CHAPTER SIX

The impact of CALT

Bennett, a measurement researcher and an enthusiastic advocate of technology, writes about the transformative impact of technology on large-scale educational assessment:

> New technology will permit [a] transformation [in assessment] by allowing us to create tests that are more firmly grounded in conceptualizations of what one needs to know and be able to do to succeed in a domain; by making performance assessment practical and routine through the use of computer-based simulation, automatic item generation, and automated essay scoring; and by changing the ways in which we deliver, and the purposes for which we use, large-scale tests. (Bennett, 1999a, p. 11)

The seeds for Bennett's enthusiasm about the transformative power of technology for assessment were planted much earlier, and many of these ideas have been hinted at by researchers in educational measurement for years (e.g., Bejar, 1985; Cole, 1993; Bejar & Braun, 1994). Although Bennett and other enthusiasts typically do not refer specifically to second language tests, they regularly include writing tests in their discussions. In reading these predictions, second language teachers, test developers, and researchers cannot help but consider whether or not our assessments are part of the revolution in assessment, and if so whether or not this revolution has happened, is in progress, or is yet to come.

In this chapter we will suggest that in second language assessment, despite the significant changes and advances made through the use of technology, the revolution portrayed by Bennett has not yet occurred.

Although automated scoring and computer-assisted test delivery are realities, we were unable to show evidence for performance assessment made practical through widespread use of simulation, authentic item generation, or significant changes in testing purposes through technology. Such revolutionary changes need to be prompted and supported by conceptual advances in our understanding of language, language use, and language learning. More than 15 years ago, Alderson made this point in his discussion of individualized classroom testing:

> Possibilities for diagnosis and remediation raise an important problem for applied linguists and language teachers. The limitation on the development of such tests is not the capacity of the hardware, or the complexity of the programming task, but our inadequate understanding of the nature of language learning and of language use . . . the challenge of [CALT] is more to the linguist and applied linguist to provide appropriate input on the nature of branching routines, and on the hints, clues and feedback that would help learners, than to the computer programmer to produce adequate software.
> (Alderson, 1990, p. 25)

Alderson was focusing on CALT in the service of classroom learning, but analogous comments could have been made about other types of tests as well. Today there is little evidence to suggest that great progress has been made toward building the specific types of knowledge that could fuel the revolution. In fact, the recent discussion of the DIALANG project (Alderson, 2005), a large-scale diagnostic computer-delivered test, makes the same point. DIALANG was developed on the basis of the levels of the Common European Framework, but because research on second language acquisition has not examined development of grammar and vocabulary in view of these levels, the would-be diagnosis of specific linguistic forms and functions does not have a clear basis in either theory or research. In this case, the attempt is being made to expand test uses radically by providing learners with diagnostic information that can give precise guidance about what to study; however, the professional knowledge of second language acquisition falls short of providing the basis for such a revolutionary change.

Despite the fact that the revolution has not occurred, in language-testing practice today the idea that technology has changed language testing holds some credibility. After all, many language tests are delivered by computer and books on every aspect of language assessment predict a profound role for technology in the future. On the surface, anyway, it seems that many test takers are affected by CALT. We would suggest that

these changes would most appropriately be considered evolutionary, because they have not significantly advanced and changed the way testing is conducted or the ways tests are used. Instead, through the use of CALT, researchers have made incremental advances in addressing some of the perennial problems in language assessment. Rather than describing a revolution characterized by new types of tests and roles for testing, we have written about the evolutionary advances and the issues they have raised about test development and validation. In this chapter, we consider some of the implications of these advances for applied linguistics. We also suggest future directions and discuss what revolutionary changes might entail.

Advances in language assessment through CALT

In the previous chapters we have described CALT used in practice today, research intended to increase understanding of and improve CALT, methods used to develop CALT, and validation issues for CALT. The first chapter argued that technology plays such a significant role in everyday assessment practices that knowledge and an understanding of technology-related issues is essential for language teachers, professional test developers, and language-testing researchers.

Chapter 2 demonstrated the ways that CALT is being used in many testing programs and situations, expanding the test developer's options for constructing test tasks. We described ways in which technology affects the test method characteristics including physical and temporal test circumstances, the test rubric, input and response characteristics, the interaction between input and response, and the characteristics of assessment. We showed that the most salient differences are to be found in the characteristics of the input and response, the interaction between them, and assessment. Advances in test delivery and access were evident in the examples of rich contextualized input, the variety of response techniques, computer adaptivity, and automated scoring made possible by computer. However, we were unable to report on revolutionary changes such as performance assessment made practical through the use of simulations.

