

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.