Language Learning & Technology
http://llt.msu.edu/vol2num2/article4/

January 1999, Volume 2, Number 2
pp. 77-93 online
(page numbers in PDF differ and should not be used for reference)

CONSIDERATIONS IN DEVELOPING OR USING SECOND/FOREIGN LANGUAGE PROFICIENCY COMPUTER-ADAPTIVE TESTS

Patricia A. Dunkel
Georgia State University

ABSTRACT

This article describes what a computer-adaptive test (CAT) is, examines its roots, and points out some of the challenges this innovative approach to assessment presents. A number of issues involving CATs are then discussed in terms of (a) the basic principles of assessment embodied in the CAT, (b) the special psychometric and technical issues peculiar to the CAT as opposed to traditional or paper-and-pencil tests, (c) the hardware and software used in the CAT, and (d) the administration of the CAT. Each of these issues is discussed in terms of the questions CAT score users should ask (and have answered), and developers must consider when designing L2 CATs. The questions should highlight the need for published CAT tables of specifications, models, and blueprints; should help design evaluative criteria for assessing the reliability, validity, and utility of L2 CATs; may help developers fashion more valid, reliable, and authentic L2 CATs; and might familiarize CAT score users with some the complexities involved in interpreting CAT scores, as opposed to more traditional paper-and-pencil scores.

INTRODUCTION

Today, powerful microcomputers are not only beginning to affect a redesign of the structure and content of school curricula and the entire process of instruction and learning,[1] but they are also having a decided impact on the types of tests created and used to assess that learning. In fact, computerized testing is increasingly being viewed as a practical alternative to paper-and-pencil testing (Kingsbury & Houser, 1993). Tests administered at computer terminals, or on personal computers, are known as computerized tests. Given the advantages of individual, time-independent language testing, computer-based testing will no doubt prove to be a positive development in assessment practice (Brown, 1997, p. 46).

Proponents of computer-based testing (CBT) in general, and computer-adaptive testing (CAT)[2] in particular, view the increasing use of computers in testing to be a natural evolution in assessment practice. Reckase (1989) points out that standardized adaptive tests have been available since the early 1900s. However, it was in the 1970s that pioneer Frederick Lord (1970, 1980) succeeded in working out the theoretical structure, as well as the practical technicalities, of creating mass-administered tailored tests using the computer. Today, numerous L2 researchers, practitioners, and experts in testing organizations are actively engaged in designing, developing, and using computerized placement and achievement tests (Carlson, 1994; Chaloub-Deville, 1996; de Jong, 1996; Dunkel, 1996, 1997; Eignor, Way, Stocking, & Steffen, 1963; Henning, Johnson, Boutin, & Rice, 1994; Johnston, 1996; Kaya-Carton, Carton, & Dandonoli, 1991; Larson, 1996; Larson & Madsen, 1985; Linacre, 1996; Luecht, 1996; Madsen, 1991; Nagata, 1998; Stevenson & Gross, 1991; Taylor, Jamieson, Eignor, & Kirsch, 1998; Weiss, 1996; West, 1995; Yao, 1995; Yoes, 1996).

Although support for and use of computerized testing is gaining momentum, the more dispassionate supporters, as well as the dubious skeptics, voice concern about the trend toward greater use of computers in the assessment process. Some are concerned about the appropriacy of CBTs for assessing particular skills such as reading comprehension (Bernhardt, 1996); others are worried about the fidelity and comprehensiveness of computerized tests (McNamara, 1996). Still others are concerned about the degree

to which construct-irrelevant (or nuisance-ability) variables, such as computer-familiarity or computer anxiety, might be injected into the assessment process to impact examinee performance in negative ways.[3]

Nevertheless, it seems very likely that computerized testing will continue to be more frequently used in assessment circles in the coming decades. In fact, CAT, a subcategory of computerized testing that involves the dynamic selection of test items to match the actual performance of an examinee during the taking of a test, has finally become a readily accessible methodology for use in standardized testing programs (Reckase, 1989, p. 11). Large-scale testing programs (e.g., the Graduate Record Examination and the Scholastic Aptitude Test) now use adaptive testing *pro forma* in the U.S., and numerous CATs of L2 learning loom large on the horizon.[4]

Commercial software systems for developing adaptive tests (e.g., those created by the IBM-based Assessment Systems Corporation of St. Paul, Minnesota, and the Macintosh-based Computer-Adaptive Technologies Corporation of Chicago, Illinois) now make it feasible to create CATs for L2 language assessment without the CAT developer having to start completely *de novo*. It is estimated that a significant number of the standardized tests that students take in the future will, in fact, be in CBT[5] or CAT formats. Today, the question no longer seems to be, "Should we use or create a CBT or a CAT?" but rather, "What do we need to know about computer-based or computer-adaptive testing to design or use such tests?" In other words, we need to ask, "What are the issues CAT developers need to consider, and what questions do they need to ask and answer before or while trying to construct a valid and reliable CAT?"