In Chapter 3, we discussed the potential problems raised by CALT in terms of the way they affect validity. We noted six concerns that are often expressed as potential threats of CALT: different test performance, new task types, limitations due to adaptive item selection, inaccurate

automatic scoring, compromised security, and negative impact. We noted that few studies have attempted to better understand the meaning of test scores from CALT and the consequences of CALT use. We also noted that the potential threats to validity were framed in terms of suspicions about how technology might undermine the validity or fairness of testing. In other words, the most public discussion of CALT has been framed by a skeptical view that asks how technology might undermine current practices rather than an innovative perspective that seeks to discover how technology can contribute to a revolution which significantly improves the overall usefulness of assessment.

In Chapter 4, we discussed how would-be CALT developers might work with authoring tools such as WebCT, pointing out that the tools required for test development depend on the purpose of the assessment and such practical issues as the resources of time, money and expertise. The reality of language assessment is that limitations in money and expertise for developing authoring tools specific to language assessment limit the degree to which revolutionary innovations are likely to be developed. Consequently, ideal authoring systems have not been developed for language assessments, but this is an active area of inquiry.

Chapter 5 suggested that computer-based testing should be evaluated against standards that are consistent with those used to evaluate other tests, but that technology-related issues need to be highlighted. The specific technology-related issues identified by CALT researchers should be placed within a broader framework of test evaluation, focusing on aspects of test usefulness, which highlights factors of particular concern in language assessment such as authenticity as a key issue for test development and evaluation. We explored an interpretive argument which laid out a structure into which specific test-taking and interface concerns might be situated. However, this discussion was necessarily tentative in view of the limited amount of research reported on empirical validation of CALT from current perspectives.

In sum, the reality of CALT today is not what one could call a revolution in language assessment. Certain characteristics of CALT methods are substantively different from those of tests involving other means of delivery and response, but technology has not radically reconfigured the role of assessment in teaching and learning. Thus far we have seen CALT as an evolution in assessment, expanding what we do in testing, rather than a revolution, changing what assessment is in relation to language education and research. A revolution may be coming sometime in the future, but in the meantime, in view of the central role that language

assessment plays in applied linguistics, the changes brought about by technology intersect in important ways with other areas of applied linguistics.

CALT in applied linguistics

The development, use, and evaluation of CALT challenges and expands the imaginations of applied linguists because of the new options opened by testing through technology. One issue is the need to consider the nature of the language abilities that are called upon in technology-mediated interactions and communication, and therefore, the need to rethink test constructs. A second issue is that the precision of the information about learners that can be analyzed on the basis of examinees' constructed test responses prompts test designers to consider what to do with such capabilities. For example, should test developers reconsider how research on SLA can inform the development of tests that provide more detailed information than tests relying on human raters? A third issue is that the flexibility of the technology for anytime, anywhere testing and record keeping appears to afford powerful opportunities for improving instruction through assessment. These three issues, which have persisted throughout this volume, are worthy of additional discussion.

Language ability and use

Investigation of CALT underscores the fact that the language constructs underlying score interpretation need to be considered in view of the context in which the language is used. Applied linguists would therefore speak of language ability as the ability to choose and deploy appropriate linguistic resources for particular types of situations. But today we might replace such a conception of language ability with one that encompasses the ability to select and deploy appropriate language through the technologies that are appropriate for a situation. Email is good for some things; a phone call or a face-to-face conversation is better for others. The language user often makes the choice. The spell-checker is informative sometimes; it needs to be ignored at others. The language user makes the choice. These choices ultimately depend on the language user's technological and strategic competence, which together with linguistic competence may be the type of construct of relevance to language use through technology.

