Language Learning & Technology May 2001, Vol. 5, Num. 2
http://llt.msu.edu/vol5num2/roever/default.html pp. 84-94

WEB-BASED LANGUAGE TESTING

Carsten Roever
University of Hawai'i at Manoa

ABSTRACT

This article describes what a Web-based language test (WBT) is, how WBTs differ from traditional computer-based tests, and what uses WBTs have in language testing. After a brief review of computer-based testing, WBTs are defined and categorized as low-tech or high tech. Since low-tech tests are the more feasible, they will constitute the focus of this paper. Next, item types for low-tech WBTs are described, and validation concerns that are specific to WBTs are discussed. After a brief overview of the marriage of computer-adaptive and Web-based tests, the general advantages as well as design and implementation issues of WBTs are considered before examining the role that testing consequences play in deciding whether a WBT is an appropriate assessment instrument. It is argued that WBTs are most appropriate in low-stakes testing situations; but with proper supervision, they can also be used in medium-stakes situations although they are not generally recommended for high-stakes situations. Some possible areas for future research are suggested.

INTRODUCTION

Interest in Web-based testing is growing in the language testing community, as was obvious at recent LTRC conferences, where it was the topic of a symposium on the DIALANG project (Alderson, 2001), a paper (Roever, 2000), several in-progress reports (Malone, Carpenter, Winke, Kenyon, 2001; Sawaki, 2001; Wang et al., 2000), and poster sessions (Carr, Green, Vongpumivitch, & Xi, 2001; Bachman et al., 2000). Web-based testing is also considered in Douglas's recent book (Douglas, 2000). It is the focus of research projects at UCLA and the University of Hawai'i at Manoa, and a number of online tests for various purposes are available at this time and are listed on Glenn Fulcher's Resources in Language Testing Web site (Fulcher, 2001). This paper is intended to advance the Web-based language testing movement by outlining some of the fundamental theoretical and practical questions associated with its development. Simply defined, a Web-based language test (WBT) is a computer-based language test which is delivered via the World Wide Web (WWW). WBTs share many characteristics of more traditional computer-based tests (CBTs), but using the Web as their delivery medium adds specific advantages while their delivery medium complicates matters.

COMPUTER-BASED AND WEB-BASED TESTS

The pre-cursor to Web-based language tests (WBTs) are computer-based tests (CBTs; for a recent discussion see Drasgow & Olson-Buchanan, 1999), delivered on an individual computer or a closed network. CBTs have been used in second language testing since the early 80s (Brown, 1997), although the use of computers in testing goes back a decade (Chalhoub-Deville & Deville, 1999). Computers as a testing medium attracted the attention of psychometricians because they allow the application of item response theory for delivering adaptive tests (Wainer, 1990), which can often pinpoint a test taker's ability level faster and with greater precision than paper-and-pencil tests. Based on the test taker's responses, the computer selects items of appropriate difficulty thereby avoiding delivering items that are too difficult or too easy for a test taker, but instead selects more items at the test taker's level of ability than a non-adaptive test could include. But even for non-adaptive testing, computers as the testing medium feature significant advantages. CBTs can be offered at any time unlike mass paper-and-pencil administrations

which are constrained by logistical considerations. In addition, CBTs consisting of dichotomously-scored items can provide feedback on the test results immediately upon completion of the test. They can also provide immediate feedback on each test taker's responses -- a characteristic that is very useful for pedagogical purposes. The seamless integration of media enhances the testing process itself, and the tracing of a test taker's every move can provide valuable information about testing processes as part of overall test validation.

On the negative side, problems with CBTs include the introduction of construct-irrelevant variance due to test takers' differing familiarity with computers (Kirsch, Jamieson, Taylor, & Eignor, 1998), the high cost of establishing new testing centers, and the possibility of sudden and inexplicable computer breakdowns.

Types of WBTs

A WBT is an assessment instrument that is written in the "language" of the web, HTML. The test itself is 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 as FORM data, or their responses to dichotomously scored items may be scored clientside by means of a scoring script written in JavaScript. 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 continuum of 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, which tends to involve complex code writing and requires close cooperation with server administrators. In a low-tech WBT, the server only holds the test or the item pool while the selection of the next test item is accomplished by means of a script located clientside. 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, for example, by having the server handle item selection through adaptive algorithms or by placing a database program on the server to collect and analyze test-taker responses. Both tasks require 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, complex adaptive algorithms are used, item banks are large, and budgets allow for the purchase of expensive software and the hiring of computer professionals.