In a continuation of the discussion opened by Brown (1997) concerning present and future research on computers in language testing, the author of this paper examines a number of basic assessment, psychometric, technical, and administrative issues that need to be taken into account when an L2 CAT is designed or used, particularly for L2 placement or achievement purposes. The author mentions a number of design issues raised by Brown (1997) on a computer-adaptive language test (e.g., how to make decisions about cut-points), but she also highlights issues not addressed by him (e.g., validity concerns and issues related to Item Response Theory). Although all the factors considered in the present article can affect in positive and negative ways the effectiveness, appropriacy, and equitableness of the L2 CAT, the author does not purport to highlight all the major issues related to CAT design or evaluation. Rather, she poses a number of specific questions that developers need to address, and CAT score users and examinees need to understand and carefully consider, when CAT becomes a heuristic of assessment.

The answers to the questions posed in the present discussion are neither simple nor singular, but solutions must nevertheless be found if examinees and score users are to benefit from this inventive, efficient approach to L2 assessment.

COMPUTER-ADAPTIVE TESTING AND L2 CAT: AN INNOVATIVE APPROACH TO ASSESSMENT

L2 CAT is a technologically advanced method of assessment in which the computer elects and presents test items to examinees according to the estimated level of the examinees, language ability. The basic notion of an adaptive test is to mimic automatically what a wise examiner would normally do. Specifically, if an examiner asked a question that turned out to be too difficult for the examinee, the next question asked would be considerably easier. This approach stems from the realization that we learn little about an individual's ability if we persist in asking questions that are far too difficult or far too easy for that person. We learn the most about an examinee,s ability when we accurately direct our questions at the same level as the examinee's ability (Wainer, 1990, p. 10).

Thus, in a CAT, each examinee's first item is usually of a projective medium-difficulty level for the test population. If the examinee responds correctly, the next item received is more difficult. If, however, the examinee misses the item, an easier question is presented in turn. And so it goes, with the computer

algorithm adjusting the selection of the items interactively to the successful or failed responses of the test taker.

In a CAT, each examinee takes a unique test that is tailored to his or her own ability level. Avoided are questions that have low information value about the test taker's proficiency. The result of this approach is higher precision across a wider range of ability levels (Carlson, 1994, p. 218). In fact, CAT was developed to eliminate the time-consuming and inefficient (and traditional) test that presents easy questions to high-ability persons and excessively difficult questions to low-ability testees. Other advantages of CAT include the following:

- ***Self-Pacing***. CAT allows test takers to work at their own pace. The speed of examinee responses could be used as additional information in assessing proficiency, if desired and warranted.[6]
- ***Challenge***. Test takers are challenged but not discouraged or bothered by the presentation of items that are far above or below their ability level.
- ***Immediate Feedback***. The test can be scored immediately, providing instantaneous feedback for the examinees.
- ***Improved Test Security***. The computer contains the entire item pool, rather than merely those specific items that will make up the examinee,s test. As a result, it is more difficult to artificially boost one's scores by merely learning a few items or even types of items (Wainer, 1990). However, in order to achieve improved security, the item pool must be sufficiently large to ensure that test items do not reappear with a frequency sufficient to allow examinees to memorize them. For a description of ways to ensure item bank depth, see Bergstrom and Stahl (1992) and Brown (1997).
- ***Multimedia Presentation***. Tests can include text, graphics, photographs, and even full-motion video clips, although multimedia CAT development is still in its infancy. For a more in-depth discussion of new technologies such as CD-ROM and interactive video that make it possible for students to interact with a computer during language testing, see Brown (1997).

Green and colleagues (1995) point out that a variety of agencies and groups could tap the power of computerized testing and benefit therefrom. These include:

- Educators considering the use of a published (or an in-house-developed) CAT to assess student achievement in individual, large-enrollment second/foreign language classrooms or programs
- Licensing boards needing to hire a private test developer or consulting firm to develop a CAT to help them select candidates who have met specific performance standards for licensure, such as the Occupational English Test (OET) developed on behalf of the Australian Government (McNamara, 1991), and the National Council of State Boards of Nursing (Stenson, Graves, Gardiner, & Dally, 1991)
- Agencies (e.g., Educational Testing Service) preparing users' guides for their computer-adaptive achievement tests, such as the Graduate Record Examination
- State departments of education wishing to develop a CAT version of statewide minimum competency tests
- Departments of modern foreign languages wanting to create a proficiency CAT for entrance into or exit from required language courses, such as the Ohio State University's Multimedia Computer-Adaptive Test (MultiCAT) of French, German, and Spanish

COMPUTER-ADAPTIVE TESTING: THE ROOTS AND CHALLENGES

In the 1960s and 1970s, the U.S. Department of Defense perceived the potential benefits of adaptive testing and supported extensive theoretical research in CAT and Item Response Theory (IRT),[7] the family of psychometric models underlying computer-adaptive testing (Wainer, 1990, p. 10). IRT is based on probabilistic theory; that is, it calculates the probability of a given person getting a particular item right (Alderson, Clapham, & Wall, 1995). Examinees, scores, and item total statistics are transformed into one scale so that they can be related to each other. If a person's ability is the same as the difficulty level of the item, that person has a 50-50 chance of getting that item right. If the ability is less, that probability decreases. The relationship between the examinees' item performance and the abilities underlying item performance is described in an item characteristic curve (ICC). As the level of students' ability increases, so does the probability of a correct response (see Alderson, Clapham, & Wall, 1995, p. 90).