In other words, communicative language ability needs to be conceived in view of the joint role that language and technology play in the process of communication. Rassool (1999) brings communicative competence into the modern era by suggesting that "communicative competence refers to the interactive process in which meanings are produced dynamically between information technology and the world in which we live" (p. 238). From the perspective of language assessment this statement raises the polemical issue of a context-dependent language construct. Literacy researchers such as Tyner (1998) begin to explore what this could mean: they see the need to splinter the construct of literacy to express their belief that technology affects the nature of literacy required for language use *with different technologies*:

> New approaches to literacy teaching and learning suggest that instead of approaching literacy as a monolithic concept . . . it is more useful to break literacy down into any number of multiple literacy modes, each with distinctive characteristics that reveal a variety of social purposes . . . These multiple literacies have been called *technology literacy*, *information literacy*, *visual literacy*, *media literacy*, and so on . . . As contemporary communication media converge into sensory soup, the particular features of each of these literacies also converge and overlap . . . (Tyner, 1998, p. 60)

Such a proliferation of literacies may reduce the term to denote any ability, whether or not it entails language. The implication for language assessment is not completely clear. However, the idea that various technologies might affect the abilities of interest in language assessment seems like an issue that needs to be considered in the process of test development and validation today whether or not CALT is involved. Bruce and Hogan (1998) express this idea in terms of technology being an integral part of communication: they point out that anyone who is not competent in using the technology is not competent in communication in many important situations. A similar issue is evident in attempts to define a construct underlying a test of language for specific purposes. It is both the specific purpose knowledge and the linguistic and strategic competence that work together to accomplish communication (Douglas, 2000). Likewise, it is probably the combination of language and strategic competence together with technological knowledge that accomplishes communication through technology.

Elaborating on the view of multiple technology-defined literacies, Warschauer argues that the literacy skills learners need to acquire in today's world are qualitatively different from those they need to participate

in literate life that does not involve technology. Warschauer (2000) describes new language and literacy skills needed for effective communication by replacing the constructs of reading and writing with the abilities that he refers to as reading/research and writing/authorship, respectively (p. 521). These constructs, which include aspects of the strategic competence required to perform successfully in some electronic environments, force test developers and users to confront how strategic competence is to come into play. This is not a new problem, but it is one that is exposed and amplified through CALT. In this way, CALT provides both the need and the opportunity to better understand the language abilities called upon in computer-mediated communication. In terms of the interpretive argument explained in Chapter 5, the test developer would need to express the score interpretation in terms of ability to gather visually presented information on the Internet rather than in terms such as "reading ability" in general.

In attempting to formulate theory-based perspectives on the abilities required for use of language through technology, language-testing researchers face the challenge of integrating the seemingly incompatible discourses of language assessment and literacy studies. Literacy studies take a social practice perspective entailing description of behavior rather than the more cognitive perspective that underlies much language assessment work. From the perspective of social practice, electronic literacy and multimodal literacy, for example, are seen as what people do with language through technology rather than what they need to know about language and the strategies they need to use language through technology. Some argue that an ability perspective is incommensurable with the social practice perspective of new literacy studies because the former typically entails defining a theoretical ability that is responsible for performance across contexts, and the latter is based on description of specific context-based performances. It may be that exploration of interpretive arguments for CALT will prompt applied linguists to better understand the abilities underlying electronic literacy, or multimodal literacy, in a manner that is measurable and that yields interpretable and useful scores.

Second language acquisition

Development and evaluation of CALT underscores the need to strengthen connections between second language acquisition (SLA) and language assessment. Anyone who has attempted to design the specifics of testing

method for a computer-delivered test or to tackle the problem of assigning partial-score values to examinee responses can grasp the issue at stake. In providing written or aural input to the examinee, should help be offered as well? In assigning a score from one to six to a written essay, what makes a five better than a four? What are the specific linguistic, rhetorical, and content features that prompt the assignment of a particular score on the six-point scale? In developing a rationale for issues such as these, one would hope that theory and research on SLA would be informative in at least three ways.

First, it would seem that research investigating developmental sequences of acquisition should play a role in developing and scoring some assessments in which grammar plays a role in performance. For example, Norris (in press) wrote items for a grammar-screening test on the basis of studies in SLA that have demonstrated the relatively earlier development of some grammatical knowledge over other aspects. Low-level examinees were expected to perform better on items requiring the ordering of words in single-clause declarative sentences than on items requiring ordering of words in sentences with embedded noun clauses. These items were scored as either correct or incorrect, and no linguistic production was scored, but one would hope to be able to explore the analysis of learner's production in light of work in SLA.