In this paper, I 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.

What to Test on the Web and How to Test It

The first step in any language testing effort is a definition of the construct for what is to be tested. Will the test results allow inferences about aspects of students' overall second language competence in speaking, reading, listening, and writing (Bachman, 1990; Bachman & Palmer, 1996). Or will the test directly examine their performance on second language tasks from a pre-defined domain (McNamara, 1996;

Norris, Hudson, Brown, & Yoshioka, 1998; Shohamy, 1992, 1995), such as leaving a message for a business partner, writing an abstract, or giving a closing argument in a courtroom.

Whether a test focuses on aspects of second language competence or performance, its construct validity is the overriding concern in its development and validation. To that end, the test developer must be able to detect sources of construct irrelevant variance, assess whether the construct is adequately represented, in addition to considering the test's relevance, value implications, and social consequences (Messick, 1989). Also, they must examine the test‘s reliability, authenticity, interactiveness, impact, and practicality (Bachman & Palmer, 1996).

In the following section, appropriate content and item types for WBTs will be discussed and some WBT-specific validation challenges briefly described.

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. However, a WBT or any computer-based test can never truly simulate situations like "dinner at the swanky Italian bistro" (Norris et al., 1998, pp. 110-112). Rather than analyzing the possibilities of Web-based testing primarily along the lines of the competence-performance distinction, it is more useful 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 using radio buttons to create multiple choice items, cloze tests and C-tests with textfields 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. If the test items are dichotomous, they can be scored automatically with a scoring script. Such items can be contextualized with images (but see Gruba, 2000, for some caveats). They can also include sound and video files, although the latter are problematic: These files are often rather large, which can lead to unacceptably long download times, and they require an external player, a plug-in, which is beyond the tester's control. This plug-in allows test takers to play a soundfile repeatedly simply by clicking the plug-in's "Play" button.

Probably 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. Microphones are of course available for computers with soundcards, but recording and sending a sound file requires so much work on the part of the test taker that the error potential is unacceptably large.

Validation of WBTs

Quantitative and qualitative validation of WBTs does not differ in principle from validation of other types of tests. This is described in detail by Messick (1989) and Chapelle (1998, 1999). However, there are specific validity issues introduced by the testing medium that deserve attention in any WBT validation effort.

Computer familiarity. It is well established that test takers' varying familiarity with computers can influence their scores and introduce construct-irrelevant variance (Kirsch et al., 1998). Tutorials to increase computer familiarity can eliminate this effect (Taylor, Jamieson, Eignor, & Kirsch, 1998) and the use of standard web-browsers in WBTs increases the likelihood that test takers are already acquainted with the testing environment. For example, Roever (2001) found no significant correlation between self-assessments of Web browser familiarity and scores on a Web-based test of second language pragmatics

taken by 61 intermediate-level English as a second language (ESL) learners in the English Language Institute at the University of Hawai'i: Browser familiarity only accounted for 1%-3% of the variance in scores.

Typing speed. Differences in test takers' typing speed are potentially more serious sources of error variance and are not amenable to quick training. In oral debriefings, test takers in the Roever (2001) study complained about having too little time for the discourse completion section of the test, which required typing brief utterances and allowed 90 seconds per item. On average, test takers completed 83% of the brief response section, whereas they completed 99% of each of the test's two multiple-choice sections, in which they were allotted 60 seconds per item. Although a simple time increase for brief response items seems like an obvious option, the fact that no member of the native-speaker (NS) comparison group had the same problem, raises the question of whether and how typing speed and second language proficiency are related.

Delivery failures and speededness. One issue in the development phase of a Web-based test is to ensure that the test does not "skip" items during delivery due to technical problems. This can happen if the test taker accidentally double-clicks instead of single-clicking a button, or if there are errors in the algorithm that selects the next item in an adaptive or randomized test . It can be difficult to "tease apart" whether an item was not answered because the test taker ran out of time or because the computer did not deliver the item.