Early attempts to build adaptive tests by the U.S. Army, Navy, and Air Force were often less than successful, very expensive, and used large-scale computers. However, by the early 1980s, personal computers had acquired the power of the large-scale computers of earlier years, and the pioneering efforts of IRT theorists such as Lord (1970) and others had perfected the psychometric model underlying CAT. In the late 1980s, CAT finally moved out of the realm of theory and supposition into the sphere of possibility and implementation with the advent of the College Board's CAT Graduate Record Examination, and with the work of in-house researchers in foreign language education at the Defense Language Institute and universities throughout the U.S., Britain, and the Netherlands, to name but a few of the countries.

Today, as a result of software development companies assisting developers with their own institutional L2 CATs, computer-adaptive testing has finally become a viable alternative to conventional paper-and-pencil testing. Commercial CAT programs such as those offered by the Assessment Systems Corporation (St. Paul, Minnesota) and Computer-Adaptive Technologies (Chicago, Illinois) make it easier for developers to create L2 CATs using software templates rather than having to start programming and development from scratch. It is anticipated that in the future more and more commercial companies and academic institutions will be producing testing shells that can be used to create CATs for placement, achievement,[8] and licensing purposes.

COMPUTER-ADAPTIVE TESTING: THE NEED FOR CLEAR GUIDELINES, SPECIFICATIONS, AND BLUEPRINTS

As mentioned previously, the issue to be decided in the future no longer seems to be whether to test examinees with a CAT, but how to best design, validate, and implement the CAT and how accurately to interpret CAT scores. Unfortunately, there are few available blueprints, models, or design beacons to guide L2 CAT developers, whether they are constructing the CAT with commercial testing-shell software or proceeding from scratch. Furthermore, there exist few published guidelines to help developers create L2 CATs, or to help users understand and evaluate the complexities, strengths, and weaknesses of any particular CAT. The need for such guidelines to aid the developer and the user alike is great, and will become greater as CAT becomes more prevalent in the coming century.

As a developer of an L2 listening comprehension CAT in the mid-1980s, the author of this article would have greatly benefited from having the opportunity to examine and review established L2 CAT guidelines and evaluative criteria. It would also have helped to be able to examine (a) tables of specifications for developed CATs, (b) models of other L2 CATs, (c) blueprints for created L2 CATs, and (d) a listing of the basic technical features that needed to be included in a CAT for any particular language skill (e.g., reading comprehension, listening comprehension, grammar, etc.). It would have greatly aided the L2 CAT development effort if I could have had this kind of information available when trying to create and

evaluate the quality and utility of available CAT software and the L2 CATs being developed at the time.[9] However, when work in listening comprehension CAT development started in 1987, few CAT (and ever fewer L2 listening CAT) guidelines were available for consultation. The situation has not changed appreciably in the past ten years although there is increased awareness of the need for tables of specifications for paper-and-pencil tests.[10] Although it may be understandable that CAT test developers need to keep confidential the tables of specifications for their tests, this confidentiality makes CAT development more difficult. At the very least, developers need to share information about general models and blueprints for targeted L2 tasks, item types, and item-bank contents. We need to establish general guidelines for CAT development if we are to create valid and useful L2 CATs specifically, and to improve their validity, utility, and usability in general. It is hoped that the issues discussed in this article will help CAT developers to create appropriate guidelines, specifications, and models that can be shared among professionals in the field.

As mentioned earlier, a CAT needed by a licensing board, a state department of education, or a department of modern foreign languages might be developed by the score user, or it might be created under contract to meet the specific purposes of a particular language program. It might be purchased as an existing CAT from a commercial vendor with or without online support, or the user may buy the test, and the testing company (e.g., Assessment Systems Corporation) may or may not provide support services to the score user. Each of these situations warrants special consideration by persons responsible for making decisions about the structure and ultimate use of the test, including developers of the CAT assessment programs (Green et al., 1995). Whether one is using a commercially developed CAT or an in-house product, these special considerations fall into four broad categories involving:

- The basic principles of assessment embodied in any well-designed test, for example, essential measurement qualities such as reliability and construct validity since they provide the major justification for using test scores as the basis for making inferences or decisions (Bachman & Palmer, 1996), as well as authenticity.[11]
- The special psychometric as well as technical issues peculiar to CAT as opposed to paper-and-pencil tests, for example, use of the appropriate IRT model, choice of the item-selection algorithm,[12] selection of the test-stopping rule,[13] and the field testing or trailing of the items to obtain statistical calibrations associated with each item.
- The basic issues of the hardware and software used in the CAT.
- Issues surrounding the administration of the CAT.

