CHAPTER ONE

The technology thread

It would be difficult to estimate how many second language learners today have taken or will take a language test delivered by computer, but many high- and low-stakes tests are delivered by computer and the number is rapidly increasing. This fact of language testing in practice is reflected in a thread that runs through the Cambridge Language Assessment Series. The author of each book in the series suggests that computer technology plays a role in language assessment, and particularly in its future. In his survey of vocabulary assessments, for example, Read (2000) includes the computer-based Eurocentres Vocabulary Size Test and the Test of English as a Foreign Language. Beyond this discussion of computer-delivered tests, however, he points out that computer-assisted methodologies are essential for an understanding of vocabulary that is needed to move vocabulary assessment forward. Similarly, Buck (2001) suggests that a critical issue for the future of listening comprehension assessment is presentation of oral language with the support of computer-delivered multimedia. Weigle's (2002) discussion of the future of writing assessment touches upon both the technology-assisted methods of writing assessment, such as computer scoring of written language, and the effects of technology on writing. Alderson (2000) discusses development of a large-scale Web-based test, computer-assisted testing methods for reading, as well as the construct of reading online. Douglas discusses "the pitfalls of technology" (Douglas, 2000, pp. 275ff.) in *Assessing Languages for Specific Purposes*.

Taken together, the strands of the technology thread point to an important change in the fabric of language assessment: the comprehensive

introduction of technology. In this volume, we examine the important developments implied by the new uses of technology for language assessment and explore the changes in professional knowledge required by the use of technology. Throughout the book we use the terms "test" and "assessment" interchangeably as we discuss a full range of high-stakes and low-stakes uses of assessments that draw on technology for constructing test tasks and scoring examinee performance. We have not included the many other uses of computers for data handling, and statistical analysis of what we refer to as traditional or non-computer tests (see Davidson 1996 and Bachman 2004, respectively, for discussion of these topics). In related areas of applied linguistics, such as the study of language use, second language acquisition research and second language teaching, technology has had notable impacts on professional knowledge and practice. In all of these areas, research and practice demonstrate that technology expands and changes the conceptual and practical demands placed on those who use it, and that the new demands can often probe users' understanding of their work in applied linguistics.

In language assessment, as well, exploration of technology for testing has increased to the point that today no matter where second language learners live, they will sooner or later take a computer-assisted language test. One of the largest and best-known second language testing programs in the world, the Test of English as a Foreign Language (TOEFL), is delivered by computer in many countries, and several hundred thousand candidates take it annually (Educational Testing Service, TOEFL Program: http://www.ets.org/toefl/). Likewise in many classrooms and language programs online learning materials such as Longman English Interactive (Rost, 2003) incorporate assessments that serve as diagnostic or achievement tests. The mention of computer-assisted assessment in the other books in this series along with the growing number of testing and instructional programs offering online assessment suggest the importance of technology for the future of assessment. In this book, we expand on this suggestion by discussing the differences that computer-assisted language testing (CALT) makes for language assessment. The people most affected by the changes are test takers, of course, because they are the ones who ultimately use the technology. However, the intended readers of this volume are the professionals who work to help the learners and therefore we begin in this chapter by outlining some of the implications of CALT for teachers, test developers, and language-testing researchers.

Language teachers

Language teachers need a solid understanding of assessment because they help learners to develop self-assessment strategies, test learners in the classroom, select or develop tests for language programs and prepare learners to take tests beyond the classroom and language program. Many teachers meet their responsibility for preparing learners to take high-stakes computer-based language tests with some feelings of anxiety and even anger because of the possibility that taking a language test online may disadvantage learners, keeping them from demonstrating the full extent of their ability. Issues of fairness to examinees are only one set of the concerns that technology raises for the testing process. Others include the knowledge required for selection, use and development of computer-assisted tests. At the same time, teachers and learners may benefit by having access to assessment for placements and diagnosis which may or may not be connected to online instruction, and may offer possibilities for response analysis, feedback, and record keeping beyond what is feasible with traditional assessments.