Loading time and timer. If the test is not delivered clientside but via the Web, download times can be negligible or considerable, depending on server traffic, complexity of the page, client computer speed, and a host of other factors beyond the test designer's control. It is therefore important for timed tests to stop the timer during downloads and restart it when the page is fully displayed.

A Special Case: CATs on the Web

Computer-adaptive tests are possible on the Web and do not pose many technical problems beyond those encountered in linear tests but it cannot be emphasized enough that the design of a sophisticated CAT is a very complex undertaking that requires considerable expertise in item response theory (IRT; for general introductions, see Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). Issues in designing and implementing CATs in second language assessment contexts have been discussed at length elsewhere (Chalhoub-Deville & Deville, 1999; Dunkel, 1999), so the following will only discuss issues specific to Web-adaptive tests (WATs).

Like general WBTs, CATs and WATs can be designed at various levels of sophistication. A very simple WAT could display sets of items of increasing difficulty and break off when a test-taker scores less than 50% on a set. The test-taker's ability would then roughly lie between the difficulty of the final and the preceding set. This is fairly easy to realize on the Web, since all that is required is a count of the number of correct responses. However, such a test does not save much time for high-ability test takers who would have to proceed through most difficulty levels. So instead of starting at the lowest difficulty level, initial items could be of mid-difficulty. Subsequent sets would be more or less difficult depending on a test taker's score until the 50% correctness criterion is met.

On the sophisticated end of CATs, complex algorithms re-compute ability estimates after every test taker response and select the best next item from a large item pool. Even these algorithms can run clientside, determine which item parameters are desirable for the next item, select an item from a list, and request that item from the server. This does not address the issue of item exposure, which is a major consideration in the item selection process since it potentially comprises test security: An overexposed item could be reconstructed by test takers after the test and communicated to others. However, this is hardly a concern for WATs, which are most appriopate for low-stakes situations (discussed later in this article). In the event that WAT used in a medium or high-stakes situation necessitates exposure control, the simplest way

of limiting exposure is by means of a randomization function, which selects an item from a pool of equivalent items with the same parameters (for a more complex approach, see Stocking & Lewis, 1995). However, this means that the item bank has to be quite large: Stocking (1994) recommends an item bank that is 12 times the test's length; Stahl and Lunz (1993) content themselves with 8-10 times.

WHY WBTs IF WE ALREADY HAVE CBTs?

Low-tech WBTs offer advantages over traditional CBTs with regard to their practicality (Bachman & Palmer, 1996), logistics, design, cost, and convenience.

"Anyplace, Anytime": The Asynchrony Principle

Probably the single 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.

An important caveat is called for here, which will be elaborated further in the section on stakes. In high-stakes situations, test takers stand to gain an advantage by cheating, if uncontrolled and unsecured access is not feasible. In such cases, monitored and supervised testing facilities must be used, where the degree of supervision and standardization of the physical environment again depends on the stakes involved. Even if high stakes are involved, there are still advantages to delivering the test via the Web, that is, no specialized software necessary, existing facilities like computer labs can be used as testing centers. However, just the convenience of "any place, any time" access no longer holds.

"Testing Goes Grassroots"

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 -- language testers do not have to be computer programmers to write a WBT. This is largely due to HTML's not being a true programming language but only a set of formatting commands, which instruct the client's Web browser how to display content. In addition, HTML contains elements that support the construction of common item types, such as radio buttons for multiple-choice items, input boxes for short response items, and text areas for extended response items (essays or dictations). Free or low-cost editing programs are available that further aid test design.

Of course, just because it is easy to write WBTs does not mean that it is easy to write good WBTs. retty pictures and animated images do not define test quality, and any test design and implementation must follow sound procedures (Alderson, Clapham & Wall, 1995) and include careful validation.

Testing Goes Affordable

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. Once a test is written, it can be uploaded to a server provided by the tester's institution or to one of many commercial servers that offer several megabytes of free web space (for example, www.geocities.com, www.tripod.com, www.fortunecity.com). Since WBTs tend to be small files of no more than a few kilobytes, space on a free server is usually more than sufficient for a test. The use of images, sound, or video can enlarge the test considerably, however, and may require the simultaneous use of several servers or the purchase of more space.