The remainder of this paper discusses each of these issues and raises related questions that developers and users alike need to carefully address when beginning the creation or use of an L2 CAT. It is hoped that the assembly of such questions will aid not only novice CAT developers, but also evaluators, in assessing the utility, validity, and appropriacy of the particular L2 CAT.

QUESTIONS AND RELATED ISSUES INVOLVING THE BASIC PRINCIPLES OF ASSESSMENT IN THE CAT

A number of questions need to be addressed when considering the basic principles of assessment in the CAT. They including the following:

Is the Computerized Testing System Appropriate for the Purpose of the Test?

L2 CAT developers need to clearly identify and specify the assessment purpose of the L2 CAT. This is important since, according to the American Council on Education (Green et al., 1995), CATs can be used for a wide variety of purposes. Each of the numerous and varied purposes available produces a specific type of CAT whose major function(s) might be one or more of the following:

- Identifying whether an individual has met the specific course objectives of a basic L2 language or literature course
- Indicating an individual's level of achievement in a skill domain (e.g., listening comprehension or knowledge of L2 grammar)
- Identifying specific areas in which a student needs additional educational experiences (e.g., knowledge and use of specific grammatical points or recognition of specific idioms and vocabulary items)
- Diagnosing an individual's skill-area strengths (e.g., the ability to recognize main ideas presented in a spoken mini-lecture) and weaknesses (e.g., inability to recall specific details from a short conversation about an academic topic)
- Detecting whether candidates have met minimum course requirements as demonstrated in a mastery test. In the case of a mastery CAT, when scores are compared to a certain set standard, the actual performance level of the examinee is often not as important as the relative location of the performance level to that standard set. For example, with mastery testing, a student's score is often evaluated against a minimum score (or cut-score)[14] to ascertain whether the examinee has or has not met the skill or content requirements. The particular cut-score may not be relevant if the student's score is far from the standard set. What is important is whether the examinee,s score is above or below the cut-score established. The accuracy of the assessment at the cut-point (the pass/fail standard set) is critical, albeit less precise for those scores that cluster around the cut-score. A less precise assessment might well be tolerated in certain testing situations for scores falling far above or below the cut-point. However, when scores fall slightly above or below that cut-point, the decision concerning mastery or non-mastery for these examinees can have tremendous positive or negatives consequences (Green et al., 1995, p. 2).

Although relatively few mastery CATs are used in L2 education, achievement or placement tests could serve as "mastery" CATs for L2 programs. However, it is necessary that users of the scores be able to read and understand (a) the stated purpose of the test as formulated by the developers, (b) the test tasks designed to assess the target skill, and (c) the level of mastery or the particular cut-score set required to make pass/fail decisions about examinees. In other words, users should be able to examine the particular criteria used to determine the achievement scores or placement levels of examinees.

Thus, it is important that end users of CAT scores be familiar with the exact purpose of the test, have access to a published Table of Specifications indicating the exact focus of the content and types of language tasks tapped by the items in the item pool, and understand the mastery-level scale with the particular level cut-points set for the CAT.

In addition to clearly stating the purposes of the test and identifying examinee-ability scale and cut-points, the American Council on Education (Green et al., 1995) states that the CAT must also be able to measure the examinees, true proficiency level. To achieve this goal, the L2 CAT must provide examinees with a sufficiently broad range of L2 content areas and skill-tasks to ensure that their true proficiency is indeed measured in the test items taken. Because examinees may be of high or how proficiency levels, the CAT must be designed in such a way as to provide adequate assessment for the entire range of ability, from low to high, as represented in the examinee population (Green et al., 1995, p. 2). This objective may most easily be accomplished by obtaining or designing a CAT that includes the entire range of ability in its item pool. For example, in the case of a general listening proficiency CAT, items in the pool must cover low to high listening ability levels. In addition, the items need to include a variety of listening tasks, such as comprehension of the main ideas of a conversation or mini-lecture, recognition and recall of details of a conversation, identification of specific words and phrases used in a passage, and so forth.

To achieve both objectives, the item selection algorithm must constrain the selection of items not just on the basis of the statistical parameter associated with the test item (such as the difficulty level), but it must also be able to present a variety of designated listening comprehension tasks to the examinees.

Does the CAT Embody the Basic Principles of Reliability?

Reliability refers to the precision and consistency of scores derived from the CAT instrument. It is a function of general, situational, and individual factors (Cohen, 1994) that can be used to frame evaluative questions for the developers of the test. The answers to these questions can, in turn, determine or establish the reliability of the CAT.

General Factors Influencing Reliability.

- How extensively does the CAT item pool sample the stated test objectives?
- Does the CAT lessen the ambiguity in interpreting the test items?
- Are the instructions for the examinees clear and explicit?
- Are the graphics and photos clear enough and of sufficient quality that examinees do not misinterpret their intent?
- Are the examinees sufficiently familiar with the format of the CAT before taking it? In other words, is there a practice session or tutorial that examinees can take before beginning the CAT?

