Annual Review of Applied Linguistics (2007) **27**, 115–132. Printed in the USA.
Copyright © 2008 Cambridge University Press 0267-1905/08 $12.00
doi: 10.1017/S0267190508070062

6. ASSESSING LANGUAGE USING COMPUTER TECHNOLOGY

Dan Douglas and Volker Hegelheimer

In this article, we propose to follow up on the most recent ARAL survey article on trends in computer-based second language assessment (Jamieson, 2005) and review developments in the use of technology in the creation, delivery, and scoring of language tests. We will discuss the promise and threats associated with computer-based language testing, including the language construct in relation to computer-based delivery and response technologies; computer-based authoring options; current developments; scoring, feedback, and reporting systems; and validation issues.

In her 2005 review of trends in computer-based language assessment, Jamieson first discussed early developments, which meant, really, the delivery of paper-and-pencil tests by computers, the main perceived advantages being convenience and standardization of delivery. She then outlined "value-added" benefits of computerization, focusing on computer-adaptive technologies, which theoretically provided tests that could be shorter, more efficient, and better suited to test takers' abilities. The next steps, according to Jamieson, involved more substantive advances in how language ability is perceived. This involves the exploitation of computer technology in presenting input and response formats that reflect more accurately relevant aspects of the target language use domain than is possible in paper-and-pencil formats. Technology also enhances the ability to provide more tailor-made reports to test takers and score users that interpret specific aspects of the test taker's performance that are lost in paper-and-pencil formats. Finally, Jamieson noted a number of possible future applications of computer technology, including the use of corpora to generate vocabulary lists and specific purpose discourse, realistic simulations by means of multimedia and virtual reality, and the capture of many aspects of a test taker's progress through the test. She speculated, though, that the automated rating of written essays is perhaps the most likely near-future technological development, and concluded with a brief discussion of the use of innovative item types and multimedia technologies in language testing.

Jamieson brought her survey of trends to a close with a combined note of pessimism and hope:

> We can also see that the future envisioned by many of these commentators is the distant future. . . . Although we may be mesmerized by panoramic visions, a journey begins on a path where we place one foot in front of the other, making what appears at first to be modest progress. (p. 238)

In the few years since Jamieson's article, what modest progress has been made in assessing language using computer technology? We will discuss the promise and threats associated with computer-based language testing, including the language construct in relation to computer-based delivery and response technologies; computer-based authoring options; current developments; scoring, feedback, and reporting systems; and validation issues.

The Language Construct in Relation to Computer-Based Delivery and Response Technologies

A fundamental issue since the advent of computer-based language testing (CBT) has been whether and how the delivery medium changes the nature of the construct being measured. This was particularly a concern in the early days when existing paper-and-pencil language tests (PPTs) were simply mounted on a computer, and little was known about how such issues as reading on a screen rather than on a printed page, clicking with a mouse rather than darkening a box with a pencil, or keyboarding an extended response rather than handwriting it might affect test performance and the inferences that might be made as a result. A number of studies over the years have suggested that there is little effective difference between performance on PPTs and CBTs by groups of test takers (e.g., Breland, Lee, & Muraki, 2004; Choi, Kim, & Boo, 2003; Coniam, 2006; Wolfe & Manalo, 2005). However, it is almost certainly the case that subgroups of test takers may be differentially affected by computer technology and therefore perform differentially on a CBT, though there is very little research on this issue. Wolfe and Manalo (2005) did find that female test takers, examinees whose native language is not based on a Roman/Cyrillic alphabet, and examinees with lower English proficiency were more likely to write essays by hand rather than keyboarding, and that handwritten essays tended to receive slightly higher ratings than those written by word processor. Apart from this, we have not been able to find a single study directly investigating the impact of computer familiarity on CBT performance, and the upshot is that we still do not know with any certainty how computer technology in language tests affects individual test takers who may be advantaged or disadvantaged with respect to computer skills. In any case, even if it were true that there are no significant differences between performances on PPTs and CBTs, this begs the question, as Chapelle and Douglas (2006) argued, of whether the costs of computer technology justify alleged gains in efficiency and accuracy. Surely, the switch from paper-based to computer-based tests ought to be undertaken to achieve a *better* measure of the construct, not simply a more efficient one, which brings us to a consideration of the potential of computers to provide what Jamieson (2005) called “computerized tasks that better represent authentic language use” (p. 233).

In what sense can computers engage test takers in more authentic language use tasks? Certainly computers, the Internet, and related technologies, such as text messaging and podcasting, have become pervasive in how we communicate. The ability to use information technology, including e-mail, word processing, data handling, and computer presentation graphics (PowerPoint), is a necessary part of what it means to be literate in professional and social life today. Chapelle and Douglas (2006) have suggested that “communicative language ability needs to be conceived in view of the joint role that language and technology play in the process of communication” (p. 108), and recommended that language ability needs to be defined in terms of an interaction between language and technology: “the ability to select and deploy appropriate language through the technologies that are appropriate for a situation” (p. 107).

There are times when e-mail is an appropriate means for communication, while at others, a telephone call might be better, and still others when face-to-face communication is required. In this way of viewing the language construct, what has been traditionally defined simply as “reading ability,” for example, might now need to be more specifically defined as the ability to gather visually presented information from the Internet—as well, of course, as the ability to gather visually presented information from the printed page. For language testing there is no doubt that the addition of technology means a more complex construct, and this leads to difficulties for test developers: Do we need truly multimedia tests in which the test taker must switch from print to electronic text to voice mail?

