A Critical Review of Selected Computer Assisted Language Testing Instruments

Magdolna F. Silye – Troy B. Wiwczaroski

University of Debrecen, Centre of Agricultural Sciences,
Faculty of Agricultural Sciences,
Centre of Technical Languages Instruction, Debrecen

SUMMARY

Over the past few decades, a fairly large literature examining the effectiveness of computer-assisted language learning (CALL) has been developed. The findings indicate that language learners have generally positive attitudes toward using computers in the classroom. Less is known, however, about the more specific areas of computers in language testing. The purpose of this article is to examine recent developments in language testing that directly involve computer use. After a brief overview of computer-based testing (CBT) in general, web-based testing (WBT) is defined and certain issues reviewed.

1. COMPUTERS IN LANGUAGE TESTING

An abundance of literature on computer-assisted language learning supports that both language learners and instructors have generally positive attitudes toward using computers in the classroom, but less is known about a more specific area of computer use, i.e. language testing.

In reviewing the literature on computers in language testing, one can find three recurring sets of issues: (1) item banking, (2) computer-assisted language testing, and (3) computer-adaptive language testing.

1.1. Item Banking

Item banking covers any procedures that are used to create, pilot, analyze, store, manage, and select test items so that multiple test forms can be created from subsets of the total "bank" of items. With a large item bank available, new forms of tests can be created whenever they are needed.

While the underlying aims of item banking can be accomplished by using traditional item analysis procedures, a problem often occurs because of differences in abilities among the groups of people who are used in piloting the items, especially when they are compared to the population of students with whom the test is ultimately to be used.

One can assume that item banking is not without potential problems. Henning (1991), for example discusses specific problems that may be encountered with the validity of item banking techniques in language testing settings. Another serious limitation is the large number of students that must be tested before it can responsibly be applied. Typically, the item response is only applicable for full item analysis (that is, for analysis of two or three parameters) when the numbers of students being tested are very large, that is to say, over one thousand.

Minimal item banking can be done without computers by using file cards, and, of course, the traditional item analysis statistics can be done (using the sizes of groups typically found in language programs) with no more sophisticated equipment than a hand-held calculator. Naturally, a personal computer can make both item banking and item analysis procedures much easier and much faster. For example, standard database software can be used to do the item banking (e.g., Microsoft Access, 1996; or Corel Paradox, 1996).

An example of a software program specifically designed for item banking is the PARTest (1990) program. If PARTest is used in conjunction with PARScore (1990) and PARGrade (1990), a completely integrated item banking, test analysis, and record-keeping system can be set up and integrated with a machine scoring system.

1.2. Computer-Assisted Language Testing

Tests that are administered at computer terminals, or on personal computers, are called computer-assisted tests. Receptive-response items-including multiple-choice, true-false, and matching items-are fairly easy to adapt to the computer-assisted testing medium. Relatively cheap authoring software like Testmaster (1988) can be used to create such tests. Even productive-response item types-including fill-in and cloze-can be created using authoring software like Testmaster. Unfortunately, the more interesting types of language tasks (e.g., role plays, interviews, compositions, oral presentations) prove much more difficult to develop for computer-assisted testing.

The new technologies such as the CD-ROM and interactive video discussed in Brown (1992a) do make it possible for students to interact with a computer. Hence, no technical reason remains why interactive testing like role plays, interviews, compositions, and presentations cannot be done in a computer-assisted mode.

All in all, two primary benefits can be gained from computer-assisted language testing:

1. Computer-assisted language tests can be individually administered, even on a walk-in basis. Thus group-administered tests and all of the organizational constraints that they impose will no longer be necessary.
2. Traditional time limits are not necessary. Students can be given as much time as they need to finish a given test because no human proctor needs to wait around for them to finish the test.

1.3. Computer-Adaptive Language Testing

Computer-adaptive language tests are a subtype of computer-assisted language tests because they are administered at computer terminals or on personal computers. The computer-adaptive subtype of computer-assisted tests has three additional characteristics: (a) the test items are selected and fitted to the individual students involved, (b) the test is ended when the student's ability level is located, and, as a consequence, (c) computer-adaptive tests are usually relatively short in terms of the number of items involved and the time needed (Madsen, 1991; Wainer, 1990).

This flexi-level strategy eliminates the need (usually present in traditional fixed-length paper-and-pencil tests) for students to answer numerous questions that are too difficult or too easy for them. In fact, in a CALT, all students take tests that are suitable to their own particular ability levels-tests that may be very different for each student (Larson and Madsen, 1985).

1.4. Advantages and limitations

Brown (1992b) and others looked in more detail at both the advantages and limitations of using computers in language testing.

