Language Learning & Technology May 2001, Vol. 5, Num. 2
http://llt.msu.edu/vol5num2/deville/default.html pp. 95-98

LANGUAGE TESTING AND TECHNOLOGY: PAST AND FUTURE

Micheline Chalhoub-Deville
University of Iowa

Technology is increasingly being promulgated as a powerful mechanism that can transform education. It is not surprising then, that the second language (L2) testing field adopted "language testing and technology" as the theme for its 2001 annual conference, the Language Testing Research Colloquium (LTRC). Technology is not a novel topic for language testers. The 1985 LTRC was also organized around this theme. Given the recurring theme and the growing use of technology, it is appropriate to reflect on what has transpired in the L2 testing field related to this topic and to situate the developments within the larger language testing, general measurement, and educational contexts.

The L2 field's first concerted effort in terms of computer-based testing (CBT) emerged in the mid-80s with the 1985 LTRC. The conference proceedings were published under the title *Technology and Language Testing* (Stansfield, 1986). The proceedings indicate that several papers presented at the conference dealt with CBT and the application of latent trait models to item-bank construction, item selection, and computer adaptive testing (CAT). The general measurement profession had been working with CBT and, more specifically, with CAT since the early 70s. The first conference on CAT was held in 1975. Perhaps the main reason the L2 field has lagged behind in this area is because it has long promoted performance-based assessment, a form of assessment that does not lend itself as readily to computerized administration as do more traditional test formats. In fact, the second section of the Stansfield volume deals primarily with performance-based assessment. So, whereas general measurement researchers, especially those working with CAT, have concerned themselves more with selected-response item types, the L2 field has continued to promote performance-based assessment. Even today, CBT performance-based assessment continues to be a challenge.

After Stansfield (1986), the most significant work on language testing and technology is an edited volume by Dunkel (1991). Dunkel's volume is devoted to computer-assisted language learning and CAT. The CAT section documents how knowledge in the L2 field has progressed in the area of CBT. For the most part, the studies continue to explore the value of the adaptive format and report on various CAT developments and validation research efforts. What is most significant about this volume, however, is the variety of CAT applications for assessing L2 proficiency in school and university settings, a clear indication of the growing use of CAT.

Indeed, since the Dunkel volume, the number of CBT and CAT instruments developed by academic and testing organizations has continued to increase. Many universities and virtually every major language testing organization are engaged in the development of CBT or CAT. Brigham Young University was a pioneer in developing French, German, and Spanish CAT instruments used for placement at universities. The Educational Testing Service in 1998 launched CBT TOEFL in the in US and in numerous countries around the world. The University of Cambridge Local Examinations Syndicate (UCLES) also developed CAT instruments in various languages and for various purposes. More specifically, UCLES developed CommuniCAT primarily for language programs in academic settings and BULATS (Business Language Testing Service) for the corporate sector. Languages targeted in these tests include English, French, German, and Spanish. In addition, the Council of Europe has sponsored the DIALANG project, which provides diagnostic assessment in 14 languages.

Coupled with this proliferation of CBT and CAT instruments has been a steady flow of publications on the topic. These publications include those by Brown and Iwashita (1996, 1998); Young, Shermis,

Brutten, and Perkins (1996); Shermis (1996); Burstein, Frase, Ginther, and Grant (1997); Brown (1997); Dunkel (1997); Chalhoub-Deville, Alcaya, and Lozier (1997); and Chalhoub-Deville and Deville (1999).

The most recent book dealing with CAT appeared in 1999. The Chalhoub-Deville (1999) edited volume is devoted entirely to CAT issues, with experts from both the L2 and general measurement fields invited to share their research knowledge and experience in three interrelated areas: the L2 reading construct, L2 CAT applications, and IRT (internet-related technologies) measurement issues. The volume highlights the importance of a construct-driven approach to CAT work. This is exemplified most clearly in the discussion chapters that explore and identify links among the various areas to advance more systematic and construct-based CBT and CAT development.

Over 15 years have passed since LTRC first focused on technology, and one can argue that great strides have been made in the area of CBT and CAT. The computerized delivery of tests has become an appealing and a viable medium for the administration of standardized L2 tests in academic and non-academic institutions. Additionally, a reasonable body of research exits in this area. Given the growing use of and research on CBT, an important issue to consider is the nature of the change that this mode of assessment has introduced to L2 testing.

An examination of the changes brought forth by L2 CBT shows that technology has been intended primarily to help make assessment more efficient and serviceable, what Christensen (1997) calls "sustaining" innovations. CBT allows, among other things, more flexible and individualized test administration, tracking of student performance, immediate test feedback, new item/task types, and enhanced test security. Perhaps one of the most exciting capabilities of CBT is the adaptive approach, which one might argue, using Christensen's terminology, is a "disruptive" technology, that is, a technology that changes how we think of and implement our operations. CAT permits the tailoring of item difficulty to the test taker's performance, allowing a more accurate assessment of the examinee's L2 ability. But apart from the adaptive innovation, CBT has been utilized mainly to facilitate test delivery and administration. What is needed, therefore, is to explore how technology can engender fundamental changes in the L2 CBT endeavor. In a forthcoming paper (Chalhoub-Deville, in press), I discuss various issues in areas of L2 CBT can be regarded as a disruptive use of technology. Areas covered in this discussion include the representation of the L2 construct, overall test design, item/task construction, and test purpose.

In terms of construct representation, it is well documented that the L2 construct is multidimensional and involves a variety of interacting components and processes (Bachman, 1990; Bachman & Palmer, 1996). Language testers need to utilize technology to design measures that increasingly explore and better measure such critical aspects of the construct. Additionally, researchers have argued that some abilities and processes, which are critical for beginning language learners, become less salient for more proficient learners, for whom yet other aspects of the construct begin to emerge (Bernhardt, 1991). Technology provides an excellent capability to trace test takers' language development thus enabling researchers to better understand how aspects of the construct evolve across different ability levels.