Another aspect of the integration of computers and authentic language use involves the ability of computers to present contextualization cues in the form of still images, audio, animation, and video. This, as it turns out, is not as straightforward an issue as it might appear on the surface. It seems reasonable to assume that the use of, say, a video of two people in conversation in a listening test would enhance authenticity by providing a context and nonverbal cues for the test takers. However, there have been questions raised about whether test takers actually look at the video, preferring to focus solely on the audio input (e.g., Gruba, 1997). There is also debate about whether nonverbal information provided by video input should be a part of the listening construct. With regard to the first concern, Wagner (2007) conducted a small study to investigate the extent to which test takers watched the video in a listening test. He found that on average, across a variety of video texts, participants in his study oriented to the video 69% of the time, with some watching as little as 15% of the time and others as much as 90% to 100%. This is the first study to date that has provided empirical evidence on the question of test-taker performance with regard to video input.

The question of whether nonverbal input—gestures, physical orientation, facial expressions, and vocal articulation—should be part of the listening construct can be resolved by reference to the inferences that are to be made based on test performance. There are situations when nonverbal information is not a relevant part of the construct, for example, if the test is to be used to make inferences about a call center employee's ability to speak on the telephone, or an air traffic controller's ability to engage in radio communication with a pilot. However, there are also many situations, perhaps the majority, when visual input and verbal input are "inextricably intertwined in the communication of the total meaning of an interpersonal exchange" (Burgoon, 1994, p. 47), as in face-to-face conversations and lectures and, of course, while watching films and videos.

Finally, there is the question about the extent to which multimedia input actually enhances the authenticity of a language test. To our knowledge, there has been no empirical research directly aimed at exploring this issue, at least in the sense called for by Gruba (1997), who advocated the abandonment of comparative studies of tests with and without video input in favor of those that investigate how nonverbal input is used by learners and test takers, through such methods as verbal reports, observations, and interviews. Wagner (2007) also called for research into the extent to which attention to video input makes a difference to test performance by correlating test takers' time of orientation to the visual input with their scores on the test.

There are certainly complex questions associated with the construct being measured by computer-based language tests, and we have reviewed what seem to us to be the major unresolved issues in need of empirical investigation: the joint role of technology and language in defining the construct to be measured, the effect of nonverbal input as part of the listening construct, and the effect of multimodal input on authenticity.

Computer-Based Authoring Options

Although numerous tools of varying complexity are available for the development of computer-based and/or Web-based language tests, an exhaustive review is beyond the scope of this article. Instead, a cross section of tools that are being used for language assessment purposes is presented.

Particularly for Web-based testing, developers frequently rely on tools that are either embedded in course management systems (CMSs) or available as add-on extensions because at least some tracking of student performance on the language assessment tasks is automatically included. One increasingly used freely available open source course management system for the delivery on online courses in general, but also specifically for courses related to language learning and teaching, *Moodle* (http://www.moodle.org/), presents developers with several options to develop language tests including response types such as multiple-choice, multichoice, true-false, fill-ins, matching, and short answers using an exact-matching strategy (Brandl, 2005). Other free CMSs that enjoy a loyal following include Drupal (http://drupal.org/) and Joomla (http://www.joomla.org/).

Other very commonly used commercial CMSs include Desire2Learn (http://www.desire2learn.com/), Blackboard (http://www.blackboard.com/us/index.

Bb), and WebCT (http://www.webct.com/), which, like Moodle, provide options for test construction. Sim, Horton, and Strong (2004) outlined numerous interface issues when WebCT is used as the delivery mechanism for assessments, which mirrors the experience of other language testers. Among other problems, Sim et al. pointed out that answering and proceeding to the next question is far from intuitive, it is not possible to change the answer once it has been submitted, and the choice to click on "Next Question" on the last item takes learners to the beginning of the test; and they found inconsistencies between messages (instructions) and required actions.

Although still time-consuming and not always intuitive to manage, the advantages of using a CMS can outweigh the limitations. The advantages include tracking if and when learners log on to the course Web page and the assessment tasks, tracking responses to individual tasks, imposing time restrictions, and setting the number of attempts. The future will likely see the development of modules and test generation tools specifically aimed at the production of language tests.

In contrast to CMS-related tools, stand-alone applications such as *Hot Potatoes* (http://hotpot.uvic.ca/) continue to be used in language testing situations, with an emphasis on low-stakes tests (e.g., self-assessment tests) that do not require item-level data collection and security features such as authentication (see, e.g., Winke & MacGregor, 2001). Another example of a stand-alone, though not free, application, *Flashform Rapid eLearning Studio* (Rapid Intake: http://www.rapidintake.com/index.htm) takes advantage of the increased prevalence of Flash on the World Wide Web. Creating quizzes is straightforward, items can be randomized, and feedback can be provided; however, it is limited in terms of test tasks.

Discovery School's Quiz Center (http://school.discovery.com/quizcenter/quizcenter.html) is a free online tool that allows developers to put language tests online that require multiple-choice, short-answer, true/false, or essay-type test tasks. Quizzes can also be randomized, different types of matches for short answer questions (exact, inexact match, or multiple possible correct answers) are possible, and users have the option of sending actual responses and scores to the test taker and the test administrator.

