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.