Situational Factors Influencing Reliability. Is the testing environment suitable for computer-adaptive testing? In a listening CAT, for example, environmental noise should be kept to a minimum and headphones should be provided to examinees**.**

Individual Factors Influencing Reliability. These include transient factors (e.g., the physical and psychological health of test takers) and stable factors (e.g., their experience with similar tests).

- Is the tension as a result of taking the CAT reduced as much as possible?
- Does the orientation sufficiently prepare the examinees for the test?

In addition to the factors above, specific design features such as the "tailored" aspect of the CAT can influence and even threaten the reliability of the test in significant ways. The sources of error which can cause unreliability arise from item sampling in the adaptive test itself since each examinee takes a unique test. Concepts of reliability and measurement error are more complex in IRT-based CATs than in Classical-Item-Analysis-based paper-and-pencil tests in which error variance is assumed to be the same for all scores. In a CAT, scale scores may have a different degree of precision as indicated by the size of the errors of measurement at different levels of proficiency (Wainer, 1990, p. 165). Thus, measurement precision in an IRT system usually varies as a function of proficiency, and in any examination of the reliability of a CAT, it is necessary to determine the following:

CAT Item-Sampling Variation. This is equivalent to alternate form reliability in traditional paper-and-pencil testing. Has item-sampling variation been examined by inspecting how subjects do on alternate item pools from the CAT system? In other words, have examinees been tested twice, and have the correlations between the scores on the two administrations been compared?

CAT Stability over Time. This is equivalent to test-retest reliability in paper-and-pencil tests. Have examinees been tested with the tailored CAT items on two occasions, and have the correlations between the two scores been computed?

CAT versus a Paper-and-Pencil Version of the Comparisons. Although the paper-and-pencil version of an L2 test will in all likelihood be based on Classical Item Analysis as opposed to Item Response Theory, developers should compare the CAT at the test level with an older form of the paper-and-pencil test. In some cases, pencil-and-paper tests may not be available for comparison, especially as more and more CATs are developed without recourse to adapting paper-and- pencil models to the CAT environment.

Does the CAT Embody the Basic Principles of Validity?

Validity refers to whether the CAT actually measures what it purports to measure. It relates to the appropriacy of the inferences made on the basis of the test scores. Validity is often divided into content,

construct, criterion, concurrent, and predictive validity. CAT developers and users need to examine issues related to each of these types of validity, and to pose relevant questions vis-à-vis the validity and design of the CAT.

Content Validity. This relates to the adequacy with which the CAT samples the content of the subject matter. Questions to be asked by the evaluator or user of the CAT scores include the following:

- Was there a Table of Specifications used by the developers? Can it be perused?
- Can one review a listing of the content and the L2 model used in the test construction?
- Is there evidence that both the entire item pool and the items selected for each individual examinee represent the skill domain specified for that particular CAT (Wainer, 1990)?
- Is there evidence that a change in the mode of presentation of the items (e.g., from paper-and-pencil to CAT) has no effect per se on the item content and difficulty of the items?

Construct Validity. This refers to the degree to which test scores permit accurate inferences about the underlying traits being measured, and whether the CAT is an accurate operationalization of the underlying theory or trait being measured. Questions to pose include:

- Is there evidence presented that the item pool measures a single factor?
- Is the construct underlying the observed response used consistently? Wainer (1990) suggests that the pattern of correlations among tests hypothesized to measure related constructs be examined (e.g., by comparing patterns of responses on the TOEFL listening test with responses on a CAT, testing listening skills similar to those represented in the listening section of the TOEFL).

Criterion-Related Validity. This indicates how well the respondents, performance on specific sets of objectives on the test parallels their performance on another instrument or defined criterion measure of interest (Suen, 1990). Wainer (1990) suggests that criterion-related validity quantifies the relationship between a test used for selecting an educational program or occupation and subsequent real-world performance (p. 196). Criterion-related validity can thus provide evidence that the examinee who performs well in a testing situation should also do well in a similar real-world situation. For example, it could be assumed, though not always correctly, that the examinee who achieved a high score on an academic listening CAT would perform equally well in a university lecture hall after taking tests on the content of the lecture heard.

Concurrent Validity. This involves issues relating to how well test takers perform on the individual CAT and on another construct-valid CAT. Since to date so few CATs have been fully implemented, it is difficult to assess their concurrent validity. However, future development in L2 CAT should alter this situation appreciably and allow for the examination and demonstration of concurrent validity. For instance, the computer-based, and eventually computer-adaptive, TOEFL should allow developers of academic listening CATs to examine the concurrent validity of their instrument against the CBT/CAT TOEFL.

Predictive Validity. This relates to issues of how well CAT test scores can predict ensuing success on a trait, skill, or other ability of interest. Suen (1990) notes that the frequently made distinction between predictive and concurrent validity is moot. The only difference is in the exact time when the criterion measurement is made (pp. 140-141).