Technology might also help test developers move beyond conventional test design procedures, which provide scores primarily to rank students, to other procedures that can facilitate a more systematic test design approach, one which creates interrelations among task characteristics, test takers' performances, and inferences about intended underlying abilities and processes. Such an approach would produce a richer and more meaningful depiction of test takers' abilities. An example of such an integrated and construct-based approach to CBT design is Portal (see Mislevy, 1996; Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999). Portal utilizes computer technology and alternative measurement models to examine test takers' performance on test tasks with documented features and provides rich information about underlying language components and processes.

In a similar vein, technology can also facilitate a more meaningful approach to task development and enable test developers to draw more defensible inferences by establishing a closer link between task creation and underlying abilities. Prototype tasks with identified characteristics, based on a systematic analysis, can be fed into a database used to generate new tasks with the desired linguistic, cognitive, situational, and measurement characteristics. Additionally, computer technology advances now permit the use of more complex tasks in L2 tests. For example, simulation tasks allow test developers to elicit contextualized, integrated performances that closely resemble those in real-life L2 interactions. With the aid of technology, simulations with identified characteristics allow relevant features to be manipulated in a structured manner in order to target intended ability levels.

Spolsky (1997) argues that the main purpose of today's L2 standardized tests reflect the "gate keeping" needs that emerged because of the increased educational demands and limited instructional resources experienced at the beginning of the 20th century. Technological advances, however, are fast transforming L2 learning opportunities. The proliferation of distance learning programs will likely result in a decreased need for selection testing and an increased need for assessments that grant credentialing or certification. Similarly, computer-delivered tests that assess and diagnose a person's language and skill development will be in greater demand. In fact, such tests are already available. DIALANG, mentioned above, is an example of this new generation of computer-delivered assessments, which have been developed to meet the needs of learners in non-conventional classroom settings.

In conclusion, computer technology has enhanced the efficiency of many of our L2 testing practices and introduced notable innovations such as CAT. But most L2 CBT and CAT instruments available on the market fall short in providing any radical transformation of assessment practices. Advances in technology should encourage test developers to move beyond the thinking that has long dominated paper-and-pencil testing and inspire the use of "disruptive" applications, by which assessments are conceptualized and implemented in innovatively different ways.

ABOUT THE AUTHOR

Micheline Chalhoub-Deville is an Associate Professor of Foreign Language and ESL Education at the University of Iowa. She has published in the journals Language Testing and Language Learning and System. She also has directed federally-funded computer adaptive test projects. Dr. Chalhoub-Deville has received the International Language Testing Association 1995 Award for Best Article on Language Testing and the TOEFL 2001 Outstanding Young Scholar Award.

E-mail: m-chalhoub-deville@uiowa.edu

REFERENCES

Bachman, L. F. (1990). *Fundamental consideration in language testing.* Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice.* Oxford, UK: Oxford University Press.

Bernhardt, E. (1991). *Reading development in a second language: Theoretical, empirical, and classroom perspectives.* Norwood, NJ: Ablex Publishing Corporation.

Brown, A., & Iwashita, N. (1996). Language background and item difficulty: the development of a computer-adaptive test of Japanese. *System*, *24*(2), 199-206.

Brown, A., & Iwashita, N. (1998). The role of language background in the validation of a computer-adaptive test. In A. Kunnan (Ed.), *Connecting fairness and validation in language testing* (pp. 195-207). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 44-59. Retrieved August 15, 1998 from the World Wide Web: http://llt.msu.edu/vol1num1/brown/default.html.

Burstein, J., Frase, L., Ginther, A., & Grant, L. (1997). Technologies for language assessment. *Annual Review of Applied Linguistics, 16,* 240-260.

Chalhoub-Deville, M. (Ed.). (1999). *Issues in computer adaptive testing of reading proficiency.* New York: Cambridge University Press.

Chalhoub-Deville, M. (in press). Technology in standardized language assessments. In R. Kaplan (Ed.), *Handbook of Applied Linguistics* (pp. x-x). Oxford, UK: Oxford University Press

Chalhoub-Deville, M., Alcaya, C., & Lozier, V. M. (1997). Language and measurement issues in developing computer-adaptive tests of reading ability: The University of Minnesota model. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 546-585). Jyväskylä, Finland: University of Jyväskylä.

Chalhoub-Deville, M., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics, 19,* 273-299.

Christensen, C. (1997). *The innovator's dilemma: When new technologies cause great firms to fail.* Boston, MA: Harvard Business School Press

Dunkel, P. (Ed). (1991). *Computer assisted language learning and testing: Research issues and practice.* New York: Newbury House.

Dunkel, P. (1997). Computer-adaptive testing of listening comprehension: A blueprint for CAT development. *The Language Teacher Online, 21,* 1-8. Retrieved August 15, 1998 from the World Wide Web: http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/97/oct/dunkel.html.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379-416.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive analysis, with implications for designing simulation-based performance assessment. *Computers in Human Behavior, 15,* 335-374.

Shermis, M. (1996). Computerized adaptive testing for reading placement and diagnostic assessment. *Journal of Developmental Education, 38*(1), 45-52.

Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learned in one hundred years? *Language Testing, 14,* 242-247.

Stansfield, C. (1986). *Technology and language testing.* Washington, DC: TESOL.

Young, Y., Shermis, M. D., Brutten, S. R., & Perkins, K. (1996). From conventional to computer-adaptive testing of ESL reading comprehension. *System, 24,* 23-40.