In Chapter 3 we discussed the tension felt by test developers in need of rationales underlying procedures for scoring responses with multiple possible responses. Even the non-linguistic responses of the text-sequencing task investigated by Alderson, Percsich, and Szabo (2000) needed to have a basis for assigning the scores to the many various orderings that the examinees might give. But the issue was exacerbated by the enormous variations that might be entered by examinees taking Coniam's (1998) dictation test, which used an automatic scoring routine, the reading test with open-ended responses developed by Jamieson, Campbell, Norfleet, and Berbisada (1993), or the essays discussed by Powers, Burstein, Chodorow, Fowles, and Kukich (2001). In all of these cases, and many more that we might imagine, the developers should benefit from a scientific basis upon which they can consider some performance as evidence of advanced knowledge and some other performance as lower level. Ideally, some professional knowledge about acquisition could be drawn upon for CALT.

The work that has attempted to identify linguistic features associated with levels of ESL writing might offer some suggestions for developing rationales for response analysis in assessment. Wolfe-Quintero, Inagaki,

and Kim (1998) review the issues and findings of research seeking syntactic performance indicative of levels on a developmental index; they consider development along the dimensions of fluency, accuracy, and complexity. Hinkel (2003) identifies lexical choices made by ESL writers that contribute to the perception of simplistic and imprecise writing. Both of these lines of research appear to be productive in developing a better understanding of levels of performance. At the same time, such research is necessarily limited if it is interpreted as suggesting that linguistic knowledge is acquired in an invariant order or that linguistic knowledge is impervious to the conditions under which it is displayed. Without disregarding empirical results that provide evidence for more, or less, linguistic knowledge, language-testing researchers need to take into account the cognitive and contextual factors that also come into play during performance.

A second way in which SLA research might fruitfully inform test development and response scoring is through research identifying the effects of processing conditions on performance. The assumption underlying this strand of SLA research is that performance needs to be explained in view of not only the knowledge of the examinee, but also the conditions under which performance was obtained. Researchers such as Skehan (1998) and Robinson (2001) therefore hypothesize particular task characteristics that should be expected to produce more, or less, difficult conditions for performance. For example, a requirement to produce a written response quickly vs. slowly would interact with the examinee's level of linguistic knowledge to produce a higher or lower level of response. In other words, dimensions other than level of knowledge need to be taken into account in interpreting performance.

A third dimension that needs to be integrated is the context constructed for the examinee's performance. The extended language production desired in tests involving speaking and writing is produced in response to a prompt intending to create a particular contextual configuration (Halliday & Hasan, 1989) for the examinee. This in turn cues the examinee to produce genre-appropriate language to accomplish a communicative function (Paltridge, 2001). Such tests, if they are well designed, help the examinee to understand and create a discourse domain (Douglas, 2000), which includes the topic of the response, the audience, and its communicative function. As a consequence, the job of analyzing the language produced on such tasks is not at all an open-ended, general problem of language analysis, but rather a problem that can be characterized empirically by the functional grammarian's description (e.g., Halliday, 1994) of

the ways in which learners at various levels of ability deploy their limited linguistic resources to construct meaning in a well-defined context. This is the type of work that has been conducted in computational linguistics and artificial intelligence for over 40 years, except that such research has been concerned with typical proficient-speakers' performance in defined contexts rather than levels of learners' performance.

Working toward a better understanding of these three factors in shaping test performance seems to be an essential step for today's machine-scored assessments as well as for the more revolutionary intelligent assessment of the future. Bennett (1999b) describes "intelligent assessment" (p. 99) as an integration of three lines of research: constructed-response testing, artificial intelligence, and model-based measurement. He explained:

> This integration is envisioned as producing assessment methods consisting of tasks closer to the complex problems typically encountered in academic and work settings. These tasks will be scored by automated routines that emulate the behavior of an expert, providing a rating on a partial credit scale for summative purposes as well as a qualitative description designed to impart instructionally useful information. The driving mechanisms underlying these tasks and their scoring are . . . measurement models [grounded in cognitive psychology] that may dictate what the characteristics of the items should be, which items from a large pool should be administered, how item responses should be combined to make more general inferences, and how uncertainty should be handled. (p. 99)

To support such research in second language testing, however, would require substantial collaboration between language assessment and SLA at least in the three areas outlined above. Such connections are not altogether implausible (e.g., Brown, Hudson, Norris & Bonk, 2002) despite the fact that the two areas of applied linguistics seem to speak different languages. Even the most basic working constructs such as units of analysis are different, with assessment researchers talking about reading, writing, listening, speaking, and SLA researchers talking about the tense and aspect system, the negation system, or polite requests, for example. Attempts to bring measurement concepts to bear on the complex data of interest in SLA (e.g., Chapelle, 1996; Bachman & Cohen, 1998) need to be developed into a more systematic program of research. This is what Norris and Ortega (2003) suggest in their review of measurement practices in SLA: The SLA community needs to "engage in a comprehensive approach to all of the stages in the measurement process [in order to] find

itself much better able to make theoretically meaningful interpretations about constructs and to pursue the accumulation of scientifically worthwhile knowledge" (p. 749).

