tests, (3) note current global standards for the assessment of English and other second language abilities, and (4) identify several exemplary frameworks for guiding the development of large-scale, high-stakes assessments of English as a foreign language (EFL) ability.

II. Recent Developments in Major High-stakes Tests of EFL ability

In APEC economies where English is the priority foreign language, English tests frequently perform a gate-keeping function that significantly affects test-takers' educational, employment, and career advancement opportunities (Ross, 2008). When the scores on tests are used to make decisions that have serious consequences, they are considered *high-stakes* tests (Kane, 2002). The principal high-stakes, international tests of EFL ability used in APEC member economies include the International English Language Testing System (IELTS), Test of English as a Foreign Language™ (TOEFL®), and Test of English for International Communication (TOEIC) and there have been some important recent developments related to them.

1. IELTS

The current version of the IELTS was launched in 1995. Enhanced rating procedures, assessment criteria, and scale descriptions were introduced in the speaking component in 2001 and the writing component in 2005. A computer-delivered version of the test (CB-IELTS) was introduced at selected test centers in 2005, and test takers who elect to take the CB-IELTS have the option of handwriting their responses to Writing section tasks or composing them on the computer. The Speaking section for the paper-based and CB-IELTS is delivered in the same manner, using an interviewer and a face-to-face format. The IELTS is designed to assess test

takers' ability to use English for academic or employment purposes in contexts where English is the language of communication. There are two forms of the test (academic and general training). All test takers take the same listening and speaking components and complete the reading and writing components for either the academic or general training form. Cambridge ESOL (C-ESOL) maintains an active research and development program that supports the interpretations and use of IELTS scores for the test's intended purpose. Reports on IELTS validation activities are available on the publisher's Web sites (http://www.cambridgeesol.org/rs_notes or <http://www.ielts.org>), and they provide valuable descriptions of current trends in the design of large-scale language proficiency measures.

The IELTS uses a variety of selected response tasks (multiple choice, fill-in-the gap, true/false, and matching) in the Listening and Reading sections, and it uses performance-based tasks in the Speaking and Writing sections that require test takers to construct oral and written responses to spoken, written, and/or visual prompts. The speaking component employs a particularly noteworthy test method in that an interlocutor engages the test taker in a three-part oral interview. Prapphal (2008) reports that the rapid expansion of English medium programs at the undergraduate and graduate level in Thailand has led to increased use of standardized English assessments and the IELTS has become a popular alternative to the TOEFL in recent years. Since 2002, Hong Kong, China has used the IELTS (academic form) to assess the English proficiency of all graduating university students (Qian, 2008).

2. *TOEFL iBT*

The iBT (Internet-based test) TOEFL was launched in 2005 following a decade of research and development activities that support the design and proposed interpretations and uses of test scores. As the iBT becomes available in the various regions of the world, it will replace the paper- (PBT) and computer-based (CBT) versions of the test. The iBT is designed to assess test takers' English language proficiency and ability to use English in an academic context. Much of the evidence available to support the use of the iBT for its intended purpose is contained in research reports that are available on the ETS TOEFL Web site (<http://www.ets.org>). Those engaged in the development of local high-stakes tests of L2 ability will find ETS research reports and monographs to be an excellent source of information on current trends in the design of large-scale language proficiency measures.

There are some notable developments in the iBT that distinguish it from previous versions of the TOEFL. First, the grammar component has been eliminated and grammar is now assessed in the context of test takers' performance on speaking and writing tasks. Second, a speaking component was added and test takers respond to multiple speaking tasks. Third, the Speaking and Writing sections contain tasks that require test takers to engage more than one language skill and use language in ways that approximate real-world situations. For example, in the integrated speaking tasks, test takers read a short passage, listen to discourse on the topic, and respond orally to questions related to the topic or situation. In integrated writing tasks, test takers read a passage, listen to a short lecture, and compose a summary. These modifications to the test method reflect current trends in how L2 ability is conceptualized and assessed in large-scale, high-stakes tests. Additionally, performance descriptions were developed for the iBT (Educational Testing Service, 2004), and iBT scores were mapped to the Common European Framework of Reference

(CEFR)(Tannenbaum & Wylie, 2005). Both of these developments will make it easier for test users to interpret and use iBT scores.

3. *TOEIC*

Since the introduction of the TOEIC in 1979, the number of test takers has steadily expanded. In 2007, over 5 million registrants in 92 economies took the test (Educational Testing Service, *TOEIC speaking and writing sample tests*, 2007, p. 2). Numerous APEC economies in Asia use TOEIC scores to make decisions related to test takers' education, employment, and career advancement (Choi, 2008; Gottlieb 2008; Kaplan & Baldauf, 2005; Prapphal, 2008). A revised version of the TOEIC Listening and Reading (L&R) test was launched in 2005, and it continues to be a paper-based assessment. Additionally, TOEIC Speaking and Writing tests were introduced in late 2006, and these are computer-delivered assessments that are administered separately and at different times than the TOEIC L&R test. The TOEIC is designed to assess the everyday English ability needed to communicate with others in international business contexts (Educational Testing Service, *TOEIC examinee handbook: Listening and reading*, 2007).

