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.