Two more flexible and more powerful—and also more expensive—developments aimed at educational testing in general and applicable to language testing in particular are Question*mark*^{TM} Perception and LidgetGreen. Question*mark* Perception (http://www.questionmark.com/us/index.aspx) is described as a "comprehensive Assessment Management System to Author, Schedule, Deliver and Report on Surveys, Quizzes, Tests and Exams." Like many other commercially available applications, it can be integrated with other systems. An authoring wizard makes it easy to create 22 different question types (http://www.questionmark.com/us/perception/authoring_windows_qm_qtypes.aspx), including standard multiple-choice, drag and drop, matching, cloze, text-matching, and the like. It is also possible to add multimedia files including audio, video, *Adobe Flash,* and *Adobe*

Captivate to Question*mark* Perception tasks. Although Question*mark*^{TM} Perception is used by educational and corporate organizations, its use in language testing was subject of a brief debate on LTEST-L (LTEST-L Archive, 2006b) in which some shortcomings of the product for language testing were pointed out. LidgetGreen (http://www.lidgetgreen.org/) focuses on educational assessment and provides services that include test design and development, item writing, validation, and research. LidgetGreen also offers a placement test aimed at university-level intensive English programs: college academic English (http://ept-college.com/). Its *Test Generation and Online Delivery System* (TGOD), written in Java, consists of three main functions: an administration module, an online test creation component, and an online test delivery component. There are 10 different reading item types and 9 listening item types. Speaking and writing item types are currently under development.

Both companies provide more sophisticated options in terms of test tasks and provide users with authentication services, which make them possible choices for higher-stakes testing. However, both approaches require financial resources that may be beyond what individual language test developers typically have access to. Language test developers with a sufficient budget may find that creating their own authoring systems will be more responsive to their needs.

All these tools equip even novice developers to arrive at functional computer-based or Web-based language tests. However, a crucial aspect that has not received as much attention, according to Fulcher (2003), is that of interface design issues related to computer-based language testing. Fulcher provided detailed checklists dealing with issues such as hardware, software, and interface design issues such as navigation, page layout, text, color, toolbars, graphics, and help facilities.

Decisions about which computer-based authoring options to pick are mandated by needs, expertise and finances. Freely available test development programs (e.g., Hot Potatoes, Quiz Center, Flashform Rapid eLearning Studio) continue to be useful and appropriate for small-scale, low-stakes test. Embedding language tests in course management systems adds an additional layer of security and tracking possibilities, while commercial products such as TGOD and Question*mark* Perception offer more flexibility at higher cost. Developers of authoring systems could begin paying attention to innovations in language testing (e.g., DIALANG's examples of innovative language testing item types: http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/expmenu.htm) and provide flexible templates for the types of tasks that are more relevant to interactive language use.

In spite of the availability of a number of authoring systems as well as expertise to build tailor-made authoring tools, we have not been able to find any examples of truly innovative, groundbreaking new computer-assisted language test (CALT) projects in the last 4 or 5 years. For informative reviews of a number of currently available CALTs, see Stoynoff and Chapelle (2005).

We might also point out that there is currently little work in the area of computer-adaptive testing (CAT) as discussed in a recent thread on LTEST-L (LTEST-L Archive, 2006a). The basic point is that there are appropriate (e.g., low to medium stakes, specifically defined context and audience) and inappropriate (e.g., large-scale, high-stakes, broadly defined context and audience) uses of CATs. Thus, what was once thought to be the wave of the future in language testing, that is, computer-adaptive tests, has been tempered by the realization of the limits of this particular technology.

Scoring, Feedback, and Reporting Systems

As Jamieson (2005) suggested, perhaps the most likely computer-related innovation in the near term is the development of automated essay scoring systems, and indeed, since 2005, several studies investigating automated essay scoring applications have been published. These include survey articles (e.g., Dikli, 2006) highlighting various essay scoring systems such as *Project Essay Grader* (PEG: http://134.68.49.185/pegdemo/), *Intelligent Essay Assessor* (IEA: http://www.knowledge-technologies.com/prodIEA.shtml), *E-rater* and *Criterion* (http://criterion4.ets.org/cwe/), *IntelliMetric*® (http://www.vantagelearning.com/intellimetric/), *MyAccess* (http://www.gomyaccess.com/myaccess/do/log), and the *Bayesian Essay Test Scoring System* (BETSY: http://edres.org/betsy/). First proposed in the 1960s, computer scoring has seen different approaches from the PEG, which relied on statistical procedures and regression analyses on various variables in an effort to predict essay scores assigned by human raters, to the IEA, which uses *Latent Semantic Analysis* (LSA), an approach that requires initial training on sample essays so that essays can be characterized by LSA vectors, which Dikli (2006) explained as a "mathematical representation of the essay" (p. 6).

Researchers have also continued their examination of specific scoring engines (e.g., Pearson Knowledge Technologies, 2006; Rudner, Garcia, & Welch, 2005) focusing on psychometric research often sponsored by the system developers, primarily conducted for validation reasons. Thus, these should be interpreted with caution.