Selection of tests

Teachers are likely to have the opportunity to choose from among a variety of computer-assisted tests and therefore need to have an idea of how such tests can best be evaluated. Do guidelines from educational measurement for analyzing reliability, validity, practicality, and authenticity, for example, cover all the relevant considerations for evaluation of computer-based language assessment? As in the case of the evaluation of computer-assisted language materials (Susser, 2001), evaluation checklists have been proposed for computer-based tests (Noijons, 1994). They include factors that one might find on any test quality checklist (e.g., clear instructions) with modifications pertaining to the technology (e.g., information about help options). Other points, however, are unique to the physical and temporal circumstances of computer-assisted testing (e.g., security of test response data upon test completion). Such checklists have been drawn primarily from educational measurement (e.g., Green, 1988), and therefore they are expected to form a solid foundation but we should also question the extent to which they include all of the concerns relevant to language assessment. For example, tests in other areas very rarely include any spoken language, and therefore the issues concerning speaking and

listening through the computer are likely to be under-analyzed in such frameworks. We will discuss the evaluation of CALT in Chapter 5.

Classroom assessment

More and more frequently, teachers have access to computer-assisted language tests that are included as part of online language courses, or to the authoring software that allows teachers to create their own tests. Such classroom assessments raise interesting possibilities for assessing student learning systematically and with provision for detailed feedback. This possibility has been identified as one of the potential attractions of CALT from the early days of the use of technology for language learning (Otto, 1989).

An early example was the French curriculum on the PLATO computer system at the University of Illinois, which kept records on the learners' performance during each session of their work over the course of the semester and provided them with summary information about their performance when they requested it (Marty, 1981). The example Marty provided, called the "General Performance Analysis," could be requested by the learner at any point during the semester. The analysis would tell the student, for example, that he or she had worked on 298 grammar categories, and that overall a score of 77% had been obtained across all categories. Upon request, the learner could obtain a more detailed analysis by asking to see the categories in which he or she had scored below 40%. Figure 1.1 depicts the type of feedback that appeared on the screen in response to such a request. The first column refers to a grammar code, the

1	30%	12	Assez with modifier
21	30%	13	De-verb + partitive
37	20%	19	Verb + de + infinitive
42	10%	14	Ne pas not split with infinitive

Press DATA to enter a different score

Press SHIFT-LAB to review a grammar item

Figure 1.1 Analysis of learners' errors from French learning materials (from Marty, 1981, p. 39).

second is the percentage correct, and the third is the number of items that the learner completed on the particular grammatical point. In addition to the grammar code, the learners were given a description of each grammatical point that they would recognize from instruction.

These diagnostic assessments were built over a period of 20 years in an environment where research and development on French language learning and teaching went hand in hand. The complexity inherent in computer-assisted diagnostic assessment calls for a sustained research agenda rather than a one-time project, as description of the large-scale DIALANG project reveals (Alderson, 2000). Commercial publishers with the resources to develop sophisticated online materials are beginning to draw on some of these ideas about diagnostic assessment or achievements designed to match the courses. Online courses in English, such as Market Leader (Longman, 2002), have an integrated assessment component throughout the courses to give pre- and post-test information to learners and teachers. Such tests are developed through application of the well-known principles of criterion-referenced testing, but the example from the French course illustrates that these basic principles can play out differently for development of online tests.

Whenever language instruction is offered online, it makes sense for teachers to at least consider online assessment as well. However, even some stand-alone tests might best be administered by computer when detailed diagnostic information is desired. For example, years ago, Molholt and Presler (1986) suggested that their pronunciation analysis might be used to identify specific aspects of pronunciation in need of instruction. Canale (1986) advocated looking toward intelligent tutoring systems which would be able to gather diagnostic information about learners as they worked online, and a number of such systems have been described for language learning, but such research has largely emphasized the instructional potential of the systems without fully exploring them as assessments (e.g., Holland, Kaplan & Sams, 1994). Future exploration of the detailed information obtained through diagnostic assessment offers interesting challenges to language assessment as a discipline. As Clark (1989) pointed out, diagnostic tests are developed according to different specifications from those used to construct a proficiency test from which a single score is to be obtained. However, the large part of the theoretical and practical knowledge about developing and interpreting assessments has been cultivated for proficiency-type tests, leaving issues of diagnosis somewhat uncharted territory. As more and more people become interested in and capable of developing and using

computer-assisted diagnostic assessments, the issues are likely to be better understood (see Alderson, 2005, for a discussion of these issues).