For the test taker, the only expenses incurred are phone charges and charges for online time, but since many phone companies in the US offer flat rates for unlimited local calls and many Internet service providers have similar flat rate plans for unlimited web access, test takers may not incur any extra costs for a testing session. However, the situation can be markedly different outside North America, where phone companies still charge by the minute for local calls. In such cases, a version of the test that can be completed entirely offline should be provided and distributed via email or download.

ISSUES AND LIMITATIONS OF USING WBTs

The following are some issues that should be considered during the conceptualization and the early stages of WBT development.

Cheating and Item Exposure

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. That limits the possible applications of unsupervised WBTs to low-stakes testing situations.

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. This is not a problem if items are created "on the fly" or if the item pool is constantly refreshed and each item is only used a few times.

Of course, cheating and item confidentiality are less relevant to low-stakes situations and can be prevented if the test is taken under supervision. This reduces the "anyplace, anytime" advantage of a Web-based test, but it may be a viable option for medium-stakes tests or tests taken only by few test takers, where the establishment of permanent testing centers would not be cost-effective and trustworthy supervisors can be found easily at appropriate facilities.

Self-Scoring Tests and Scripts

Using JavaScript to make tests self-scoring is an attractive approach because it can save a great deal of tedious scoring work, but there is a potential problem associated with this 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, but with a little searching, even that can be found in the browser cache. Solutions to this problem are supervision, scoring by the tester (e.g., by means of SPSS syntax), or serverside scoring scripts which would have to be written in Java, Perl, or serverside JavaScript.

Data Storage

Requirements for secure data storage differ by the type and purpose of the WBT. If the test is taken clientside only, for example, as a self-assessment instrument without any involvement of the tester, test-taker entries should be stored for the duration of the test so that a browser crash does not wipe out a test taker's work (and score) up to that point. However, as a security feature, Web browsers are generally prevented from writing to the test taker's hard disk. The only file to which they can write is a cookie file (cookie.txt on PC, cookie on Mac), and the main content that can be written to each individual cookie is one string of up to 2,000 characters (about two double-spaced pages). This may not be enough to save a long essay, but plenty to save numerical responses, short answers, and biodata. A problem here is that cookies as a means of data backup work only in *Microsoft Internet Explorer*, which updates the cookie

physically on the hard drive every time it is modified. *Netscape Navigator* holds the cookie in memory and only updates it when the browser window is closed, so that a system crash in the middle of a testing session irretrievably erases the cookie.

If the test involves the tester, that is, if test data are sent back to the tester's server, secure data storage is somewhat easier. The response to every item can be sent as a FORM email, so that a reconstruction of test taker responses is possible even after a browser or system crash. As an additional security feature to guard against server problems, sets of responses can be "harvested" by a JavaScript function and sent to a different server, so that in fact two or several records of each testing session exist.

From Test to Spreadsheet: Think Backwards

If complex serverside scripting or manual data entry of test-taker responses into a spreadsheet is to be avoided, the most convenient way of transferring responses is simply having test-taker responses to the entire test transferred at the same time in one final FORM email as a single long string. Testers then edit their email file (after saving it under a different name) so that it consists of nothing but those response strings (e.g., by making all the response strings bold and subsequently deleting everything that is not bold), which can be read into a spreadsheet as raw, unformatted text (ASCII data).

It is important to think backwards in this design process, that is, start out by considering the requirements and limitations that the spreadsheet or data analysis programs impose. For example, SPSS delimits data points by means of commas or spaces which means that they should also be thus delimited in the response strings, and that all other commas and spaces have to be eliminated. Scripts should be devised to check test test-taker input for commas and spaces and replace them, for example, the test taker entering "Doe, John" should become "DoeJohn."

Server Failure and Browser Incompatibility

A variety of technical problems is possible in Web-based testing, but the most significant ones are server failure and browser incompatibilities.

Server failure means that the server which houses the test is "down," so that test takers cannot access the test or continue a testing session where items are downloaded one by one. A simple way around this problem is to have "mirror sites" on alternate servers. Alternatively, all items can be downloaded at the beginning of the testing session as part of a script and can then be retrieved clientside.