Warschauer and Ware (2005) pointed out that much less work with automated writing evaluation (AWE) systems has been done in languages other than English, with some notable exceptions (Cohen, Ben-Simon, & Hovav, 2003, with Hebrew; Ishioka & Kameda, 2004, with Japanese), but suggested that systems using latent semantic analysis should work well across languages because "the underlying procedures of such analysis are language blind" (p. 9). The authors noted that AWE software holds great potential for research investigating the writing processes in part because it is set up in such a way that it keeps track of writing development over time as it stores various versions of writings, making it an ideal vehicle for research possibilities. Thus, this line of research supplements research focusing on the written end product, which tends to treat the use of the software as a black box.

Monaghan and Bridgeman (2005) identified one key aspect of automated essay evaluation, public acceptance of scores assigned by automated essay rating. "Automated essay evaluation systems, such as E-rater, have a very high threshold to meet to gain people's full confidence as a valid scoring approach" (p. 3). Addressing this need, the authors outline a series of steps they have taken to: (1) compare E-rater scores to human rater scores and (2) ensure that concerns for a transparent decision-making process are met. Consequently, E-rater "run[s] in the background" (p. 2) to check scores that were assigned by human raters. They cite Burstein's research and development efforts (e.g., Burstein & Chodorow, 2002), including the development of a scoring model for more than 100 prompts of issue- or argument-type essays that yielded relatively high levels of agreement. Monaghan and Bridgeman concluded that for now, one human rater plus E-rater appears to be the most promising constellation.

Another promising area of research is in the automatic identification of off-topic essays. One difficulty is that many programs require extensive training, that is, 200–300 essays on the same topic. Higgins, Burstein, and Attali (2006) reported on the development of an algorithm that minimizes the need for such large-scale training. Their algorithm requires "only text of existing essay questions, the text of the new essay question, and the student essay in order to predict off-topicness" (p. 158).

Work with speech recognition as part of automated scoring of nonnative speech is another area showing progress since the last review by Jamieson. In light of the inclusion of speaking performance on TOEFL® (Test of English as a Foreign Language) iBT (Internet-based test), more research addressing this topic is likely. Zechner, Bejar, and Hemat (2007), guided by the question of whether automated procedures for identifying evidence of speaking performance can be developed, evaluated an off-the-shelf speech-recognition system to score speaking prompts from a 2002 *LanguEdge* field test. They concluded that even though the challenges were great, the results of their study indicated that additional investment and research were feasible in the area of speech recognition technology. Provided continuous improvement is evident, they suggested that the investment of time and finances "is reasonable because of the potential cost savings to the growing number of testing programs offering speaking assessment, and the resulting increased competitiveness afforded by the ability to score speaking tests at a lesser cost" (p. 3).

Much research since 2004 has focused on the automated scoring of writing, as Jamieson (2005) predicted. Issues such as identification of off-topic essays and the effects of essay length on scoring continue to be an ongoing focus of research (e.g., Chodorow & Burstein, 2004). Another fruitful avenue of research lies in the study of the applicability of automated writing assessment in classroom settings. We should also point out, on the topic of scoring and feedback systems, the impressive example of DIALANG (http://www.dialang.org/). This self-assessment system, sponsored by the European Union, provides learners with information and advice on their levels of attainment in reading, writing, listening, grammar, and vocabulary in 14 languages. The system provides feedback on a test taker's level on the Council of Europe Framework of Reference (CEFR) scale, an indication of what that level means, an explanation of the initial placement test in light of the test taker's own self-assessment, a review of right and wrong answers, and specific strategies for improvement. There is research to support these feedback claims (Alderson, 2005; Alderson & Huhta, 2005), though Chapelle (2006) cast modest doubt on some of them. Nevertheless, as an example of the potential for low-stakes computer-based assessment, and for the staggering accomplishment of producing putatively equivalent assessments and practical feedback in 14 different languages, the DIALANG project is worthy of our interest and admiration.

Validation Issues

In language testing, validation is a process through which evidence is collected to support interpretations of test performance. For example, on the Listening subtest of the ACT ESL Placement Test (2007a), a score between 82 and 91 is interpreted as placing test takers at Level 3 on the Listening Proficiency Descriptors, which indicates, in part, that students "are able to understand most discourse about personal situations and other everyday experiences, including conversations with basic academic and/or occupational subject matter" (ACT, 2007b). This descriptor thus constitutes a claim that needs to be supported by validation research. In principle, validation issues are no different for computer-based tests than they are for any other forms of assessment: Whatever claims may be made for interpreting test performance require evidence.

However, CBT has been associated since its inception with possible threats to validity (e.g., Canale, 1986) specific to computer and Web-based technology. Chapelle and Douglas (2006) have summarized a number of these as follows:

> 1. Performance on a computer-delivered test may fail to reflect the same ability as what would be measured by other forms of assessment.
> 2. The types of items that can be developed in computer-assisted formats are different from those that can be developed with other media.
> 3. Selection of items to be included on an adaptive test by an algorithm may not result in an appropriate sample of test content and may cause test takers anxiety.
> 4. Computer-assisted response scoring may fail to assign credit to the qualities of a response that are relevant to the construct that the test is intended to measure.
> 5. CALT may pose risks to test security.
> 6. CALT may have negative impact on learners, learning, classes, and society.
>
> —Chapelle & Douglas (2006, p. 41)