A number of important changes have been made to the TOEIC recently. Pictures in the Listening section have been updated and three spoken varieties of English (Australian, British, and North American) are now used in the listening input. Texts in the Reading section have also been updated and include email messages and a business letter. However, the most significant change has been the addition of Speaking and Writing tests to the TOEIC battery. Whereas the TOEIC L&R relies on traditional multiple-choice test tasks, the Speaking and Writing tests employ performance-based tasks that require test takers to construct oral and written responses to written, spoken, or visual prompts (Educational Testing Service, *TOEIC speaking and writing tests*, 2007). TOEIC test takers and score users will benefit from the availability of TOEIC tests that can cover a broader range of skills and that provide a more comprehensive assessment of communicative language ability. Moreover, recent work has related the TOEIC L&R to the language ability levels of the CEFR and this development will assist test users in interpreting and using TOEIC L&R scores (Tannenbaum & Wylie, 2005).

4. *ACTFL tests of spoken English*

The American Council on the Teaching of Foreign Languages (ACTFL) developed the ACTFL Oral Proficiency Interview (OPI), a standardized procedure designed to assess the functional language ability of test takers. It is offered in English and more than 60 other world languages. The OPI assesses how well the test taker functions in a language by comparing the individual's performance of various communicative tasks with the criteria listed for each of ten levels in the *ACTFL Proficiency Guidelines--Speaking (Revised ACTFL Proficiency Guidelines--Speaking*, 1999).

The OPI test method utilizes a 20-30 minute one-on-one interview that is conducted in person or by telephone with an examiner. Test takers respond to a variety of questions related to their personal experiences and interests. Test tasks are designed to elicit a range of communicative performance that is rated by two certified ACTFL examiners and interpreted as one of ten possible levels in the ACTFL Proficiency Guidelines. Refer to Chen et al. (2008) for a fuller

account of the ACTFL Guidelines and the history of the OPI.

In early 2006, ACTFL launched the ACTFL OPIc, a computer-delivered version of the ACTFL OPI accessed via the Internet. The OPIc is a semi-direct test of spoken language that elicits a 20- to 30-minute sample of ratable speech. It consists of four parts: volume check, self-assessment, background survey, and test tasks. Each OPIc is individualized on the basis of the test taker's responses to the self-assessment and background survey questions. Test takers hear each prompt twice and view images that provide a context for the communication. Responses are recorded digitally and an ACTFL rater compares responses to the criteria in the *ACTFL Proficiency Guidelines—Speaking (Revised, 1999)* and assigns a rating between Novice Low to Advanced. ACTFL reports that validation activities have established a high degree of consistency between scores on the OPIc and scores on the OPI. In 2009, an expanded version of the OPIc will be launched, and it will assess the full range of ACTFL proficiency levels from Novice through Superior.

5. *Two major EFL tests used in China*

In addition to these international tests of EFL ability, many APEC economies use locally developed, large-scale tests. The College English Test (CET) and the National Matriculation English Test (NMET) are two significant EFL tests developed and used in The People's Republic of China. Both the CET and NMET are aligned with China's national English curriculum, and they illustrate some of the challenges and practical constraints test developers confront.

The CET battery is designed to assess undergraduates' achievement of the requirements specified in the national English syllabus for non-English majors (Zheng & Cheng, 2008), and it includes the CET Band 4 (CET-4), CET Band 6 (CET-6), and the CET Spoken English Test (CET-SET). Zheng and Cheng (2008) reported that 13 million students took the CET in 2006, making it the most widely used high-stakes test of English language ability in the world. In a review of the English language testing research conducted by Chinese scholars in the past decade, Cheng (2008) summarizes the results of a number of studies that explored the CET. This empirical research represents some of the evidence available to support the interpretations and use of CET test scores. Since its introduction in 1987, the test content, format, and scoring have evolved in response to insights gained from test use and general developments in the field of language assessment. The 2006 version of the CET-4 and CET-6 contains Listening, Reading, Cloze, and Writing and Translation sections. Listening tasks entail listening to several brief conversations and selecting the correct response to questions and completing a dictation. Reading tasks entail reading passages of varying lengths, applying a variety of reading strategies, and selecting the correct responses to questions or filling in spaces with missing information. The cloze task requires test takers to identify the missing word in a passage from the set of choices. Writing tasks entail developing a short essay in response to a prompt and translating five sentences from Chinese to English. Speaking ability is assessed with the CET-SET, but this is an optional component of the test battery available to test takers who demonstrate an adequate level of language proficiency on the CET. The CET-SET uses a structured oral interview format similar to that employed in the IELTS; however, three to four test takers participate in the

interview and two trained raters (an interlocutor and an observer) assess the test takers' performances.