Advantages
The advantages of using computers in language testing are seen both from the angle of testing methodology and human considerations:
- Computers are much more accurate at scoring selected-response tests at reporting scores than human beings are.
- Computers can give immediate feedback.
- They allow testers to target the specific ability levels of individual students and can therefore provide more precise estimates of those abilities (Bock and Mislevy, 1982).
- Diagnostic feedback can be provided very quickly to each student.
- The use of computers allows students to work at their own pace.
- Computer mediated tests generally take less time to finish than traditional paper-and-pencil tests.
- Students experience less frustration than on paper-and-pencil tests because they are working on test items that are appropriate for their own ability levels.
- The testing procedure is less overwhelming (as compared to equivalent paper-and-pencil tests) because the questions are presented one at a time on the screen unlike in an intimidating test booklet with hundreds of test items.
- Many students like computers and even enjoy the testing process (Stevenson and Gross, 1991).

Limitations
The disadvantages of using computers in language testing result from physical and performance considerations.

- Computer equipment may not always be available, or be in reliable working order.
- Screen size limitations could be a problem, for example, for developing a reading test based on relatively long passages.
- The graphics capabilities of many computers (especially older ones) may be limited.
- Differences in the degree to which students are familiar with using computers or typewriter keyboards may lead to discrepancies in their performances (Henning, 1991).
- Computer anxiety is another potential disadvantage (Henning, 1991).

Computer-adaptive testing, however is only one stream of computer-related issues in education, psychology and related fields. Important developments are under way in areas like: testing in intelligent teaching systems, analysis and scoring of open-ended responses (compositions and speech samples), testing using the Web, and others.

2. WEB-BASED LANGUAGE TESTING

A web-based test (WBT) is an assessment instrument that is written in the "language" of the web, HTML. The test itself consists of one or several HTML file(s) located on the tester's computer, the server, and downloaded to the test taker's computer, the client. Downloading can occur for the entire test at once, or item by item. The client computer makes use of web-browser software (such as *Netscape Navigator* or *Microsoft Internet Explorer*) to interpret and display the downloaded HTML data. Test takers respond to items on their (client) computers and may send their responses back to the server or their responses to items may be scored clientside by means of a scoring script. A script can provide immediate feedback, adapt item selection to the test taker's needs, or compute a score to be displayed after completion of the test. The same evaluation process can take place on the server by means of serverside programs.

Many different kinds of WBTs are possible, depending on the developer's budget and programming expertise, as well as computer equipment available to test takers. On the low end of the technological sophistication are tests that run completely clientside and use the server only for retrieving items and storing responses. This type of test is the easiest to build and maintain because it does not require the tester to engage in serverside programming. In a low-tech WBT, the server only holds the test or the item pool. Test-taker responses are either scored clientside or sent to the tester's email box and stored for later downloading. This low-tech approach is preferable if limited amounts of test data can be expected, adaptivity is crude or unnecessary, item pools are small, and testers are interested in remaining independent of computer and software professionals.

A high-tech WBT, on the other hand, makes heavy use of the server, requiring testers to become highly familiar with the relevant software or involve

computer specialists in test setup and maintenance. This high-tech approach is preferable in cases where large amounts of test data have to be handled, item banks are large, and budgets allow for the purchase of expensive software and the hiring of computer professionals.

In this paper, we will focus on the low-tech versions of Web-based tests, which give testers maximum control over test design, require very small operating budgets, and make the advantages of computer-based testing available to testers at many institutions.

2.1. Item Types in WBTs

The Web is not automatically more suited for the testing of general second language competence or subject-specific second language performance than are other testing mediums. To the extent that the performance to be tested involves the Web itself (e.g., writing email, filling in forms), performance testing on the Web is highly authentic and very easy to do since testers only have to create an online environment that resembles the target one (Norris et al., 1998). It is useful, however, to consider which item types are more and which ones are less appropriate for Web-based testing.

It is fairly easy to implement:
- discrete-point grammar and vocabulary tests
- multiple choice items
- cloze tests and C-tests with text-fields for brief-response items
- discourse completion tests or
- essays with large text areas, as well as
- reading comprehension tests with frames, where one frame displays the text and the other frame displays multiple-choice or brief-response questions.

The most serious drawback of WBTs in terms of item types is that, at this time, there is no easy way to record test-taker speech.