With regard to the first threat, studies comparing performance on CBTs with that on other delivery formats (e.g., Choi et al., 2003; Coniam 2006) have shown quite small or mixed differences. For example, Coniam (2006) found some significant differences in performance on a computer-adaptive listening test compared with that on a 55-item paper-based listening test created from the same item pool favoring the CBT version. Coniam concluded that the CBT would be adequate for a low-stakes test but less so for a higher-stakes purpose. In addition, Coniam found that boys preferred the computer version, while girls preferred the paper-based test. This gender-related difference reminds us of the possibility for gender and regional differences in computer familiarity that can potentially affect test performance and therefore the validity of interpretations. The large-scale survey of computer use conducted by Educational Testing Service (ETS) in 1996–97 (Taylor, Jamieson, & Eignor, 2000) showed great regional disparities in computer use. The caution for developers of computer-based tests, for both international and local use, is that they should investigate carefully differences in computer familiarity among their intended populations of test takers.

Important validity research is centered on tasks used in the new Internet-based TOEFL (iBT). Cohen and Upton (2006) studied the reading and test-taking strategies test takers reported using while doing the TOEFL reading tasks, and Douglas and Hegelheimer (2006) similarly explored strategy use reported during the iBT listening test. In another study, Carrell (2007) investigated the effect of note taking on listening and other communicative tasks. Finally, Cumming et al. (2006) studied the prototype “integrated” writing tasks on the iBT to see whether and how writing produced on such tasks differed from that elicited by more traditional essay tasks.

Cohen and Upton (2006) found that in general, the test takers tended to approach the test tasks *as* test tasks rather than as genuine reading tasks. In other words, they were mainly interested in getting the correct answer rather than learning anything from the reading. The researchers concluded, however, that the tasks did appear to elicit academic reading-like skills, and thus found support for the ETS validity claim that the test required test takers to engage their reading skills to answer the questions. Douglas and Hegelheimer (2006) found that iBT listening test takers employed strategies similar to those hypothesized to be used in nontest listening, though in a finding similar to that of Cohen and Upton, it appeared that test takers listened for meaning only to the extent necessary to respond to the question, and thus were not truly engaged in academic listening.

In the iBT, test takers are allowed to take notes during the listening test if they wish. Carrell (2007) investigated their note taking both before and after a brief tutorial on good note-taking practices. She found that only about 20% of the information from the input that was subsequently assessed actually appeared in the notes, but that the notes of test takers who included more content words and those who recorded test answers in their notes bore a stronger relationship to test performance than did the notes of those who failed to include such information.

Cumming et al. (2006) found substantive differences in the essays produced in response to the integrated prompts (e.g., print or audio source texts) compared to those produced in response to the traditional TOEFL essay. Their findings provide support for certain of the claims made by the TOEFL developers regarding the differences between the two types of writing tasks and justification for including both task types on the iBT.

The second threat to CBT validity, related to different item and task types that can be devised for computer and Web delivery, has to do not so much with how test takers respond to the computer interface as with the limitations the medium places on test developers to create tasks for computer delivery that reflect the desired attributes of natural language use (Carr, 2006). Although we usually think of the potential for computers to *enhance* the possibilities for presenting lifelike language tasks, we have to remember that computers, like any other medium, are limited in what types of language-use tasks they can process. For example, no one has yet developed a computer-based simulation that adequately reflects truly interactive oral language (although some *chat-bots* are beginning to explore the potential: http://lauren.vhost.pandorabots.com/pandora/talk?botid=f6d4afd83e34564d). On a more fine-grained level, as Chapelle and Douglas (2006) pointed out, the TOEFL item type in which the test taker is presented with a sentence to place appropriately in a reading text, is constrained by the fact that the text must ideally fit onto a single computer screen, usually capable of presenting less text of a given font size than is a printed page.

Chun (2006) investigated the test method characteristics and validity of the SET-10 telephone-delivered, computer-scored speaking test (now called *Versant,* Harcourt Assessment, 2006) with respect to one of its stated purposes, screening job candidates' ability in spoken English. Chun questioned the claim of the publishers with regard to the use of the test for employment screening. As always, test developers and score users must consider the language use situation for which the test is being developed and determine whether task types possible on the computer or the Internet are appropriate for the inferences and decisions the test users wish to make.

The selection of computer-adaptive test items by a computer algorithm, the third threat to CBT validity, raises concerns about adequate sampling of item types and test content. For example, Coniam (2006) suggested that the algorithm driving his listening CAT might be partly to blame for the rather low correlation between the paper version and the adaptive version of his listening test. In another example, the ESL Computer Adaptive Placement Exam (SoftStudy, 2006), a medium-stakes placement instrument originally produced at Brigham Young University in 1999, is made available for sale to other institutions. It runs on a one-parameter algorithm that selects items based solely on their level of difficulty (Roever, 2005) and provides a placement recommendation at one of six levels. Although each test taker thus receives a different combination of items, and indeed, a different number of items, the fact that there are three sections (grammar, reading, and listening) that are tested independently probably minimizes the effect of differential sampling—all the test developers and score users are concerned with is determining a test taker's ability to answer questions at a given level.