Test development

Classroom assessments are frequently developed by teachers themselves so as to reflect the important points that were taught in class. Accordingly, a range of options exists for teachers wishing to develop their own online tests. The most efficient option for doing so is course management software that allows the teacher to construct units containing quizzes, that is, to construct the specific questions to be delivered on the quiz and a means for scoring and reporting scores to students and to teachers. Such authoring software is very useful in allowing teachers access to the authoring process with very little training. However, as Chapter 4 will explain, efficiency is often obtained at the expense of the specific features that would be desirable such as a variety of item types and linguistically sensitive response analysis. Nevertheless, such general-purpose authoring software provides teachers access to the authoring process and to some of the capabilities of CALT.

As a consequence, teachers can work together to develop assessments that fit into their program. For example, the English Language Institute (ELI) at the University of Surrey, in the UK, has developed a number of self-access activities designed to complement the courses they offer. The activities include short quizzes which provide instant feedback to learners so they can assess their own learning, as illustrated in Figure 1.2, from a quiz on thesis writing. Teachers and students might benefit from developing and using such an online quiz, which would not require sophisticated authoring tools.

Test developers

Professional developers of computer-assisted tests work with a much wider set of options than that which used to be available for test development including delivery options that expand the ways in which language can be assessed. New methods include computer-adaptive testing, the use of multimedia for presenting linguistic and visual input for learners, and automatic response analysis. These new methods raise questions for test developers about what the new language tests are measuring.

Thesis1
Thank you for taking the Thesis Writing Unit 1 Self-Access Quiz

· 1

- 3 out of 5

Section 1: Preparation In the preparation stage of your thesis, before you actually embark upon your research, once you have decided your topic, a number of activities are of particular importance. In the following list, select the 5 most important activities.

✔

Establishing objectives was correct
A correct answer was Writing initial outline proposals

Formulating the title helps clarify your thinking at the beginning, even if you change your mind later. You need to establish objectives as soon as possible, to make sure that your research has a clear direction. This also makes it easier to select reading! Initial outline proposals also help to clarify issues. The focus of the topic is crucial: it must not be too broad or too narrow. Finally, it is always important to write a timetable to establish deadlines for completing work.

Figure 1.2 Surrey ELI Self-Access Quiz feedback (http://www.surrey.ac.uk/ELI/sa/thesis1.html).

Computer-adaptive testing

Many professional test developers associate computers for test delivery with the development of large pools of items for computer-adaptive tests (CATs). A computer-adaptive test selects and presents items in a sequence based on the test taker's response to each item. If an examinee gets the first question correct, a more difficult question is selected from a pool and presented next; if this one is answered correctly, a more difficult one is selected. If the candidate misses a question, the algorithm selects an easier one for the next question, and so on. A CAT program "learns"

about the examinee's level by monitoring the difficulty of the items the test taker gets right and wrong and thus begins to select only those items at the candidate's level of ability. When the program has presented enough items to be able to estimate the test taker's ability at a predetermined level of reliability, the test ends and a score can be reported. CATs are efficient because they present items to test takers close to their level of ability, thus avoiding items that are either too easy or too difficult and which consequently would not offer much information about a test taker's abilities.

Test developers were introduced to the advantages of computer-adaptive testing at least 20 years ago. Tung (1986) outlined the following advantages: they require fewer items than their paper counterparts, they avoid challenging examinees far beyond their capability by selecting items at the appropriate difficulty level, and they offer improved security by selecting from an item pool to construct individualized tests. CATs became possible through developments in measurement theory called Item Response Theory (Lord, 1980; Hambleton, Swaminathan & Rogers, 1991), a means for obtaining robust statistical data on test items, and through advances in computer software for calculating the item statistics and providing adaptive control of item selection, presentation and evaluation (Green, Bock, Humphreys, Linn & Reckase, 1984; Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg & Thissen, 1990; Brown, 1997). See Bachman (2004, Chapter 3) for an accessible conceptual introduction to IRT.