Consequential Validity. Citing issues raised by Hulin, Drasgow, and Parsons, Bachman (1990) argues that test developers and users must consider the potential consequences of testing, and must take into account the rights and interests of test takers, the institutions responsible for testing and making decisions based on those tests, the public interest, and the value systems that inform the particular test use. He further notes that since test takers, test developers, and test users all have value systems in any given testing situation, these systems may coincide completely, overlap to some degree, or be antithetical to each other. Those developing tests in general, and CATs in particular, need to be sensitive to a host of

issues involving the political, social, and ethical consequences regarding the interpretation and use of the scores. Test developers and users must ask and answer questions such as the following:

- How does the developer, in this case the CAT designer, balance the demands of reliability, validity, and practicality against the inevitable limitations on resources for test development (Bachman, 1990, p. 281)?
- To what extent does the test reflect the personal bias of the test user, and to what extent does this differ from that of the test taker (Bachman, 1990, p. 281)? This issue is particularly important when one considers that, at present, CATs are primarily being developed and used in technologically advanced countries by developers often interested in the *efficiency* of CATs for testing skills and content areas.
- How might the use of structured-response formats (such as multiple choice), as opposed to constructed responses, in L2 achievement tests lead to increased emphasis on memorization and analysis in teaching and learning at the expense of divergent production and synthesis (Messick, 1988)? This is particularly problematic for L2 CAT development since so many L2 CATs presently rely on structured response formats. Unfortunately, this article does not allow for a fuller examination of consequential validity issues, but can merely touch upon the problem to raise the consciousness of readers about this important kind of validity. For a lengthier discussion of this issue, the reader is referred to Bachman (1990) and Messick (1988).

IRT-Related Validity. In addition to threats to the above-mentioned types of validity, a major threat to CAT validity lies in the use of unidimensional Item Response Theory models, given the likely multidimensionality of language proficiency. Assessment of dimensionality is a central issue for CAT since non-multidimensional IRT is based on the assumption that items within tests measure the same dimension of individual differences. Hence, it is a matter of indifference which items are presented to the examinee (Brock, Givvons, & Muraki, cited in Wainer, 1990, p. 212). However, one must ask whether developers have shown evidence that factor-analytic procedures of inter-item correlations have been applied to assess whether unidiminsionality or multidimensionality actually exists in the CAT item pool.

QUESTIONS AND ISSUES SURROUNDING THE PSYCHOMETRIC AND TECHNICAL ISSUES PECULIAR TO CAT

Is the Psychometric Model Underlying the CAT Appropriate for the Construct Being Measured?

As mentioned earlier, a number of psychometric models of Item Response Theory (e.g., the one-, two-, and three-parameter models) can be used in the construction of CATs. Users and developers of L2 CATs should understand the assumptions underlying these models, and select the one that accords most effectively with the purpose of the assessment and the resources available for trialing (field testing) of the test items in the development process. For example, the unidimensional Rasch model serves the needs of many small-scale developers who have hundreds, rather than thousands, of subjects available for trialing.[15] However, it must be understood (especially by scores users) that use of this model may limit the kinds of interpretations that can be gleaned from the test scores. The Rasch model would not, for instance, be appropriate to measure a multidimensional construct such as listening proficiency unless that construct was broken down into unidimensional subdomains (e.g., comprehending the main ideas or details of minitalks, comprehending the details of a short conversation, etc.). It is debatable, of course, whether the construct of listening comprehension can be viewed in unidimensional terms and, ergo, that the underlying IRT model used in a listening proficiency CAT can be the Rasch model. However, this researcher has evidence that listening can be construed in unidimensional terms when the types of tasks and content of the items are controlled to ensure that unidimensionality is reflected in the items, or when the subdomains of the tasks and content are written to achieve unidimensionality.

If thousands of subjects are available for trialing of the target item bank, and if the construct being assessed is multidimensional or has distinct domains that cannot support use of a unifying dimension, then developers must choose the three-parameter IRT model as the underlying psychometric model of the CAT. In such an event, each of the domains in the CAT may need to be scaled separately. If, however, the sub-domains of CAT items are articulated for content coverage purposes, and yet a unifying construct can be supported, then the simpler Rasch model can be utilized. In this case developers can use factor analytic methods to investigate the dimensionality structure of the construct (see Dunkel, in press). If one factor emerges as dominant, then use of the Rasch model is warranted. Green, Bock, Humphreys, Linn, and Reckase (1984) comment on a Rasch model by Lord (1980) that, with as few as 150-200 cases, yielded more reliable parameters than the three-parameter model, even when the simulated data were known to follow the three-parameter model.

Until CAT becomes more common, and until developers fully understand the assumptions and requirements of Item Response Theory, a major concern for both developers and users of CAT will remain in the determination and selection of appropriate IRT models. According to Green et al. (1995), L2 test developers should identify the IRT model used to drive their CAT, and clarify for users (a) why the model chosen is appropriate for the stated purpose and structure of the CAT, (b) how test items fit the model after trialing, and (c) how items that do not fit the model are handled.