The fourth threat to the validity of CBTs lies in what Jamieson (2005) referred to as "the most probable advancement in technology that will occur in large scale language assessment in the near future" (p. 235), the automatic scoring of responses, including such tools as those provided by E-rater, IntelliMetric, and Ordinate (2006). We have discussed current issues concerning this feature of computer-based language testing in the previous section. Perhaps the most vexing aspect of this validity threat is the problems associated with polytomously scored responses, specifically deciding how to weight partially correct responses. Chapelle and Douglas (2006) pointed out that dichotomously scored responses avoid the difficulty of such decisions as what makes a response correct and which responses might be more correct than others. These problems become particularly acute when test developers decide to assign partial scores to responses, because it must be determined not only how much weight to assign various responses but also on what basis to make the decision. Alderson, Percsich, and Szabo (2000) provided an instructive demonstration of the type of research needed to respond to these issues.

The next threat to CBT validity involves security issues, particularly to do with the identity of test takers, the vulnerability of the Internet to computer hackers, and the exposure of items in CATs. Regarding the first two, which are especially important in high-stakes, large-scale testing, as in the TOEFL program, we can only await developments in technology to help us refine our efforts to ensure that the person whose score is reported is the person who actually took the test, and that the servers on which the tests reside are as impervious to thieves as possible.

Concerning the particular problem of exposure of items in computer-adaptive tests to test takers who may memorize them and report them to friends or unscrupulous organizations, the response so far has been to resort to very large item pools. But, as has been pointed out in a discussion of computer-adaptive testing on the language testing research Internet discussion list (LTEST-L), on paper-and-pencil tests all test takers see the same limited set of items—if they collectively memorize them, they have memorized only, say, 100 items on a 100-item test. On a CAT, however, since each test taker sees a different constellation of items from a large pool, a team of test takers will potentially be able to memorize several times as many items as on a linear test (Green, 2006). This problem led Fulcher (2005) to predict the demise of computer-adaptive language testing in large-scale programs: "We have learned that computer adaptive testing is hideously expensive in large-scale tests. Item pools became vats, then lakes, then oceans, just to maintain test security. . . . The era of adaptivity in mass international language testing is dead." (http://education.guardian.co.uk/tefl/story/0,5500,1645011,00.html)

Fulcher was responding to the decision in the TOEFL program to abandon the very costly computer-adaptive sections in the computer-based TOEFL in favor of a return to linear subtests in the more recently introduced Internet-based test (iBT). There are, of course, relatively large-scale CAT language tests that have been able to overcome this problem, most notably, perhaps, the Business Language Testing Service (BULATS: http://www.bulats.org/) test (Cambridge ESOL, 2007; see also the discussion of CATs in Carr, 2006). The security problem must be weighed in consideration of the testing context and purpose.

The final threat to computer-based testing is that of possible negative consequences to learners, programs, and society owing to the often huge cost of computerization, which may divert resources away from other needs. Computerization can also produce negative washback on teaching and learning when learners, teachers, and materials writers focus on practicing the limited skills required by some computer-based tests rather than on a broader range of communicative activities. As Chapelle and Douglas (2006) pointed out, few of these negative consequences have actually been documented in the research literature, though there is ample anecdotal evidence among professionals in the language-testing field. Most notable is a discussion of the negative consequences of computer-adaptive testing on LTEST-L in 2006 (LTEST-L Archive, 2006a).

Conclusion

In this review of the construct of computer-assisted language tests (CALTs), authoring systems, automated scoring, and threats to validity associated with CALT, we have seen that, in general, computer-assisted language tests have not fully lived up to their promise, but that research and development of CALTs continues in interesting and principled directions. Specifically, we noted a need for empirical research to resolve such questions as the joint role of technology and language in defining the construct to be measured, the effect of nonverbal input as part of the listening construct, and the effect of multimodal input on authenticity. We called for the development of authoring systems that allow the implementation of innovative test tasks specifically for language tests. We suggested that more research is needed on the automated scoring of speaking and writing using natural language processing (NLP) approaches to capture the complexities of natural language use, as well as on the acceptance of such scoring systems by test users. Finally, we join Chapelle and Douglas (2006) in calling for much more research on the impact of CALTs on learning, teaching, and policy.

ANNOTATED REFERENCES

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

This book has as its main aim the starting of a debate about how diagnostic language tests might be developed. The author pointed out that although the term “diagnostic test” is not unusual in the fields of language teaching and assessment, the term has not been clearly or completely defined and there are few examples of truly diagnostic language tests available. He points to a paradox in the field that, on the one hand, we believe in the concept of “diagnosis in order to improve learning,” but on the other hand, there is little research describing in any detail how learners actually develop language proficiency and consequently no tests that explicitly claim to be

diagnostic. The book provides detailed discussions of what we know about second language acquisition in speaking, listening, reading, writing, and grammatical development, exemplifying the points with reference to the monumental DIALANG project, computer-based diagnostic tests in 14 European languages. This is thus at once a theoretically based discussion of the relationship between second language acquisition and language test development and an empirically based exemplification of how the theory might be operationalized in CALTs.

Carr, N. T. (2006). Computer-based testing: Prospects for innovative assessment. In L. Ducate & N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching* (CALICO Monograph Series, Vol. 5, pp. 289–312). San Marcos, TX: CALICO.