For such knowledge to ultimately inform the design of CALT, however, language-testing researchers need to be able to distinguish between the SLA knowledge connected to theoretical debates in SLA and that which can inform assessment. An analogous distinction has usefully been made in SLA in general and the more focused area of "instructed SLA." The latter focuses on the conditions for language performance and acquisition that pertain to instructional decisions. Of particular relevance is research aimed at increasing the research-based knowledge about pedagogic tasks (e.g., Crookes & Gass, 1993; Bygate, Skehan & Swain, 2003). It is this area that is appropriately focused to serve as a basis for hypotheses and empirical research about computer assisted language learning (CALL). Similarly, an area that one might call "assessed SLA" is needed to focus on aspects of acquisition that can be empirically observed in performance under particular cognitive conditions and in defined contexts.

Language teaching

In the first chapter we suggested that 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. However, perhaps the most provocative vision for language assessment in the classroom is the potential for assessments to help students to become better, more autonomous learners.

In Chapter 2, we described some examples of CALL programs such as Longman English Interactive, and Market Leader that contained testing and feedback to learners within the instructional packages. The idea is that if learners can be regularly informed about the quality of their knowledge and progress as they proceed through instruction, they can make better choices about studying, and ultimately become more self-reliant. Of course, if these capabilities are to be implemented, courseware developers need to have a firm understanding of the principles of assessment in the service of learning.

Moreover, assessments are not necessarily used simply because publishers have produced them. Teachers need to learn about the potentials of computer-assisted assessment if they are to introduce them to learners.

In other words, the divide that seems to exist between language testers and language teachers is dysfunctional with respect to the aim of expanding the uses of assessment in revolutionary ways. Stoynoff and Chapelle (2005) argue that it is essential to move beyond this divide and that language teachers need to become assessment literate in order to select and construct tests for learners. The potential of new uses for assessments integrated into computer-assisted learning materials creates an additional motivation for teachers' assessment literacy. In this respect, CALT might be seen as providing a powerful opportunity for positive impact within the profession that goes beyond the types of washback that have been the focus of recent research (e.g., Cheng, Watanabe & Curtis, 2004).

Future directions

Today's CALT raises issues that must be explored if it is to evolve sufficiently to become part of a revolution in assessment. Current technologies represent an embarrassment of riches for test developers – from test delivery at a distance, precise control over timing and multimedia input for examinees to natural language processing and student models. The tools for test building have become extremely sophisticated. If test developers are to make appropriate use of such tools, research needs to be guided by a clear agenda in applied linguistics which is supported by cross-disciplinary knowledge.

A cross-disciplinary project

Although the precise issues raised by technology fall squarely within the domain of problems that applied linguists should know how to address, the tools for addressing them need to be developed and tested in an arena where cross-disciplinary collaboration is brought to bear on the issues. In Chapter 4 we discussed authoring tools such as WebCT, Respondus, Hot Potatoes, Quiz Center, Blackboard, and Questionmark. Whereas these systems provide tools for developing tests in general, we saw that they did not contain specific language-related features, most notably capture of spoken linguistic responses and a means of analyzing constructed responses to assign a rationale-based partial score. Software tools specific to the needs of second language testing need to be developed based on the limitations of existing tools for language testing.

Teachers and software developers have been creating individual tests using general purpose authoring or specific programming languages for over 30 years (e.g., Boyle, Smith & Eckert, 1976). However, if this experience and knowledge base is to be developed in a cumulative fashion, professional quality authoring tools are needed for the applied linguistic community to use. Development of a robust set of appropriate tools requires a group of professionals comprising at least software engineers, language assessment specialists, and designers. Without the basic software tools that graduate students can use to learn about testing, it seems that the level of discussion about test design is confined to a level of unprofessional speculation about what *might* work and what *would be* interesting. For example, the Dutch CEF Construct Project (Alderson, Figueres, Kuijper, Nold, Takala & Tardieu, 2004) is an example of a piece of software that is intended to help test designers develop and analyze test tasks according to a construct-based framework (like the Common European Framework of Reference – CEFR). Projects such as DIALANG have taken some steps to develop a variety of item types and hopefully will develop authoring tools as well that will allow other authors to experiment with them (Alderson, 2005).