QUESTIONS AND ISSUES INVOLVING THE HARDWARE AND SOFTWARE USED IN THE CAT

Is the Equipment (e.g., the Computer Platform and the Testing Software) of Sufficient Quality and Adequacy to Ensure the Quality of the CAT?

Determinations involving both the visible and invisible features of the system and the software need to be made early in the CAT development project, and the congruence between the equipment and the CAT software needs to be made known to the users and examinees. According to Alessi and Trollip (1991), initial decisions need to be made concerning the type of delivery platform used, the capacity of the system needed, the type of display or monitor required, and the requirements of the software. The following are some initial questions that need to be addressed by the developer at the start of a test development project.

The Delivery Platform. Should a stand-alone or a networked system be used? A stand-alone system means that each station must be started separately, must store the data needed for a testing session, and at the end of a test, the supervisor must manually retrieve the data from each station. By contrast, a networked system can be started from the control panel, and the test administrator can retrieve the results more easily.

The Capacity of the System. Is there enough storage to include graphics and sound in the CAT? If not, why?

The System Display and Interface. Does the display allow for multimedia presentations? Are earphones available if listening comprehension is a target skill? Does the CAT software contain quality animation, graphics, and full-motion video? Does it have adequate input devices? Does the examinee, for example, type the response or touch the screen to give a response to a test item? Does the software contain menus for the examinee? Does it have adequate student control? Does it allow for record keeping, data analysis, and generation of new items for institutional purposes, if appropriate? Is the software secure and accessible to the test administrator? (See the discussion below concerning administration issues.)

QUESTIONS AND ISSUES INVOLVING ADMINISTRATION OF THE CAT

Do the Examinees Have an Opportunity to Become Familiar with the Computer, the CAT System, as well as the Structure, Organization, and Content Domains of the CAT?

Examinees should be given the time and opportunity to become thoroughly familiar with both the computer and the testing system. For first-time computer users, there should be an orientation to the functioning of the computer (e.g., using a mouse, calling for questions, answering questions, adjusting the audio volume, scrolling, etc.). An orientation to the structure and types of items they will encounter during the CAT should also be required for all examinees. Here the practice items should be equivalent in structure and content to those contained in the item bank. Some of the critical questions to be asked at this point include:

- Does the test taker understand what an adaptive test is, and does the software explain how a CAT operates?
- Do the test directions disclose the following information:
  - How to use the testing system?
  - What restrictions affect the administration of the test (e.g., How much time is allowed to respond? How many times can spoken input be played? What resources are permitted to the student during the test?)?
  - What content and skills will be tested?
  - Are the test takers able to review the instructions as many times as needed?
  - Are clear directions for leaving the testing system provided, including an explanation of what occurs if they do leave?
  - Are the examinees able to change their answers after making them?
  - If the software has varied presentation modes (e.g., graphics, videos, text heard, text read, multiple-choice items, selection-of-graphic items, etc.), are the examinees fully apprised of the various presentation modes and question formats?
  - Is the system easy for the examinees to use?
  - Does the system allow for adequate user control (e.g., self-pacing)?

Is the Item Pool of an Appropriate Quality to Support the Test Purpose(s) and to Measure the Identified Ability of the Examinee Population?

The depth and breadth of the item pool from which individual items are drawn strongly affects the validity and utility of the resulting CAT scores. As a result, in addition to ensuring that the items tap the variety of specific tasks and content areas pertinent to the identified purpose of the CAT, the developers and users of the scores need to be able to specify exactly what the items in the bank assess. For instance, in an academic listening proficiency CAT, the designers could specify that all examinees demonstrate the following: (a) comprehension of the main ideas of a mini-lecture; (b) comprehension of the details of a short dialog. They may also wish to set other specific skills for certain ability levels. For instance, advanced listeners should be able to understand the implied meaning of utterances.

Establishing a clear link between the focus and structure of the items and the purpose of the CAT is no easy task. However, it is necessary that a genuine correspondence exist between the stated goals of the CAT and the nature of inferences about examinee ability that are drawn from the test scores. These inferences are derived as a result of the types and quality of items passed or failed.[16] In addition, as mentioned previously, the pool of items needs to be sufficiently large so that overexposure of certain items does not occur. For a commercial CAT, for example, thousands of items may be needed to ensure that individual items are not repeatedly presented to examinees. If repeated presentation of certain items

occurs, test security is quickly breached. Such was the case with Educational Testing Service's Graduate Record Examination in the early 1990s.

Is the Stopping Rule Appropriate for the Purpose of the Test?