Following examples in the early 1980s at Brigham Young University developed by Larson and Madsen (1985), other computer adaptive language tests were reported throughout the 1990s (e.g., Kaya-Carton, Carton & Dandonoli, 1991; Burston & Monville-Burston, 1995; Brown & Iwashita, 1996; Young, Shermis, Brutten & Perkins, 1996). Through these projects, important issues were raised about the way language was being measured, about the need for independent items, and about their selection through an adaptive algorithm. In an edited volume in 1999, Chalhoub-Deville brought together a range of theoretical and practical perspectives to discuss computer-adaptive testing for L2 reading. Theoretical papers emphasized the multidimensionality of the reading construct, whereas descriptions of testing practice spoke to the need for unidimensional scores, particularly for placement (e.g., Dunkel, 1999; Laurier, 1999). Results from this work suggest that computer-adaptivity can be used to construct efficient language tests to test language abilities such as reading comprehension, but at the same time most would agree

that such tests fail to take advantage of the range of capabilities that the computer offers.

The notion of adaptivity continues to be explored and expanded, and now can refer to any form of branching, or alternative path options, that are chosen for students to take within a program based on their responses. For example, tests of the future might expand on a current example, Longman English Assessment, which branches to either general-purpose or specific business content, depending on the examinee's response to an interest questionnaire at the beginning of the test. In this case, the content of the language of the input is adapted to students' interests, to some extent. In other cases, test tasks might be adapted based on the examinee's level of performance on preceding sections of the test. In short, test developers have barely begun to scratch the surface of the ways in which a test might be tailored to fit the examinee. This is an area in which technology challenges test developers to construct tests that are suited to the needs and interests of learners.

Multimedia tasks

Another potentially powerful option that computers offer test developers is the provision for rich multimodal input in the form of full motion video, text, sound, and color graphics, potentially enhancing authenticity of both input and response. Test developers are concerned with enhancement of two aspects for authenticity: *situational authenticity*, which defines authenticity in terms of the features of context including setting, participants, content, tone, and genre, and *interactional authenticity*, which defines authenticity in terms of the interaction, between the test taker's language knowledge and the communicative task (Bachman 1991). In some cases, multimedia can help to portray these aspects of a non-test situation on a test. For example, a placement test, in the Web-based Language Assessment System (WebLAS), at the University of California, Los Angeles, developed to provide information about placement, progress, diagnosis, and achievement in second and foreign language teaching programs at UCLA, uses video to present lecture content for comprehension tasks. The use of the video is intended to enhance the situational authenticity of the test by depicting the features of academic context such as a classroom, white board, and PowerPoint slides. One can envisage other situations such as following a tour guide, checking in at a hotel, or participating in a business meeting where the video would also

add to the non-test context that the test is intended to portray. Examples of such scenarios are contained in multimedia software for language learning, which provide good examples of possibilities for test developers.

Automatic response analysis

Tests which call on the examinee to produce language hold the potential for increasing interactional authenticity over those that require selected responses since the former typically require a greater breadth and depth of language knowledge and background knowledge, and more sophisticated use of strategic competence. Some language test developers have explored the use of natural language processing technologies to construct scoring procedures for examinees' linguistic production. An automated speaking assessment, PhonePass (Ordinate Corporation, 2002b), for example, scores the accuracy of repeated words, pronunciation, reading fluency, and repeat fluency, based on a computer speech recognition system containing an algorithm derived from a large spoken corpus of native speakers of various English regional and social dialects. The Educational Testing Service, which produces the TOEFL as well as a number of other academic and professional tests, has developed an automated system, Criterion (2005a), for rating extended written responses, based on natural language processing (NLP) technology that syntactically parses input, identifies discourse structural information of selected units of text, and analyzes topical vocabulary, to produce a holistic rating of an essay on a six-point scale.

New test methods, new constructs?

In the first collection of papers on CALT, Canale (1986) pointed out that the use of the computer held the promise of providing a better means for measuring different language constructs than that which was possible with traditional test methods. However, research and development has tended to focus on the goals of increasing efficiency and authenticity of testing, whereas to date few researchers have explored the intriguing questions of how the computer might be used to assess different abilities, or constructs, than those currently assessed by traditional methods. These issues were discussed by Alderson, who outlined computer capabilities relevant to exploring an innovative agenda for CALT: