Asia-Pacific
Economic Cooperation

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

APEC Human Resources Development Working Group

September 2008

HRD 08/2008A

Printed in February 2009

Produced by
Dr. Stephen J. Stoynoff
Minnesota State University
Mankato, MN
USA
stephen.stoynoff@mnsu.edu

For
APEC Secretariat
35 Heng Mui Keng Terrace Singapore 119616
Telephone: (65) 6891-9600
Fax: (65) 6891-9690
Email: info@apec.org
Web site: www.apec.org

APEC#209-HR-01.1

Table of Contents

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 ACTLF 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

may be more appropriate to design local tests of EFL ability that are more closely aligned with the content and aims of the local English curriculum than to use a highly recognized international proficiency test.

In cases where the purpose of an international English test is consistent with the inferences and uses of test scores in local contexts, it is important to recognize that most major international assessments of English ability privilege a variety of Standard English (SE) that may not be spoken in all APEC economies. This raises a fairness concern and the question of whether some of these widely used international tests may be biased against test takers who have not been exposed to SE. Currently, there is considerable debate among applied linguists over both what norms to apply to the use of English and whether some international EFL tests are biased against test takers from particular backgrounds (Elder & Davies, 2006; Jenkins, 2006a; Taylor, 2006). A preliminary investigation conducted by Davies, Hamp-Lyons, and Kemp (2003) did not find any empirical evidence to support claims of test bias in the IELTS, TOEFL, or TOEIC, but other scholars contend these tests do not accept certain communicative language forms that are deemed acceptable in some parts of the world (Jenkins, 2006b). The question of what standards to apply to the assessment of English ability has implications for language education policymakers and test developers. Hamp-Lyons and Davies (2008) submit that there are two key questions to be answered:

(1) Whose norms are to be imposed in the test materials?
(2) What are the consequences for test-takers if the norm imposed by the test is not the "normal" variety accepted in their own society? (p. 27)

3. Conceptualizations of L2 ability

Chen et al. (2008) and Duff (2008) reported on some of the standards-based approaches (ACTFL standards, CEFR, ISLPR, Canadian Benchmarks, and TESOL Standards) to conceptualizing L2 ability. These language standards have been very useful in clarifying for language education planners and teachers what language users' can do at different proficiency levels, but some language testers have noted there is a lack of theory or empirical research to support them (Bachman, 1988; Chalhoub-Deville, 1997; Fulcher, 2004; Weir, 2005a).

Communicative second language ability is a complex, multi-faceted construct, and theoretical models can be quite useful in explicating the various factors that comprise it. For the past 25 years, L2 ability has been conceptualized as consisting of multiple subcompetencies that interact in a particular language use situation. In the decades since Canale and Swain (1980) and Canale (1883) proposed a model of communicative language ability (CLA) comprised of multiple competences, many scholars have elaborated and extended the model (e.g., Bachman, 1990; Bachman & Palmer, 1996; Chapelle, Grabe, & Berns, 1997). Although there is a lack of consensus on exactly how many factors are involved and how they are related to each other, the CLA model remains the dominant theoretical perspective used to represent the nature of L2 ability (Chalhoub-Deville, 2003; Purpura, 2008). Current approaches to language testing use theoretical rationales as well as empirical research to inform the design of high-stakes tests and to justify the interpretations and uses of the test scores. Both ETS and C-ESOL have used the CLA model to inform the design of their international EFL tests. When APEC economy

members decide that a locally developed EFL test is preferable to an international test, a theoretical conceptualization of L2 ability can assist test designers in their work. For a fuller account of how theory-based frameworks can be applied to the development of local tests of L2 ability, see Stoynoff (2007).