2.2. Advantages and limitations

Advantages
- Probably the biggest logistical advantage of a WBT is its flexibility in time and space. All that is required to take a WBT is a computer with a Web browser and an Internet connection (or the test on disk). Test takers can take the WBT whenever and wherever it is convenient, and test designers can share their test with colleagues all over the world and receive feedback.
- The use of scoring scripts for dichotomously-scored items can make the test completely independent of the tester and increases flexibility and convenience for test takers even further (Gruba, 2000). Self-scoring is an attractive approach because it can save a great deal of tedious scoring work.
- If the test items are dichotomous, they can be scored automatically with a scoring script.
- Whereas producing traditional CBTs requires a

high degree of programming expertise and the use of specially-designed and non-portable delivery platforms, WBTs are comparatively easy to write and require only a free, standard browser for their display. In fact, anybody with a computer and an introductory HTML handbook can write a WBT without too much effort, and anybody with a computer and a browser can take the test. Of course, just because it is easy to write WBTs does not mean that it is easy to write *good* WBTs. Pretty pictures and animated images do not define test quality, and any test design and implementation must follow sound procedures (Alderson, Clapham and Wall, 1995) and include careful validation.
- A WBT is very inexpensive for all parties concerned. Testers can write the test by hand or with a free editor program without incurring any production costs except the time it takes to write the test.
For the test taker, the only expenses incurred are phone charges and charges for online time.

Limitations
- The greatest limitation of WBTs is their lack of security with respect to cheating and item confidentiality. Obviously, any test that test takers can take without supervision is susceptible to cheating. It is impossible to ensure that nobody but the test taker is present at the testing session, or that it is even the test taker who is answering the test questions.
- Item confidentiality is also impossible to maintain, since test takers are not taking the test under controlled conditions, that is, they could just copy items off the screen. Also, items are downloaded into the web browser's cache on the test taker's computer, which means that they are temporarily stored on the test taker's hard drive, where they can be accessed.
- There is a potential problem associated with the scoring approach: The script contains all the answers. In other words, the answers to all items are downloaded on the test taker's computer where a techno-savvy test taker can easily view them by looking at the test's source code. This can be made a bit more difficult by not integrating the script in the HTML code but instead embedding it as a separate script file.

3. CONCLUSION

Without a doubt, we are in the middle of a monumental technological paridigm shift, one which will eventually change the way instructors teach and the students learn.

The electronic tools must be used to design instructional and assessment models that strengthen students' language awareness, enhance their language competences and develop their self evaluating mental abilities.

It should be clear, though, that neither the computer assisted language testing schemes nor the

Web themselves do not make a good language test without sophisticated expert knowledge of test writing and validation provided. The Web, as a newly emerged instrument of language testing greatly expands the availability of computer-based testing with all its advantages and will undoubtedly become a major medium of test delivery in the future.

Through all our computer-related language teaching and testing efforts, however, quality and reliability considerations of meeting the standards must be of primary importance.

REFERENCES

Alderson, J. C.-Clapham, C.-Wall, D. (1995): Language test construction and evaluation. New York: Cambridge University Press

Bock, R. D.-Mislevy, R. J. (1982): Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6. 431-444.

Brown, J. D. (1992a): Technology and language education in the twenty-first century: Media, message, and method. Language Laboratory, 29. 1-22.

Brown, J. D. (1992b): Using computers in language testing. Cross Currents, 19. 92-99.

Gruba, P. (2000): The role of digital video media in response to task demands. Paper presented at the 22nd Annual Language Testing Research Colloquium, Vancouver, BC, Canada

Henning, G. (1991): Validating an item bank in a computer-assisted or computer-adaptive test. In P. Dunkel (Ed.), Computer-assisted language learning and testing: Research issues and practice (209-222). New York: Newbury House

Larson, J. W.-Madsen, H. S. (1985): Computerized adaptive language testing: Moving beyond computer-assisted testing. CALICO Journal, 2. 32-36., 43.

Madsen, H. S. (1991): Computer-adaptive testing of listening and reading comprehension. In P. Dunkel (Ed.), Computer-assisted language learning and testing: Research issues and practice (237-257). New York: Newbury House

Norris, J. M.-Hudson, T.-Brown, J. D.-Yoshioka, J. (1998): Designing second language performance assessments. Honolulu: University of Hawai'i at Manoa, Second Language Teaching and Curriculum Center

Stevenson, J.-Gross, S. (1991): Use of a computerized adaptive testing model for ESOL/bilingual entry/exit decision making. In P. Dunkel (Ed.), Computer-assisted language learning and testing: Research issues and practice (223-235). New York: Newbury House

Wainer, H. (1990): Introduction and history. In H. Wainer (Ed.), Computerized adaptive testing: a primer (1-22). Hillsdale, NJ: Lawrence Earlbaum