A client-related problem that can be a minor or major bother is incompatibility of HTML or script features with the browser used clientside. The two major Web browsers, *Netscape Navigator* and *Microsoft Internet Explorer,* function similarly but not identically, so that the same test may work as desired on one but not the other. Even more importantly, different generations of browsers can be quite different in the kind of scripting that they can handle. The easiest way to tackle the compatibility problem is to ensure that all test takers have exactly the same browser and browser version. In that case, testers need to write and pilot the test only for that specific browser. If that is not possible, the next best solution is to offer a standard version of the test with scripting and an alternative, no-frills (no-scripts, no-frames) version that runs on any browser.

WEB-BASED TESTING OR NOT? THE CASE FOR A STAKES-DRIVEN DECISION

Whether Web-based testing is appropriate for a given assessment purpose depends largely on the consequences of the test. Generally speaking, the lower the stakes involved, the more appropriate a WBT.

Low-Stakes Assessment

WBTs are particularly appropriate for any assessment in the service of learning, where assessment serves to give learners feedback on their performance and provides them with a gauge of how close they are to

reaching a pre-specified learning goal (for an overview of the beneficial effects of assessment on learning, cf. Dempster, 1997). Such assessment can accompany classroom instruction or it can be a component of a web-based instruction system or a test-preparation system. Learners have no or little incentive to cheat on this type of assessment instrument since cheating would not be in their best interest.

A second highly appropriate use of low-tech WBTs is for second language research and specifically, research on language tests. The great flexibility of WBTs lets research participants work on the instrument wherever and whenever is convenient for them, and test developers can use scripts to record participants' every move precisely, thereby gathering valuable information on item characteristics and appropriate degree of speededness.

Finally, self-scoring instruments on the Web can be used for test preparation, either for large standardized tests or as pre-placement tests for students preparing to enroll in a foreign language program or a university in a foreign country. Such pre-placement will give test takers a general notion about how the students will perform on the test in question so that they can decide whether additional preparation is needed.

Using a WBT for low-stakes assessment preserves all the Web advantages of this test type: Test takers can take the test in the privacy of their own homes, at a time of their choice, and at their own pace. Costs for designing and maintaining the test on the Web are low to non-existent.

Medium-Stakes Assessment

Assessment situations with medium stakes include placement tests for foreign students, midterm, or final exams in classes, and other assessment situations which affect learners' lives but do not have broad, life-altering consequences. In these testing situations, test takers have an incentive to cheat, so unsupervised use of WBTs is not indicated. The test has to be administered at a trustworthy testing site, for example, in the case of a placement test for an English Language Institute at a US university, the test can be administered in the university's own computer lab under supervision of a lab monitor. Even more conveniently for students and testers, the test can be administered at a trusted remote site (e.g., another university's computer lab) before students even enter the program, thereby allowing them to register for courses in advance, and giving administrators an early overview of how many courses will be needed.

Another situation with medium stakes involves assessment for course credit. Distance education courses and classes taught via the Internet spring to mind because Web-based assessment will allow geographically dispersed test takers to take the test at a site near them.

Supervised testing reduces the "anytime, anyplace" advantage of Web-based testing, but the information value of early placement for all stakeholders may often balance this loss.

High-Stakes Assessment

High-stakes assessment is any assessment whose outcome has life-changing implications for the test taker. Admission tests for universities or other professional programs, certification exams, or citizenship tests are all high-stakes assessment situations. Obviously, such assessment requires tight security, standardized-testing environments, and the most precise and informative testing methods. Even high-stakes assessment instruments can be realized as WBTs, and such an approach can greatly increase test availability and reduce testing expenses for testers and test takers. But these situations clearly require involvement of computer experts to make test delivery glitch-free and keep the item pool hacker-proof. Generally, at this time, the author would not recommend using the Web for high-stakes testing, which is better done on closed and secure intranets.

THE FUTURE OF WEB-BASED LANGUAGE TESTING

It may seem premature to talk about the future when Web-based language testing is only now beginning to emerge as an approach to testing. However, some central issues that will have to be dealt with can already be identified:

- validation procedures for different types of media use, different types of delivery platforms, and the equivalency of test-taking in different environments,
- the potential, limits, and most appropriate uses of low-tech WBTs and high-tech WBTs,
- oral testing over the web, as real-time one-on-one voice chat or computer-generated speech,
- the possibilities of virtual reality for near-perfect task authenticity and performance-based testing.