Carr discussed the current situation of computer-based testing (CBT), including computer-adaptive tests (CAT) and Web-based testing (WBT). Carr argued that CBTs have not revolutionized testing, but are in fact very similar to the traditional paper-and-pencil tests but delivered via computer. He noted that a CBT can undermine the validity of a test if the technology distorts the way the test taker interacts with the language tasks presented. Carr also pointed out that development of automated scoring and partial scoring raises debates concerning assumptions about constructs being measured. He used the Bachman & Palmer (1996) framework of test usefulness to frame his discussion of test development and suggested that computer-delivered tests may enhance or detract from any of the qualities in the framework "a bad test delivered by a wonderful computer system is still a bad test."

Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

This book introduces computer-assisted second language testing as a subfield of language testing with unique potentials and constraints. It is intended for readers in language teaching and other areas of applied linguistics who are concerned with developing, evaluating, or using computer-assisted language tests (CALTs) for educational or research purposes. Readers in these areas are increasingly exposed to technology in their work and confronted with the need to understand how and why technology is being used to construct and deliver language tests. The book aims to address these needs by explaining issues associated with CALT, including its promise, concerns about validity, and methods of evaluating and implementing CALT, as well as the impact of CALT on the profession. The book introduces readers to *what* CALT is, *why* it is of interest from a practical perspective of language testing, and *how* it is implemented. The "what" and "why it is of interest" are covered in the first two chapters, with chapter 1 providing some examples as they relate to the interests of three types of potential readers—teachers, test developers, and researchers—and chapter 2 offering more examples, centered around the differences CALT makes in test methods. The third chapter offers the other side of the "why," considering why test users are concerned about the use of CALT. The fourth chapter illustrates how CALT can be implemented using currently available tools, while the fifth provides a discussion of how CALT can be evaluated. The final chapter deals with the impact of CALT and discusses possible future developments.

Stoynoff, S., & Chapelle, C. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.

Primarily a collection of reviews of ESOL (English for speakers of other languages) tests, including a number of CALTs, written by leading researchers in the field of language testing, this book also provides guidelines for understanding and using language tests and test results. The editors situate the various reviews by discussing the nature of language tests and test development, how to use test manuals to learn more about tests, how to evaluate language tests using the Bachman and Palmer (1996) framework, and how to develop teacher-made tests when no other options are available. The result is the first comprehensive review of language tests and language testing since that edited by Alderson, Krahnke, and Stansfield in 1987. Computer-based tests reviewed in the volume include the ACT ESL Placement Test, the Business Language Testing Service (BULATS), the ESL Computerized Adaptive Placement Exam (ESL-CAPE), and the Test of English as a Foreign Language (TOEFL) computer-based test.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 1–24.

The authors reviewed the development of several automated writing evaluation systems since the 1960s and highlighted different approaches that have been taken by three main U.S.-based companies. Warschauer and Ware pointed out that recent research in this area has focused on English and been conducted in the realm of psychometric research, which, in turn, has often been sponsored by developers of automated writing evaluation systems and should thus be interpreted with caution. The authors argued for process research in actual classrooms where AWE systems are in use. Thus, harnessing the archival potential of such systems, researchers could complement end-product-oriented research and shed light on how students use the system. Consequently, researchers would be well positioned to avoid the black box phenomenon and to document writing development.

OTHER REFERENCES

ACT. (2007a). *ACT ESL Placement Test*. Retrieved June 4, 2007, from http://www.act.org/esl/index.html

ACT. (2007b). *ACT Listening Proficiency Descriptors*. Retrieved June 4, 2007, from http://www.act.org/esl/desc/listen.html

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the common European framework. *Language Testing, 22*(3), 301–320.

Alderson, J. C., Krahnke, K., & Stansfield, C. (Eds.) (1987). *Reviews of English language proficiency tests*. Washington, D.C.: TESOL Publications.

Alderson, J. C., Percsich, R., & Szabo, G. (2000). Sequencing as an item type. *Language Testing, 17*(4), 423–447.

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

Brandl, K. (2005). Are you ready to "MOODLE"? *Language Learning & Technology, 9*(2), 16–23.

Breland, H., Lee, Y., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts: Response mode analyses* (TOEFL Research Report No. RR-75). Princeton, NJ: ETS.

Burgoon, J. (1994). Non-verbal signals. In M. Knapp & G. Miller (Eds.), *Handbook of interpersonal communication* (pp. 344–393). London: Routledge.

Burstein, J., & Chodorow, M. (2002). Directions in automated essay analysis. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 487–497). New York: Oxford University Press.

Cambridge, ESOL. (2007). *Business language testing service*. Retrieved June 15, 2007, from http://www.bulats.org/

Canale, M. (1986). The promise and threat of computerized adaptive assessment of reading comprehension. In C. Stansfield (Ed.), *Technology and language testing* (pp. 30–45). Washington, DC*:* TESOL.

Carr, N. T. (2006). Computer-based testing: Prospects for innovative assessment. In L. Ducate & N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching* (CALICO Monograph Series Vol. 5, pp. 289–312). San Marcos, TX: CALICO.

Carrell, P. (2007). *Notetaking strategies and their relationship to performance on listening comprehension and communicative assessment tasks* (ETS Research Report No. RR-07-01, TOEFL Monograph No. MS-35). Princeton, NJ: ETS.

