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.