It should be abundantly clear that the Web itself does not a good test make, no matter how flashy the Web page, how sophisticated the script, or how beautiful the animations. But the Web 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.

ABOUT THE AUTHOR

Carsten Roever is a doctoral student in the Ph.D. program in Second Language Acquisition at the University of Hawai`i at Manoa. His research interests include second language assessment, interlanguage pragmatics, and instructed SLA. He has taught several workshops and given conference presentations on Web-based language testing.

E-mail: roever@hawaii.edu

REFERENCES

Alderson, C. (Organizer). (2001, March). *Learning-centred assessment using information technology*. Symposium conducted at the 23rd Annual Language Testing Reserach Colloquium, St. Louis, MO.

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

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford, UK: Oxford University Press.

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

Bachman, L. F., Carr, N., Kamei, G., Kim, M., Llosa, L., Sawaki, Y., Shin, S., Sohn, S-O., Vongpumivitch, V., Wang, L., Xi, X., & Yessis, D. (2000, March). *Developing a web-based language placement examination system.* Poster session presented at the 22nd Annual Language Testing Research Colloquium, Vancouver, BC, Canada.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 44-59. Retrieved April 1, 2001 from the World Wide Web: http://llt.msu.edu/vol1num1/brown/default.html.

Carr, N., Green, B., Vongpumivitch, V., & Xi, X. (2001, March). *Development and initial validation of a Web-based ESL placement test*. Poster session presented at the 23rd Annual Language Testing Reserach Colloquium, St. Louis, MO.

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

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). New York: Cambridge University Press.

Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19,* 254-272.

Dempster, F. N. (1997). Using tests to promote classroom learning. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 332-346). Westport, CT: Greenwood Press.

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

Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment.* Mahwah, NJ: Lawrence Earlbaum Associates.

Dunkel, P. A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology, 2*(2), 77-93. Retrieved April 1, 2001 from the World Wide Web: http://llt.msu.edu/vol2num2/article4/index.html.

Fulcher, G. (2001). *Resources in language testing page.* Retrieved April 1, 2001 from the World Wide Web: http://www.surrey.ac.uk/ELI/ltr.html

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

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Niijhoff.

Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees.* (TOEFL Research Report No. 59). Princeton, NJ: Educational Testing Service.

Malone, M., Carpenter, H., Winke, P., & Kenyon, D. (2001, March). *Development of a Web-based listening and reading test for less commonly taught languages.* Work in progress session presented at the 23rd Annual Language Testing Reserach Colloquium, St. Louis, MO.

McNamara, T. (1996). *Measuring second language performance.* London: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

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.

Roever, C. (2000, March). *Web-based language testing: opportunities and challenges.* Paper presented at the 22nd Annual Language Testing Research Colloquium, Vancouver, BC, Canada.

Roever, C. (2001). *A Web-based test of interlanguage pragmatic knowledge: Implicatures, speech acts, and routines.* Unpublished manuscript, University of Hawai'i at Manoa.

Sawaki, Y. (2001, March). *How examinees take conventional versus web-based Japanese reading tests.* Work in progress session presented at the 23rd Annual Language Testing Reserach Colloquium, St. Louis, MO.

Shohamy, E. (1992). Beyond performance testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal 76*(4), 513-521.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics, 15*, 188-211.

Stahl, J. A., & Lunz, M. E. (1993, April). *Assessing the extent of overlap of items among computerized adaptive tests.* Paper presented at the annual meeting of National Council for Measurement in Education, Atlanta, GA.

Stocking, M. L. (1994). *An alternative method for scoring adaptive tests* (Research Report #94-48). Princeton, NJ: ETS.

Stocking, M. L., & Lewis, C. (1995). *Controlling item exposure conditional on ability in computerized adaptive testing* (Research Report #95-24). Princeton, NJ: ETS.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (TOEFL Research Report No. 61). Princeton, NJ: ETS.

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

Wang, L., Bachman, L. F., Carr, N., Kamei, G., Kim, M., Llosa, L., Sawaki, Y., Shin, S., Vongpumivitch, V., Xi, X., Yessis, D. (2000, March). *A cognitive-psychometric approach to construct validation of Web-based language assessment.* Work-in-progress report presented at the 22nd Annual Language Testing Research Colloquium, Vancouver, BC, Canada.