Annual Review of Applied Linguistics (2005) **25**, 228–242. Printed in the USA.
Copyright © 2005 Cambridge University Press 0267-1905/05 $12.00

12. TRENDS IN COMPUTER-BASED SECOND LANGUAGE ASSESSMENT

Joan Jamieson

In the last 20 years, several authors have described the possible changes that computers may effect in language testing. Since *ARAL*'s last review of general language testing trends (Clapham, 2000), authors in the *Cambridge Language Assessment Series* have offered various visions of how computer technology could alter the testing of second language skills. This chapter reflects these perspectives as it charts the paths recently taken in the field. Initial steps were made in the conversion of existing item types and constructs already known from paper-and-pencil testing into formats suitable for computer delivery. This conversion was closely followed by the introduction of computer-adaptive tests, which aim to make more, and perhaps, better, use of computer capabilities to tailor tests more closely to individual abilities and interests. Movement toward greater use of computers in assessment has been coupled with an assumption that computer-based tests should be better than their traditional predecessors, and some related steps have been taken. Corpus linguistics has provided tools to create more authentic assessments; the quest for authenticity has also motivated inclusion of more complex tasks and constructs. Both these innovations have begun to be incorporated into computer-based language tests. Natural language processing has also provided some tools for computerized scoring of essays, particularly relevant in large-scale language testing programs. Although computer use has not revolutionized all aspects of language testing, recent efforts have produced some of the research, technological advances, and improved pedagogical understanding needed to support progress.

Since 2000, six new books describing the assessment of reading, listening, language for special purposes, speaking, vocabulary, and writing have been published in the *Cambridge Language Assessment Series* edited by J. Charles Alderson and Lyle Bachman (Alderson, 2000; Buck, 2001; Douglas, 2000; Loumi, 2004; Read, 2000; Weigle, 2002). In the last chapter of every book (except for *Assessing Speaking*), the authors addressed the potential and the challenges of using the computer for assessment, following in the footsteps of Canale (1986) by alerting us to limitations while charting directions for the future. Each author mentioned some of the efficiency arguments that have, for many years now, been sugeted for computer-based assessment such as less time needed for testing, faster score

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

reporting, and provision of more convenient times for test-takers to take the test. Authors also mentioned the arguments that have been put forward for computer-adaptive assessment, such as more precise information about each individual's level of ability, which is thought to result in more challenging and more motivating experiences for test takers than traditional pencil-and-paper tests (e.g., Chalhoub-Deville & Deville, 1999; Madsen, 1986). The authors cautioned us that computer-based assessment is expensive, and that there is a danger in letting advances in both technology and in statistical procedures in psychometrics lead us down misguided paths of development. Most importantly, though, these experts articulated their visions of how computers may redirect second language assessment by providing us means to enhance the authenticity of our tests and to clarify our understanding of language abilities. In this chapter, I survey the field of computerized second language assessment mainly in terms of English as a second language (ESL), reflecting the directions of these commentators, looking at where we are, and, like the others, trying to foresee what lies ahead. For current reviews of widely used ESL tests, see Stoynoff and Chapelle (2005).

First Steps: Using Traditional Item Types and Constructs in Computerized Formats

The ease with which multiple choice or fill-in-the-blank items testing vocabulary, grammar, and reading comprehension can be adapted to a computerized environment is evident in the number of language tests on the Web. Glenn Fulcher maintains a very useful web site with links, as well as brief reviews, to many second language tests. Dave's ESL Cafe has a Quiz Center that provides students with a number of short tests which are immediately scored; these quizzes illustrate typical "low-tech" online assessments in which a hodgepodge of topics are grouped together in one site where a learner can answer a few questions, click on a button, and receive a percentage correct score. Somewhat more sophisticated examples offer quick checks of a learner's English level. There are many examples currently on the web; for the purpose of illustration, only a few examples are included here. ForumEducation.net offers learners two multiple choice vocabulary tests to estimate word knowledge as an indication of English proficiency. Wordskills.com offers visitors to its site three levels of tests (with 25 items each) which are said to approximate (disclaimer provided) the Cambridge First Certificate, Certificate in Advanced English, and the Certificate of Proficiency in English; the site has other tests including assessment of negotiation skills and e-commerce. Also, Churchill House offers online tests for students to check their level to prepare for the most appropriate test in the University of Cambridge Local Examination Syndicate (UCLES) suite of exams. All of the items are in multiple choice format. Netlanguages.com offers learners a two-part test to determine their level and appropriate placement in online courses. Learners initially select what they think their level of English is. The first section then tests grammar by having the test taker drag and drop a word into a sentence; if the test-taker's score is low after the first 10 items, for example, a message indicates that another level may be more appropriate. The second section consists of choosing a few questions from a larger set and then writing two or three sentences in response to them and submitting the responses to