Chapelle, C., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chapelle, C. A. (2006). Review of DIALANG. *Language Testing, 23*(4), 544–550.

Chodorow, H., & Burstein, J. (2004). *Beyond essay length: Evaluating E-rater®'s performance on TOEFL essays* (TOEFL Research Report No. RR-73). Princeton, NJ: ETS.

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295–320.

Chun, C. W. (2006). Commentary: An analysis of a language test for employment: The authenticity of the PhonePass Test. *Language Assessment Quarterly: An International Journal, 3*(3), 295–306.

Cohen, Y., Ben-Simon, A., & Hovav, M. (2003, October). *The effect of specific language features on the complexity of systems for automated essay scoring*.

Paper presented at the International Association of Educational Assessment Annual Conference, Manchester, England.

Cohen, A. D., & Upton, A. T. (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph Series MS-33). Princeton, NJ: ETS.

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL, 18*, 193–211.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U. S., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL* (TOEFL Monograph Series MS-30). Princeton, NJ: ETS.

DIALANG. (2003). Retrieved June 15, 2007, from http://www.dialang.org

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1). Retrieved January 30, 2007, from http://www.jtla.org

Douglas, D., & Hegelheimer, V. (2006, July). *Strategies and use of knowledge in performing new TOEFL listening tasks: The main study*. Paper presented at Language Testing Research Colloquium, Melbourne, Australia.

Fulcher, G. (2005). Better communications test will silence critics. *Guardian Weekly*, Friday, February 18, 2005. http://education.guardian.co.uk/tefl/story/0,5500,1645011,00.html

Green, T. (2006, April 3). Comment on LTEST-L. *LTEST-L Archives*. Retrieved from http://lists.psu.edu/archives/ltest-l.html

Gruba, P. (1997). Exploring digital video material. In R. Debski, J. Gassin, & M. Smith (Eds.), *Language learning through social computing* (pp. 109–140). Parkville, Victoria, Australia: Applied Linguistics Association of Australia.

Harcourt Assessment. (2006). *Versant*. Retrieved June 5, 2007, from http://harcourtassessment.ca/haiweb/Cultures/en-US/dotCom/Versant/

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering, 12*(2), 145–159.

Ishioka, T., & Kameda, M. (2004). Automated Japanese essay scoring system: JESS. *Proceedings of the 15th International Workshop on Database and Expert Systems Applications, Zaragosa, Spain*, 4–8.

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics, 25*, 228–242.

LTEST-L Archive. (2006a, April–May). *Whatever happened to adaptive tests?* Retrieved June 14, 2007, from http://lists.psu.edu/archives/ltest-l.html

LTEST-L Archive. (2006b, November). *Questionmark Perception*. Retrieved June 14, 2007, from http://lists.psu.edu/archives/ltest-l.html

Monaghan, W., & Bridgeman, B. (2005, April). *E-rater* as a quality control on human scores. *ETS RD Connections.* Retrieved January 30, 2007, from http://www.ets.org/Media/Research/pdf/RD_Connections2.pdf

Ordinate. (2006). Retrieved June 15, 2007, from http://www.ordinate.com/

Pearson Knowledge Technologies. (2006). *Reliability and validity of the KAT engine*. Retrieved January 30, 2007, from http://www.pearsonkt.com/researchVRSum.shtml

Roever, C. (2005). Review of the ESL Computer Adaptive Placement Exam (ESL-CAPE). In S. Stoynoff & C. Chapelle, *ESOL tests and testing* (pp. 66–68). Alexandria, VA: TESOL.

Rudner, L. M., Garcia, V., & Welch, C. (2005). *An evaluation of IntelliMetric essay scoring system using responses to GMAT® AWA prompts* (GMAC Research Reports RR-05-08). McLean, VA: Graduate Management Admission Council.

Sim, G., Horton, M., & Strong, S. (2004, April). *Interfaces for online assessment: Friend or foe?* Paper presented at The 7th HCI Educators Workshop: Effective Teaching and Training in HCI, University of Central Lancashire, Preston, England. Retrieved February 4, 2007, from http://www.ics.heacademy.ac.uk/events/presentations/382_grsim_interface.pdf

SoftStudy. (2006). *Web-Computer-Adaptive Placement Examination (Web-CAPE).* Retrieved February 2, 2007, from http://www.softstudy.com//products/CAPE.cfm

Stoynoff, S., & Chapelle, C. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219–274.

Taylor, C., Jamieson, J., & Eignor, D. (2000). Trends in computer use among international students. *TESOL Quarterly, 34*(3), 575–585.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology, 11*(1), 67–86.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 1–24.

Winke, P., & MacGregor, D. (2001). Review of hot potatoes. *Language Learning & Technology, 5*(2), 28–33. Retrieved January 24, 2007, from http://llt.msu.edu/vol5num2/review3/default.html

Wolfe, E. W., & Manalo, J. R. (2005). *An investigation of the impact of composition medium on the quality of TOEFL writing scores* (TOEFL Research Report No. RR-72). Princeton, NJ: ETS.

Zechner, K., Bejar, I., & Hemat, R. (2007). *Toward an understanding of the role of speech recognition in non-native speech assessment* (ETS Research Report No. RR-07-02, TOEFL iBT Report No. TOEFLiBT-02). Princeton, NJ: ETS.