The NMET is a large-scale, high-stakes English language assessment taken by over 9 million Chinese students each year (Cheng, 2008). It was introduced in 1985 and is designed to assess test takers' English language ability and scores are used to make university admission decisions. Test content focuses primarily on test takers' linguistic knowledge of English, and the Listening, Grammar/Vocabulary, and Reading sections utilize a traditional multiple-choice test method (Cheng & Qi, 2006; Qi, 2005). The practical challenge of administering a listening component in less developed areas of the economy led test developers to postpone the inclusion of this section of the NMET until 1999. Practical constraints have also precluded the inclusion of a speaking component because of the resources required to administer it to such a large number of test takers. The Writing section requires test takers to perform an editing task in which they identify and correct the errors in a text and compose a written response to a prompt. Cheng and Qi (2006) report on some of the evidence available to support the use of the test for its designated purpose.

When the NMET was introduced, it was hoped that it would promote more communicative English language teaching and learning in Chinese secondary schools. Yet, as is true in the case of many high-stakes tests, the impact of the NMET on language teaching and learning has been complex and affected by the expectations of stakeholders. NMET scores are used for university admission decisions and this led teachers, students, and parents to focus more on how to attain the highest possible test scores than on the broader aims of the curriculum (Cheng & Qi, 2006).

6. *Challenges in large-scale assessment*

In a paper presented at the 2008 APEC EDNET symposium, Duff (2008) noted the need for better alignment between high-stakes assessment practices and curriculum standards. Qian (2008) describes how Hong Kong, China, is responding to this challenge. He reports that the principal local high-stakes English test administered to secondary students (the Hong Kong Certificate of Education Examination) has included a School-based Assessment (SBA) component since 2007. The SBA is a criterion-referenced assessment conducted by the student's classroom teacher, and it is aligned with the standards-based English curriculum. Presently, the SBA component contributes 15% to the total test score. Current plans call for implementing a new English language test for secondary students in 2012 and the SBA component of the new test will contribute 20% to the total score. It is clear that some progress in aligning high-stakes assessments with curriculum standards is being made but more progress is needed.

As evidenced in the case of the NMET, practical constraints often limit the test developer's ability to create large-scale tests that are optimally aligned with curriculum standards. These constraints include the availability of the expertise, technology, and money required to develop and administer a test; the time required to take it and process the results; and the expectations of stakeholders. Additionally, in the case of high-stakes measures, test developers must balance concerns for crucial test qualities such as validity, reliability, authenticity, and impact. With the proliferation of computers and rapid advances in technology, it is likely that some of these constraints will be mitigated in the near future. In fact, recent applications of computer technology to large-scale assessment of L2 ability, as demonstrated in the iBT and IELTS, now

make it possible to assess more language skills, abilities, and processes than before and to develop and score test tasks more efficiently (Douglas & Hegelheimer, 2007; Zenisky & Sireci, 2002).

Kunnan (2008) emphasizes that the most important challenge in large-scale assessment is the issue of *fairness*. He defines fairness in terms of the use of fair content and test methods in assessing language ability and the fair use of the scores obtained from the test. Whether test users rely on international or locally developed tests, they have a responsibility to ensure adequate evidence exists to support the interpretations and use of the scores from the test. In cases where there is a lack of evidence available in the public domain for a high-stakes EFL measure, test score users should be cautious about the inferences they make on the basis of the scores.

III. Current Issues in English Language Assessment and APEC Economies

Among current trends in assessing English language ability, four issues have implications for APEC economies: (1) adoption of professional standards to the design and use of high-stakes assessments, (2) determination of the standard (norms) of English to be applied to assessment of EFL ability, (3) representation of L2 ability, and (4) inclusion of performance-based tasks of speaking and writing ability in high-stakes tests.

1. Application of professional standards to the design and use of high-stakes tests

There is general consensus in the educational measurement community that prevailing professional standards and practices ought to be applied to the design and use of high-stakes tests. Several major professional organizations with the expertise to establish standards for educational assessments have codified and disseminated the standards and practices they advocate in publications such as the *Code of fair testing practices* (Joint Committee on Testing Practices/JCTP, 2004), *Code of practice* (Association of Language Testers in Europe/ALTE, 2001), and *Standards for educational and psychological testing* (American Educational Research Association/AERA, American Psychological Association/APA, & National Council on Measurement in Education/NCME, 1999). At the very least, test developers are expected to specify the purpose of the test and present persuasive evidence obtained from multiple sources that the test fulfills its intended purpose. For test developers that embrace the standards advocated by JCTP, this means conducting a variety of validation activities that yield evidence to support the interpretations and use of test scores and integrating the evidence (both theoretical and empirical) into a compelling argument that justifies use of the test for its intended purpose. The *Standards for educational and psychological testing* advocates collecting and reporting evidence related to the test content, response processes, internal structure, relations of other variables, and consequences of testing (AERA/APA/NCME, 1999).

2. Determination of the standard of English to be applied to assessment of EFL ability

Several of the most widely used international EFL tests utilized in APEC economies have been designed to assess the English proficiency of students seeking to study in English-medium colleges and universities in North America, the United Kingdom, or Australia. The tests were not designed to assess secondary students' achievement of the local English curriculum. Hence, it