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

get scored by a human rater. Many of these tests are aligned with online learning programs. At study.com learners are provided with EFI (English for Internet) placements tests in listening, speaking, and writing in addition to vocabulary, reading, and grammar in preparation for free, online instruction.

In 2001, Irene Thompson edited a special issue of *Language Learning & Technology*, which was devoted to the topic of the theory and practice of computer-based language testing (Thompson, 2001). Several of the columns and reviews pointed out web resources (Leloup & Ponterio, 2001) and tools for web test development (Godwin-Jones, 2001; Polio, 2001; Winke & MacGregor, 2001). In his article on web-based testing in the same issue, Roever (2001) stated that these types of tests are most suitable for low-stakes decisions when the tests are used for checks on learning, test preparation, or research in which learners have little or no reason to cheat, and in situations in which test takers can take the test whenever they want, in the privacy of their own homes, and at their own pace.

In some cases, rather than creating a new set of items, a paper-and-pencil version of a test has been directly converted to a computerized format, which may be on the web or not. As in the examples given previously, language ability is often characterized in terms of performance on subtests such as vocabulary, grammar, reading comprehension, and/or listening comprehension; in other words, constructs that have been used in earlier tests are adapted to a new medium. In the case of the Test of English Proficiency at Seoul National University, new technology was used to measure English skills with accessible administration and fast score reporting (Choi, Sung Kim, & Boo, 2003). In the case of the Test of English as a Foreign Language (TOEFL) computer-based test, this was done for a research project to investigate the potential threat to validity caused by lack of computer familiarity (Taylor, Kirsch, Jamieson, & Eignor, 1999; also see ETS, 2000). Differences that may affect performance because of the ways computer-based tests and paper-and-pencil tests present tasks to test-takers are discussed briefly by both Alderson (2000) and Douglas (2000). In her comparison of paper-and-pencil and computerized reading tests, Sawaki (2001) provides us with a thorough review of many factors that need to be considered when the mode of delivery is changed.

Second Steps: "Added Value" to Computerized Assessments

Computer-Adaptive Tests

Because the cost of developing computer-based tests is much higher than that of traditional paper-and-pencil tests, Buck (2001) wrote that "there need to be significant advantages to make that worthwhile" (p. 255). In many language testing programs, a "value-added" approach to computerization of the paper-and-pencil format was the introduction of computer-adaptive sections. An added benefit of a computer-adaptive test is that theoretically test-takers are given items that are well-suited to their abilities, thus making the test both more interesting and a more accurate measure.

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

As described in Wainer et al. (2000) and Eignor (1999), information such as the content and difficulty of an item needs to be determined before it can be administered in a computer-adaptive test. Item function is characterized by item parameters—item difficulty, item discrimination, and guessing—and the measurement model used to estimate these parameters is Item Response Theory (IRT). Some computer-adaptive tests use only the first parameter, item difficulty; other tests use all three parameters. To get stable item parameters, items are typically pretested on what is considered a representative sample of test takers. Pretesting is a necessary, but expensive, condition for a computer-adaptive test. For example, the ACT ESL test used over 2,000 test-takers to estimate its item parameters (ACT, Inc., 2000).

The basic idea of computer adaptive tests is that a test taker will respond to an item with known item parameters. That response will be automatically scored by the computer so that a type of profile of the test taker can be initiated. As the test taker proceeds, his or her "profile" is continually updated, based on the correctness of response to items that are being selected as representative of a certain range of parameters. The necessity of automatic scoring for subsequent item selection was a relatively easy condition to meet for language tests using the traditional constructs of vocabulary, grammar, reading comprehension, and listening comprehension that were originally presented in a paper-and-pencil format with multiple choice items.