4. Inclusion of performance-based tasks of speaking and writing ability in high-stakes tests

One of the most significant changes to the iBT TOEFL was the inclusion of performance-based tasks in the speaking and writing components. Performance-based tasks can contribute to the authenticity of high-stakes L2 assessments and increase the kinds of language knowledge, skills, and strategies test takers engage during the test. The inclusion of performance-based tasks in high-stakes tests also increases the congruency between what students experience in language learning classrooms and what they encounter on large-scale, high-stakes tests. This in turn enhances the positive consequences of using the test. However, the factors that affect performance on these types of tasks are complex and interact in different ways and they are not fully understood. Some of the challenges of using performance-based tasks in large-scale, high-stakes tests are related to task difficulty, the adequacy of construct representation, and the ability to generalize from task performances (Bachman, 2002; Norris, 2002; Norris, Brown, Hudson, & Bonk, 2002; Wigglesworth, 2008). The application of computer technology to task development and the scoring of performance may be helpful in responding to some of the challenges, but it may also affect the validity of score inferences. Therefore, developers of high-stakes tests must present sufficient evidence that the use of performance-based tasks and any applications of technology to them do not negatively affect test takers' performance on the test.

IV. Global Standards for Assessing L2 Ability

In the past decade, a consensus has emerged among measurement specialists and applied linguists on what contributes to the construction of high-quality tests and the promotion of fair testing practices. Yet the actual standards and procedures applied to the design and use of large-scale, high-stakes EFL assessments vary greatly (Eckes, Ellis, Kalnberzina, Pižorn, Springer, Szollás, & Tsagari, 2005). Government entities can play an important role in improving the quality of locally developed high-stakes EFL tests by encouraging test developers to adopt global standards and practices and by identifying useful exemplars that can assist test developers in designing and using high-stakes tests. ETS and C-ESOL are leading centers for research on and development of international tests of English language ability and several of their tests are among the most widely used EFL assessments in the APEC economies. ETS has aligned its test development practices with those advocated in the *Code of fair testing practices* (JCTP, 2004) and the *Standards for educational and psychological testing* (1999). Moreover, ETS has detailed protocols in place to monitor the quality and fairness of its tests (Educational Testing Service, 2002). C-ESOL also aligns its test development practices with global standards, and their practices conform to the standards for test quality and fairness advocated in the ALTE *Code of practice* (2001). As a result, the English proficiency tests and supporting documentation produced by these leading test development centers not only meet current international standards, but they also represent exemplars for the global language testing community.

Based on a review of trends in high-stakes tests of EFL ability, Stoynoff (in press) avers the following generalizations can be made about current approaches to test development.

1. Test developers specify the purpose of the test. This entails specifying the kinds of inferences to be made based upon test takers' performance on the test.
2. Test developers collect evidence from multiple sources and use it to justify the interpretations and use of test scores for the test's intended purpose. The most compelling arguments for a test include both empirical evidence and a theoretical rationale for the proposed uses of the test in a particular context.
3. Test developers monitor the impact of the test (on test takers, score users, educational systems, society). This includes collecting evidence of the impact of using the test and striving to minimize the negative consequences and seeking to maximize the positive consequences of test use.
4. The process of collecting evidence is systematic, comprehensive, and ongoing.
5. Because the process is ongoing and the justification for the interpretation and use of test scores is based on the available evidence, the case for score interpretations and use will be revised as additional information is obtained and developments in language testing occur.

Government entities can advance global standards for development of high-stakes tests by encouraging test developers to comply with professional codes of practice, conduct validation activities that support use of the test for its intended purpose, and adopt exemplary processes for test development and validation activities.

V. Frameworks for Developing High-stakes EFL Tests

The professional literature contains numerous examples of test development frameworks. Most descriptions divide test development and validation activities into stages and specify the kinds of evidence that can be used to support a validity argument for the test. Bachman and Palmer (1996) offer one of the most influential approaches to developing tests of English language ability and their framework can be applied to constructing tests for different purposes and contexts. There are three general stages: "design, operationalization, and administration" (p. 86). Activities in each stage yield certain products. For instance, at the end of the first stage, a comprehensive document is produced that describes the purpose of the test, the target language users and context of language use, the construct of interest, the usefulness analysis, and the necessary resources. The second stage produces test specifications, including the test tasks, instructions, and scoring procedures for the test. In the final stage, the test is piloted and the results from the administration of it and information collected from other stages of the process become part of the evidence available to support use of the test.