Other glimpses of what is possible with sophisticated software can be found in papers about intelligent computer-assisted language learning (e.g., Chanier, Pengelly, Twidale & Self, 1992), which are the product of such cross-disciplinary research. Such a system contains the elements similar to those Bennett described as essential for intelligent assessment – analysis of learners' constructed responses, a student model which is updated on the basis of analysis of examinees' responses, and an expert system that selects probes for the learner to gain more information. It is not clear to what extent any measurement concepts come into play in this system, which is not intended specifically for assessment. But the point is that such complex systems are being explored in other areas, and that making them accessible to researchers in language assessment requires more sophisticated authoring tools than those which one finds for developing classroom tests.

An applied linguistics agenda

Despite the need to draw on expertise across the areas of educational measurement, applied linguistics, second language acquisition, and technology, the agenda needs to be set and driven by the concerns of

applied linguists for assessment. However, even within applied linguistics, a carefully articulated stance needs to be developed toward technology. Based on analysis of approaches toward developing agendas for research and practice in language assessment, Chapelle (2003) identifies three approaches that are taken, as summarized in Table 6.1.

The tunnel approach, as in "tunnel vision," refers to a metaphor from Brown and Duguid (2000), who describe technologists across all facets of society as moving single-mindedly to goals of speed and efficiency without regard for anything else. In language assessment, technology is often construed in this way – as a means of constructing more efficient tests. If efficiency is the goal, the desired results are shorter, more convenient tests. In other words, the argument to be made by test developers is that the computer-based test can do the same thing as the tests offered in other forms, except faster and cheaper.

A comparison approach to CALT treats the technology as suspect, and therefore the problem for research is to discern the differences between computer-based tests and other types of tests. Such analyses can be conducted at the level of performance on a whole test or it can be studied at the level of item performance. What underlies this perspective, however, is the view that the no-technology condition is the normal one, and then the problem is to figure out what difference the technology makes. Both the tunnel and the comparison approaches clearly seek to achieve worthwhile goals. In applied linguistics, who would suggest that more efficient

Table 6.1 *Assumptions about technology and results of tunnel, comparison, and innovation approaches* (*From Chapelle, 2003, p. 179*)

Approach	Assumption about technology in assessment	Results
Tunnel	It is an efficiency	Short tests with automatic scoring and delivery of results for existing test uses
Comparison	It should be considered suspect	A variety of types of tests for existing test uses; knowledge about how technology affects traditional tests when they are delivered online
Innovation	It should be considered a resource	A variety of types of tests and new test uses; knowledge about the intersection of technology with a variety of assessment issues

and convenient tests are not desired? Who would deny the value of better understanding how technology affects performance conditions and test results? However, while these two perspectives are clearly in line with applied linguistics and language assessment, each is limited in its capacity to revolutionize language assessment in the ways that Bennett described.

The revolution may lie within the innovative approach, which draws on technology as a resource to explore a variety of assessment issues. Chapelle (2003) suggests that such innovation entails development of a variety of tests and test uses that are not possible without technology. To do so would require the types of language testing software tools mentioned above, and should also entail the use of technology for developing knowledge about the intersection of technology with a variety of assessment issues. Educational measurement researcher Eva Baker suggests that such an agenda of innovation is at the heart of the revolution in which technology is to play an important role. She argues that "Technology applied to the service of understanding the learning we want will help us fix the presently unfixable – the deep validity problem at the heart of our testing system" (Baker, 1998, p. 22).

Conclusion

The suggestions and questions that appear within an innovative approach to CALT are many of the same ones posed by the pioneer-innovators in this area over ten years ago. Canale (1986) suggested the use of intelligent tutoring technologies to model learners' knowledge and inform instruction. Alderson (1988) pointed out that the computer can make use of language rules for analysis of learners' constructed responses. Corbel (1993) asked about the possibilities of intelligent assessment, CALT to aid in self-assessment, and strengthening links between assessment and other areas of applied linguistics through technology. Despite the evolutionary developments in assessment that have incorporated technology, we are not able to report on any revolutionary changes in assessment that might have resulted from systematic inquiry into these areas. Computer technology may in the future radically change research and practice in language assessment but doing so will require the type of research that engages with the complexity of the issues, crossing the boundaries between assessment, language, and technology for the purpose of developing paths that work toward the goals of applied linguists.