There are currently a number of computer-adaptive second language tests that reflect this trend. Examples range from the English Comprehension Level Test administered at the Defense Language Institute, the ACT ESL Placement Test (COMPASS/ESL), the Business Language Testing Service (BULATS), and the Structure and Written Expression section and the Listening Section of the computer-based TOEFL; the reading section was not computer-adaptive so that test takers could skip items and return to them later (ETS, 2000). Readers interested in issues related to computer-adaptive testing of reading are directed to the edited volume by Chalhoub-Deville (1999).

Because computer-adaptive tests for the most part rely on item response theory, a necessary assumption is that items must be conditionally independent. When several items are based on the same passage, in a reading comprehension test, for example, this situation can be handled by treating all of the items in a set as one "large" item, which is polytomously scored. In other words, instead of having six reading comprehension items each scored 0–1, there would just be one item, scored 0–6. Work reported in the mid-1990s found that polytomous IRT models could be used to combine scores from traditional, dichotomously scored items (correct/incorrect or 1/0) with these polytomously scored items (Tang, 1996; Tang & Eignor, 1997). However, the assumption cannot be met when a reading passage is not only used for reading comprehension items, but also as the content for a writing item.

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

"Adaptive" Tests in Different Guises

The idea of adapting test delivery and content to more closely match the test-taker's interests and/or ability level has been appealing to language test developers, whose tests, for one reason or another, do not meet the stringent requirements of traditional computer-adaptive tests. Consequently, there are a number of computerized language tests currently available or in the research and development stages that make adaptations from a format in which all test takers are given the same items in the same order. Two techniques which allow for the tailoring of tests to the individual are self-assessment and short screening tests.

There are cases in which computer-adaptive tests are not currently possible, such as speaking tests in which oral discourse is scored by human raters. Here, the condition of automatic scoring one item, or task, to aid in the selection of subsequent items cannot be met. Still, test developers are interested in incorporating some flexibility, or "adaptivity," into the selection of tasks. The COPI (Computerized Oral Proficiency Interview) is the latest development in assessing speaking from the Center for Applied Linguistics, following the SOPI (Simulated Oral Proficiency Interview—tape recorded) and the OPI (oral proficiency interview—humans). In the original COPI project, existing material from SOPIs formed a test bank. Test takers were asked to assess their speaking level, ranging from intermediate to superior. Seven speaking tasks were then generated from the pool—four at the level of self-assessment and three at the next level above that (Kenyon, Malabonga, & Carpenter, 2001). Contrasting test-takers' attitudes between the SOPI and the COPI, Kenyon and Malabonga (2001) reported that test-takers felt that the COPI was less difficult than the SOPI and that "the adaptive nature of the COPI allowed the difficulty level of the assessment task to be matched more appropriately to the proficiency level of the examinee" (p. 60) although, as Norris (2001) pointed out, this claim needs further investigation. The Center for Applied Linguistics web site indicates that efforts are underway to create operational versions of the COPI in Arabic and Spanish which should be available on CD-ROM in 2006.

In other cases, items can be automatically scored, but the prerequisite pilot testing of items needed to estimate item parameters is not feasible. Still, test developers are interested in "adapting" the selection of items for each test taker. Two techniques have been used recently. One technique has test takers answer survey questions to select content of interest. The other has test takers respond to items in a screening test to estimate their ability level so that subsequent test sections can be selected at an appropriate level of difficulty. *Longman English Assessment* (Jamieson & Chapelle, 2002; Chapelle, Jamieson, & Hegelheimer, 2003) illustrates both of these techniques. This low-stakes test was designed to provide English language learners with an interesting experience that would give them feedback about their level of proficiency and recommendations for improving their English. This test begins with a fifteen minute "interest and ability finder." First, test-takers answer survey questions about why they want to learn English; responses are used to select business or more general content. Second, all test-takers are then administered a set of vocabulary and grammar questions; responses are used to recommend a level

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

(i.e., beginning, intermediate, advanced) for the subsequent sections of the test. Because of the inability to pretest items in order to determine their difficulty empirically, the test developers relied on theory to create items and sections at different ability levels. For example, the vocabulary items were based on word frequency counts, the grammar items on developmental sequences, the written structures on textbook analysis, and the reading and listening comprehension items on a set of processing conditions (Rupp, Garcia, & Jamieson, 2002). The text passages were analyzed using corpus linguistic techniques together with the judgments of experienced teachers.

A final example in this section is DIALANG, which incorporates both self-assessment and a screening test to create an "adaptive" test for individual language learners. Like *Longman English Assessment*, DIALANG is a low-stakes test with which individual language learners can find out about their ability levels in vocabulary, grammar, writing, reading, and listening. It is currently available in 14 languages: Danish, Dutch, English, Finnish, French, German, Greek, Icelandic, Irish, Italian, Norwegian, Portuguese, Spanish and Swedish. The theoretical foundation for DIALANG is the Common European Framework. Test takers first select the language and the skill in which they want to be tested. Then, DIALANG begins with a vocabulary test which is used for initial screening; this is followed by a self-assessment section. These are followed by a test in whatever language skill was chosen, which is modified for the test taker based on information from the screening test and self-assessment. A pilot version can be downloaded from the DIALANG website.

In this section, I have illustrated language tests that are different from traditional tests mainly because they are administered on computer rather then with paper-and-pencil or with an interviewer, and because many of them make use of computer technology to branch test takers to different subsets of items or tasks. These tests can be adaptive or linear, and they can be administered via the web, CD, or network. Each method of delivery offers potential benefits and problems, as summarized by Brown (2004). However, many of these tests do not provide us with an alternate construct of language ability. So, although the tests are innovative in terms of technology, they are not particularly innovative in their operationalization of communicative language ability.

Next Steps: Innovation in How Language is Perceived

As Douglas wrote, "language testing that is driven by technology, rather than technology being employed in the service of language testing, is likely to lead us down a road best not traveled" (2000, p. 275). Moving beyond technical innovation and into the realm of how we perceive and measure language ability illustrates progress in our ability to make use of computers. In this section, I highlight innovations involved with language tasks and score reporting.

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

Language tasks

Considering both the nature of the input and the nature of the expected response as elements in a task, we can find examples of computerized tests that are trying to more closely replicate the target language use domain (TLU; Bachman & Palmer, 1996), addressing the test quality of authenticity and also enriching our understanding of language use in different situations. Read (2000) discussed how a broader social view of vocabulary in testing would take into account the concept of register, addressing questions such as what sort of vocabulary particular tasks require. Douglas (2000) pointed to the potential role of corpus linguistics investigating TLU contexts to increase our understanding of how grammar, vocabulary, and rhetorical structure are used in different contexts.

Informed by research in corpus linguistics, the development of the new TOEFL provides us with an example of how ideas such as these have been put into practice. Biber and his colleagues (2002) created a corpus of spoken and written academic language that was subsequently analyzed for lexical and syntactic characteristics. This corpus then served as a type of baseline for language used in different registers and genres. Spoken texts, for example, included lectures with different amounts of interaction, service encounters, and classroom management. Listening passages were developed following real-life examples. Biber et al. (2004) created tools which allowed the features of the created passages to be compared to the features of the texts analyzed in the corpus. Instead of incorporating this interactionist construct perspective, a trait perspective was used by Lanfer and her colleagues (2004a, 2004b) in their innovative vocabulary test of size and four degrees of strength of knowledge; the test was piloted in a computer-adaptive format.

Another innovation of the new TOEFL is that it includes tasks that resemble those performed by students in North American colleges and universities (Rosenfeld, Leung, & Oltman, 2001). These "integrated skills" tasks require test takers to use the information provided in the reading and listening passages in essay and/or spoken responses. The same passages also serve as the input for reading and listening comprehension questions. This use of more authentic tasks provides evidence for the representation inference in the validity argument for the new TOEFL, and it has also forced language testers to grapple with the traditional, yet conflicting, inferences of performance on language tests between underlying ability and task completion (Chapelle, Enright, & Jamieson, forthcoming). Moreover, this decision to include integrated tasks resulted in the violation of the assumption of IRT that tasks on a single scale be conditionally independent, and without IRT, the necessary psychometric base for calibrating items for computer-adaptive delivery was lost. The current need for human raters of speaking and writing tasks also precluded the use of computer-adaptive delivery (Jamieson, Eignor, Grabe, & Kunnan, forthcoming). The decision to develop computerized tasks that better represent authentic language use rather than to be constrained in task development by relying on known technology and psychometrics marks a new direction in large scale, high-stakes testing.

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

Another task factor that has been discussed in the literature is response type; associated with this is factor is whether or not the response can be scored by a computer. Up to this point, most of the examples discussed so far use multiple choice responses, which may appear as familiar types where test takers click on a circle before one of four choices, or as more novel drag and drop item types. These items share the essential quality of having one correct answer and being dichotomously scored as either correct or incorrect. More innovative item types require the test taker to identify several important points from a reading or listening passage, or to select sentences from a pool to form a paragraph. These items are often scored polytymously—on a scale from 0 to 3, for example. However, to date the rationales for associating the partial scores for each item have more to do with controlling for guessing (Enright, Bridgeman, Eignor, Lee, & Powers, forthcoming) than they have to do with the construct definition, as Chapelle illustrated in an example of decisions for partial correction in a c-test (2001). This is because the computer programs used to analyze responses in most current operational language tests are not based on analysis of linguistic output, but rather on matching the pattern of the correct response to the response given by the test-taker.

A test which stands in sharp contrast to this situation is the Spoken English Test, formerly known as PhonePass. SET-10 is administered over a telephone; it uses a speech recognition computer program to analyze and to score the phonological representation of responses to each item. "The system generates scores based on the exact words used in the spoken responses, as well as the pace, fluency, and pronunciation of the words. . . Base measures are derived from the linguistic units . . . based on statistical models of native speakers" (See ordinate website listed at the end of this chapter; Ordinate, 1999).

Score Reports

The desire to communicate meaningful interpretations of test scores is not limited only to computer-based assessment. However, recent language assessments delivered with computers have incorporated some innovations in the types of information about test-takers given to test score users.

"The possibility of recording response latencies and time on test or task opens up a whole new world of exploration of rates of reading or word recognition" (Alderson, 2000, p. 351). Although I think that Alderson was addressing a fundamental change in construct definition, the closest example of a large scale, standardized test which makes use of test takers' response times is ACT's COMPASS/ESL. ACT began development of this test in 1993, in consultation with ESL professionals. Its purpose is mainly placement of students in community colleges in the United States. Having chosen to deliver the test in an adaptive format, "students can proceed at their own pace on the computer, yet have their assessment time monitored automatically, allowing institutions to use the amount of time the student spends taking the test as an additional factor in placement decisions" (ACT, Inc., 2000, p. 132).

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

Each of the three tests in the ACT battery (grammar/usage, reading, and listening) includes a five-level score scale. The levels are mapped to the tests' specifications and proficiency descriptors are provided for each level. These proficiency descriptors were created by not only examining what was measured internally on the test, but by also examining external sources such as learning outcomes tied to college-level ESL curricula and well known benchmarks such as the ACTFL Guidelines. An example from the listening scale follows:

> Proficiency Level 2 (67-81): Students at Level 2 typically have the ability to understand brief questions and answers relating to personal information, the immediate setting, or predictable aspects of everyday need. They understand short conversations supported by context but usually require careful or slowed speech, repetitions, or rephrasing. Their comprehension of main ideas and details is still incomplete. They can distinguish time forms, some question forms (Wh-, yes/no, tag questions), most common word-order patterns, and most simple contractions, but the students may have difficulty with tense shifts and more complex sentence structures. (ACT, 2000, p. 45)

Proficiency scaling approaches assume that the abilities used to explain test scores are defined by task characteristics, which are hypothesized to be responsible for task difficulty. IRT, used as the psychometric model in the COMPASS/ESL test, allows items and person ability to be placed on the same scale. Similar proficiency scales were considered for test score reporting in the new TOEFL (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000). However, because IRT was not going to be used, other types of performance descriptors, for example, teacher assessments of what a person at a certain score level could probably be expected to do, were considered to accompany score reports (Powers, Roever, Huff, & Trapani, 2003).

Another way to make score reports meaningful to the test taker is to provide information on how to improve one's language ability. This type of feedback is provided by both DIALANG and *Longman English Assessment*. In the latter, the computer program generates a list of study strategies based on test takers responses to where they are studying English (in an English-speaking country or not), together with their level scores.

Future Steps: Computer Scoring of Constructed Responses, New Item Types

The authors in the *Cambridge Language Assessment Series* look forward to corpus-based vocabulary lists (Read, 2000) and corpora of field specific discourse (Douglas, 2001), realistic simulations, perhaps through virtual reality (Buck, 2001), and the capture of every detail of a learner's progress through a test (Alderson, 2000). All of these are possible, of course, yet it is the scoring of writing by computer (Weigle, 2002) that I think is the most probable advancement in technology that will occur in large scale language assessment in the near future.

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

Three computer scoring systems are clearly described by Weigle: Latent Semantic Analysis, Project Essay Grade (Tru-Judge, Inc.), and E-rater (Educational Testing Service, ETS). Another automatic scoring system is the IntelliMetric Essay Scoring Engine (Vantage Learning). Intelligent Essay Assessor uses Latent Semantic Analysis (formerly owned by Knowledge Assessment Technologies, its acquisition by Pearson Education was announced on June 29, 2004). According to Weigle (2002), Latent Semantic Analysis has the practical disadvantage of not taking word order into account. As described by Chodorow and Burstein (2004, p. 1), each system trains on essays that have been read by human raters who have assigned a holistic score. From the training examples, the system learns which features are the best predictors of the essay score. "The differences among the systems lie in the particular features that they extract and in the ways in which they combine the features in generating a score" (Chodorow & Burstein, 2004, p. 1). E-rater's particular features are based on four kinds of analysis: syntactic, discourse, topical, and lexical. In version 2.0, because essay length was found to be "the single most important objectively calculated variable in predicting human holistic scores" essay length as measured by number of words was included in e-rater's updated feature set (Attali & Burstein, 2004, p. 4). Although Weigle stated that this trend is controversial among writing teachers, these automated systems have none the less reported high correlations with human raters (Burstein & Chodorow, 1999).

COMPASS/ESL offers a writing section called COMPASS e-Write that is scored using IntelliMetric. An overall score ranging in value from 2–8 is given to each essay, along with feedback on content, style, organization, focus, and mechanics. This technology is available on the web; the instructional part of the software is available at MyAccess! Although not clear from the COMPASS/ESL website, MyAccess! is available for ESL students, and has been developed to score essays written by nonnative speakers of English. E-rater has been used by ETS for the Graduate Management Admissions Test Analytic Writing Assessment since 1999. Criterion Online Essay Evaluation is the instructional software using e-rater from ETS (Burstein, Chodorow, & Leacock, 2003). ETS currently uses automated scoring of essays, e-rater, in the Next Generation TOEFL Practice Test, as well as ScoreItNow. Research is underway to evaluate the feasibility of using e-rater operationally as second rater for new TOEFL independent writing samples. At present there are no plans for automated scoring of speech for new TOEFL (personal communication, M. K. Enright, October 27, 2004).

As research progresses, it will be interesting to watch for advances in computer scoring of constructed, or free, responses. Another ETS program, C-rater, has been developed to measure a student's understanding of specific content without regard for writing skills (Leacock, 2004). Also, the *Spoken English Test* has included a free response section on its Set-10 Form for many years. Information from this section has not been included in score reports, but rather has been included to collect data for research on automated scoring of spoken free responses. It was announced on September 29, 2004, that the company that developed *PhonePass,* Ordinate, was acquired by the large testing company, Harcourt Assessment, Inc.

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

Finally, to return to the topic of task types on language assessments, nothing has yet been mentioned about the use of video. Certainly this is one area in test development that seems natural to include in the creation of authentic language input. Multimedia computer-assisted language learning software such as *Longman English Interactive* (Rost & Fuchs, 2003) includes video in its quizzes and tests, but most stand-alone computer-based language tests do not. Why? It is expensive to develop, it requires high-end technology to transmit, and it is unclear how it affects the construct. Douglas (2000) commented that language testers have so far not made much progress in adapting multimedia to the goals of measurement; he cautions us that, although it is tempting to make use of video to provide authentic content and contextualization cues, we need to understand and control for test method effects associated with these technological resources. With this caution in mind, it is still intriguing to imagine new types of tasks to broaden our perspective of what is possible for language tests. (A good place to begin is at DIALANG's web site; click on "New item types.")

Summary

As we read in Burstein, Frase, Ginther, and Grant (1996), the role of computers in assessment is broad because computers have a number of functions essential for an operational testing program such as the TOEFL, including item creation and presentation; response collection and scoring; statistical analysis; and storage, transmission, and retrieval of information. This review has not gone into depth on many of these issues, but has rather highlighted a few recent trends mainly surveying the literature in which authors predict that technology will play a role in language assessment in the future. We can also see that the future envisioned by many of these commentators is the distant future. While discussing the advancement of applied linguistics through assessment, Chapelle (2004) referred to "the tunnel of efficiency" and "the panorama of theory." 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.

Web sites (Addresses functional as of December 8, 2004)

BULATS (http://www.bulats.org)

Churchill House (http://www.churchillhouse.com/english/exams.html)

COMPASS/ESL (http://www.act.org/compass/ or http://www.act.org/esl)

COMPASS e-Write (http://www.act.org/e-write/index.html)

COPI (http://www.cal.org/projects/copi/)

Dave Sperling's ESL Cafe Quiz Center (http://www.pacificnet.net/~sperling/quiz/)

DIALANG (http://www.dialang.org/)

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

DIALANG's experimental items http://www.lancs.ac.uk/fss/projects/linguistics/experimental/new/start.htm)

Educational Testing Service, Criterion, e-rater, and c-rater (http://www.ets.org/research/erater.html)

Educational Testing Service, Next Generation TOEFL (http://www.ets.org/toefl/nextgen/)

Elite Skills Ltd. (http://wordskills.com/level)

English Comprehension Level Test (http://www.dlielc.org/testing/ecl_test.html)

Forum Education (http://www.forumeducation.net/servlet/pages/vil/mat/index.htm)

Glenn Fulcher's web site (http://www.dundee.ac.uk/languagestudies/ltest/ltr.html)

Intelligent Essay Assessor (http://www.knowledge-technologies.com/)

Language projects supported by European Union (http://europa.eu.int/comm/education/programmes/socrates/lingua/index_en.html)

MyAccess! (http://www.vantagelearning.com/product_pages/myaccess.html)

Net Languages (http://www.netlanguages.com/home/courses/level_test.htm)

Ordinate, *Spoken English Tests* (http://www.ordinate.com/content/prod/prod-level1.shtml)

StudyCom English For Internet (http://study.com/tests.html)

Vantage Laboratories (http://www.vantage.com/)

REFERENCES

ACT, Inc. (2000). *COMPASS/ESL reference manual.* Iowa City, Iowa: Author.

Alderson, J. C. (2000). *Assessing reading.* New York: Cambridge University Press.

Attali, Y., & Burstein, J. (2004). Automated essay scoring with E-rater V.2.0. Available on-line at http://www.ets.org/research/erater.html

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

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36,* 9–48.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortez, V., Csomay, E., Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus.* TOEFL Monograph MS-25. Princeton, NJ: Educational Testing Service.

Brown, J. D. (2004). For computerized language tests, potential benefits outweigh problems. *Essential Teacher, 1, 4,* 37–40.

Buck, G. (2001). *Assessing listening.* New York: Cambridge University Press.

Burstein, J., & Chodorow, M. (1999, June). Automated essay scoring for nonnative English speakers. In *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*. College Park, MD. Available online at http://www.ets.org/research/dload/acl99rev.pdf

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion online essay evaluation: An application for automated evaluation of student essays. Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico. Available online at: http://www.ets.org/research/dload/iaai03bursteinj.pdf

Burstein, J., Frase, L., Ginther, A., & Grant, L. (1996). Technologies for language assessment. *Annual Review of Applied Linguistics, 16,* 240–260.

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 Publications.

Chalhoub-Deville, M. (Ed.) (1999). *Issues in computer adaptive testing of reading proficiency.* Studies in Language Testing, 10. New York: Cambridge University Press.

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

Chapelle, C. A. (2004). *English language learning and technology*. Philadelphia: John Benjamins.

Chapelle, C. A. (2001). *Computer applications in second language acquisition.* New York: Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (forthcoming). Challenges in developing a test of academic English. In C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.), *Building a validity argument for TOEFL.* Mahwah, NJ: Erlbaum.

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20,* 409–439.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays.* TOEFL Research Reports, Report 73. Princeton, NJ: Educational Testing Service.

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

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

Clapham, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics, 20*, 147–161.

Douglas, D. (2000). *Assessing languages for specific purposes.* New York: Cambridge University Press.

Educational Testing Service (ETS). (2000). *The computer-based TOEFL score user guide.* Princeton, NJ: Author.

Eignor, D. (1999). Selected technical issues in the creation of computer-adaptive tests of second language reading proficiency. In *Issues in computer adaptive testing of reading proficiency: Studies in Language Testing*, *10* (pp. 167–181). New York: Cambridge University Press.

Enright, M. K., Bridgeman, B., Eignor, D., Lee, & Powers, D. E. (forthcoming). Designing measures of listening, reading, writing, and speaking. In C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.), *Building a validity argument for TOEFL.* Mahwah, NJ: Erlbaum.

Godwin-Jones, R. (2001). Emerging technologies. *Language Learning & Technology, 5*(2), 8–12.

Jamieson, J., & Chapelle, C. A. (2002). *Longman English Assessment.* New York: Pearson Longman.

Jamieson, J., Eignor, D., Grabe, W., & Kunnan, A. (forthcoming). The frameworks for reconceptualization of TOEFL. In C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.), *Building a validity argument for TOEFL.* Mahwah, NJ: Erlbaum.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper.* TOEFL Monograph Series 16. Princeton, NJ: Educational Testing Service.

Kenyon, D., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology, 5*(2), 60–83.

Kenyon, D., Malabonga, V., & Carpenter, H. (2001). Response to the Norris commentary. *Language Learning & Technology, 5*(2), 106–108.

Lanfer, B., Elder, C., Hill, K., & Congdon, P. (2004a). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*, 202–226.

Lanfer, B., & Goldstein, Z. (2004b). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399–436.

Leacock, C. (2004). Scoring free-responses automatically: A case study of a large-scale assessment. Examens, 1, 3. English version available online at http://www.ets.org/research/erater.html

LeLoup, J., & Pontierio, R. (2001). On the net, language testing resources. *Language Learning & Technology, 5*(2), 4–7.

Loumi, S. (2004). *Assessing speaking.* New York: Cambridge University Press.

Madsen, H. (1986). Evaluating a computer-adaptive ESL placement test. *CALICO Journal, 4*(2), 41–50.

Norris, J. (2001). Concerns with computerized adaptive oral proficiency assessment. *Language Learning & Technology, 5*(2), 99–105.

Ordinate Corp. (1999). *PhonePass testing: Structure and content.* Ordinate Corporation Technical Report. Menlo Park, CA: Author.

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press

Polio, C. (2001). Review of *Test Pilot*. *Language Learning & Technology, 5*(2), 34–37.

Powers, D. E., Roever, C., Huff, K. L., & Trapani, C. S. (2003). *Validating LanguEdge courseware scores against faculty ratings and student self-assessments*. (ETS Research Report 03–11). Princeton, NJ: Educational Testing Service.

Read, J. (2000). *Assessing vocabulary*. New York: Cambridge University Press.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology, 5*(2), 84–94.

Rosenfeld, M., Leung, S., & Oltman, P. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels.* (TOEFL Monograph Series, MS-21). Princeton, NJ: Educational Testing Service.

Rost, M., & Fuchs, M. (2003). *Longman English Interactive*. New York: Pearson Education.

Rupp, A., Garcia, P., & Jamieson, J. (2002). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Language Testing, 1,* 185–216.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology, 5*(2), 38–59.

Stoynoff, S., & Chapelle, C. A. (2005). *ESOL tests and testing: A resource for teachers and administrators*. Alexandria, VA: Teachers of English to Speakers of Other Languages.

Tang, K. L. (1996). *Polytomous item response theory models and their applications in large-scale testing programs: Review of literature.* (TOEFL Monograph Series MS-2). Princeton, NJ: Educational Testing Service.

Tang, K. L., & Eignor, D. R. (1997). *Concurrent calibration of dichotomously and polytomously scored TOEFL items using IRT models.* (TOEFL Technical Report TR-13). Princeton, NJ: Educational Testing Service.

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

Thompson, I. (2001). From the special issue editor (Introduction). *Language Learning & Technology*, *5*(2), 2–3.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., Steinberg, L., & Thissen, D. (2000). *Computer adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Erlbaum.

Weigle, S. C. (2002). *Assessing writing*. New York: Cambridge University Press.

Winke, P., & MacGregor, D. (2001). Review of Version 5, Hot Potatoes. *Language Learning & Technology, 5*(2), 28–33.

https://doi.org/10.1017/S0267190505000127 Published online by Cambridge University Press