Chapelle, Jamieson, and Hegelheimer (2003) formulated a practical framework based on initial work by Read and Chapelle (2001). It divides test development into a process that begins by determining the *test purpose* (including the inferences to be made based on test performance, the use of test scores, and the intended impact of the test) and *validity considerations*. *Test purpose* and *validity considerations* in turn affect subsequent *test design* and *validation* decisions. The process culminates in the development of a validity argument for the test.

C-ESOL organizes the test development and validation process into five stages: initial planning and consultation, development, validation, implementation, and operation (Falvey & Shaw, 2006). Weir (2005b) has created a socio-cognitive framework for prioritizing and conducting crucial validation activities that enable test developers to build compelling validity arguments for tests. His framework contains five elements (context validity, theory-based validity, scoring validity, consequential validity, and criterion-related validity) and considers three dimensions (test taker characteristics, task response, and score). Weir's framework reflects current trends in the design and validation activities associated with large-scale, high-stakes EFL tests and it has informed the activities of C-ESOL test developers.

ETS operates an active program of research and development that supports its EFL tests and the results are published in a series of monographs and technical papers that are available on the publisher's Web site. The results of many of these papers were integrated into a recently published case study of the development of the iBT (Chapelle, Enright, & Jamieson (2008). The volume presents one of the most comprehensive descriptions of the evidence and validity argument for a high-stakes EFL test currently available. In the book, project participants articulate a framework for the project and summarize the validation activities that informed the design of the test and support the interpretations and use of iBT scores. One key aspect of the project was the construction of an interpretive argument for the new TOEFL and it was based on recent developments in validation theory and current standards of educational measurement.

VI. Conclusion

Language testing is increasingly acknowledged to be not only a form of educational practice but a form of social and political practice as well (McNamara, 2008; Shohamy, 2001). Given the broad impact of tests on individuals and society, language education policymakers, testing specialists, and test users are obliged to strive to minimize the negative consequences of using high-stakes tests of L2 ability and to maximize the positive consequences. This is more likely to occur in a context in which test development and use are viewed as a shared responsibility and where the highest professional standards and best practices occur. In this paper, I have reviewed some of the recent developments and current standards that are being applied to the design and use of large-scale, high-stakes tests of English language ability.

References

ALTE/Association of Language Testers in Europe (2001). *Code of practice*. Retrieved 11, June, 2008, from http://www.alte.org

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition, 10*, 149-164.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. New York: Longman.

Bachman, L. F. (2002). Some reflections on task-based language performance assessments. *Language Testing, 19*(4), 453-476.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
Canale, M. (1983). On some dimensions of language proficiency. In J. Oller (Ed.), *Issues in language testing research* (pp. 333-342). Rowley, MA: Newbury House.
Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.
Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks, and test construction. *Language Testing, 14*(1), 3-22.
Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369-383.
Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.
Chapelle, C. A., Grabe, W., & Berns, M. (1997). *Communicative language proficiency: Definitions and implications for TOEFL 2000* (TOEFL Monograph Rep. No.10). Princeton, NJ: Educational Testing Service.
Chapelle, C. A., Jamieson, J. M., & Hegelheimer, V. (2003). Validation of a Web-based ESL test. *Language Testing, 20*(4), 409-439.
Chen, H., Sinclair, P., Huang, S-y, & Eyerman, L. (January, 2008). APEC EDNET project seminar on language standards and their assessment: Background research paper. Symposium conducted at the APEC EDNET meeting, Xi'an, China.
Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15-37.
Cheng, L., & Qi, L. (2006). Description and examination of the National Matriculation English Test. *Language Assessment Quarterly, 3*(1), 53-70.
Choi, I-c. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing, 25*(1), 39-62.
Crystal, D. (1997). *English as a global language*. Cambridge: Cambridge University Press.
Davies, A., Hamp-Lyons L., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes, 22*(4), 571-584.
Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics 27*, 115-132.
Duff, P. A. (January, 2008). APEC second/foreign language standards and their assessment: Trends, opportunities, and implications. Symposium conducted at the APEC EDNET meeting, Xi'an, China.
Eckes, T., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., & Tsagari, C. (2005). Progress and problems in reforming language examinations in Europe: Cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France and Germany. *Language Testing, 22*(3), 355-377.
Educational Testing Service. (2002). *The ETS standards for quality and fairness*. Princeton, NJ: Educational Testing Service.
Educational Testing Service. (2004). *English language competency descriptors*. Princeton, NJ: Educational Testing Service.
Educational Testing Service. (2007). *TOEIC examine handbook: Listening and reading*. Princeton, NJ: Educational Testing Service.
Educational Testing Service. (2007). *TOEIC speaking and writing sample tests*. Princeton, NJ: Educational Testing Service.

Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics 23*, 282-301.
Falvey, P., & Shaw, S. (2006). IELTS writing: Revising assessment criteria and scales (Phase 5) (*Research Notes 23*). Cambridge: Cambridge ESOL.
Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253-266.
Gottlieb, N. (2008). Japan: Language policy and planning in transition. *Current Issues in Language Planning, 9*(1), 1-68.
Hamp-Lyons, L., & Davies, A. (2008). The English of English tests: Bias revisited. *World Englishes, 27*(1), 26-39.
Jenkins, J. (2006a). Current perspectives on teaching world Englishes and English as a lingua franca. *TESOL Quarterly, 40*(1), 157-181.
Jenkins, J. (2006b). The spread of EIL: A testing time for testers. *ELT Journal, 60*(1), 42-50.
Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, D.C.: AERA.
Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(2), 31-41.
Kaplan, R., & Baldauf, R. (2005). Language-in-education policy and planning. In E. Hinkle (Ed.), *Handbook of research in second language teaching and learning* (pp. 1013-1034). Mahwah, NJ: Erlbaum Associates, Publishers.
Kunnan, A. J. (2008). Large scale language assessments. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education (2nd ed.)*, (Vol. 7, Language testing and assessment, pp. 135-155). New York: Springer.
Norris, J. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing, 19*(4), 337-346.
Norris, J., Brown, J. D., Hudson, T., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395-418.
McNamara, T. (2008). The socio-political and power dimensions of tests. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education (2nd ed.)*, (Vol. 7, Language testing and assessment, pp. 415-427). New York: Springer.
Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Language Testing, 25*(1), 127-143.
Purpura, J. E. (2008). Assessing communicative language ability: Models and their components. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education (2nd ed.)*, (Vol. 7, Language testing and assessment, pp. 53-68). New York: Springer.
Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142-173.
Qian, D. (2008). English language assessment in Hong Kong: A survey of practices, developments and issues. *Language Testing, 25*(1), 85-110.
Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1-32.
Ross, S. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing, 25*(1), 5-13.
Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Pearson.

Stoynoff, S. J. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching, 42*(1), 1-40.
Stoynoff, S. J. (2007). Assessing communicative competence: From theory to practice. In J. Lui (Ed.), *English language teaching in China* (pp. 127-149). London: Continuum.
Tannenbaum, R., & Wylie, E. C. (2005). *Mapping English language proficiency scores onto the Common European Framework* (TOEFL Research Report Rep. No. 80). Princeton, NJ: Educational Testing Service.
Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal, 60*(1), 51-60.
Weir, C. J. (2005a). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281-300.
Weir, C. J. (2005b). *Language testing and validation: An evidence-based approach*. Basingstoke, UK: Palgrave.
Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education (2nd ed.)*, (Vol. 7, Language testing and assessment, pp. 111-122). New York: Springer.
Zenisky, A., & Sireci, S. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362.
Zheng, Y., & Cheng, L. (2008). College English Test in China. *Language Testing, 25*(3), 408-417.