A CAT generally terminates when the computer algorithm for the stopping rule[17] indicates that the examinee's ability has been estimated within a specified level of precision, or when a certain number of items, or subsets of items, has been administered. In a *variable-length* adaptive test, the number of items administered to each examinee differs depending on the number of correct/incorrect responses given by him or her to the items presented. In a *fixed-length* computerized mastery (pass/fail) adaptive test, stopping occurs after a specified number of items are administered.[18]

The decision or stopping rule selected is critically important to the interpretation of the resulting scores (Green et al., 1995), and should be both practical and consistent with the intended uses of the scores. Green et al. note that when a test has a cut-score, it should be taken into account when deciding on the algorithm for terminating the test. Confidence intervals may then be used to ascertain whether the examinee's estimated proficiency is sufficiently similar to or deviant from the cut-score to specify the decision outcome (pass/fail). Thus, the importance of the pass/fail decision should also be taken into account when the stopping rule is implemented. The greater the importance of the decision based on the test score, the higher the test-score precision should be in the algorithm that specifies when to terminate the CAT administration.[19] An easy-to-understand discussion of the CAT's stopping rule should be included in the CAT documentation. If it is not included, the user should be able to request information on the stopping rule used.

In sum, test developers need to disclose the following information in the documentation:

- Type of CAT (fixed-length or variable-length)
- Stopping rule used to terminate the test
- Scoring algorithm
- Instances in which an examinee's proficiency level might not be correctly estimated

All these conditions require that score reporting to the users and administrators be as transparent as possible. Test developers often spend a great deal of time creating item banks and trialing the items, but few spend a concomitant amount of time devising report systems that are understandable and pertinent to individuals who are not experts in assessment.

Does the Test Security, Especially for High-Stakes Tests, Protect the Validity and Usability of the CAT Scores, as well as the Integrity of the Examinees' Score Records?

Breaches of test security can threaten the validity of the test. Only authorized persons should have access to the testing computers, and they must understand the importance of ensuring security. Security entails safe proctoring of the test as well as protecting of the item pool, the administrative systems, and examinee records. Green et al. (1995) state that depending on the purpose of the testing program, the examinees' session could be recorded on video- and/or audiotape to enhance security (p. 11). In addition to encrypting all examinee records, the system should allow only authorized persons to access the records, and should keep record of all persons who access the database. Features must also be written into the program to prevent computer hackers from retrieving items from the pool. Pilot testing of the administrative and security system must be done by developers and users to uncover potential problems and glitches.

Finally, as mentioned previously, it is critical that the number of items in the pool be sufficiently large to protect against their overexposure and to ensure test security. Green et al. (1995) state that depending on the algorithm used to select items and stop the test, the usable number of items in the pool may be far less than the physical number of items in it (p. 12). As a result, the test developer must provide information

about the *depth* of the item pool at various levels of ability, and identify whether and why certain items are likely to occur more often in administration as a result of the specific item-selection algorithm used. This information is especially important for users employing the scores for high-stakes decisions.

CONCLUSION

Computer-adaptive testing shows promise in becoming a regular component of standardized foreign language assessment in the coming century, particularly for licensing and certification purposes. Many benefits accrue to examinees and administrators alike when using CAT. However, to reap such benefits, numerous checks and balances need to be put into place so that the pitfalls in the development and uninformed use of CAT are eliminated. Developers and users alike need to understand fully what a CAT is and how it operates. They also need to be aware of what the underlying psychometric model used in their CAT posits in terms of the unidimensional or multidimensional IRT model selected. They need to understand what the selected IRT model means in terms of the dimensionality of the content and tasks associated with the items. They need to be familiar with how the IRT statistical parameters of the test items are estimated after their trialing. Above all, they must know what is necessary to implement a valid and reliable CAT. Bergstrom and Gershon (1994) sum up what is needed to produce such a CAT:

- A calibrated item bank large enough to administer valid, reliable, and well-targeted items across the range of candidate ability
- Test features such as content specifications, targeted test difficulty, and appropriate stopping rule
- A software program which administers items from the item bank according to design specifications
- Hardware and software adequate for speed, storage, and necessary graphics
- Educational programs to ensure that candidates understand how the CAT functions
- Report systems that are secure, useful, and transparent to users and examinees (p. 25)

Fulfilling these requirements may help to strengthen the validity, reliability, and utility of L2 CATs, however, mishandling them could have equally adverse effects (Green et al., 1995). One of the greatest dangers facing CAT users and developers alike stems from the innovative and complex aspects of the system. CAT initiates a new paradigm of testing that requires users and developers to become knowledgeable and comfortable with the characteristics that differentiate it from conventional or paper-and-pencil testing (e.g., IRT). It is the job of language testing researchers to inform and educate others about these aspects so that the strengths of CAT can indeed be harnessed and the pitfalls avoided.

NOTES

1 For an analysis of the research base and the methods of investigating the effectiveness of computer-assisted instruction (CAI) and computer-assisted language learning (CALL), see Chapelle (1997) and Dunkel (1991).

2 In a computer-adaptive test (CAT), each test item is presented by the computer which, given the response made, scores the item and then selects the next item most appropriate for the candidate,s skill level. Questions that are too easy or too difficult for the candidate are not presented (Green, Bock, Humphreys, Linn, & Reckase, 1984). Adaptive testing thus seeks to present only items that are appropriate for the test taker,s estimated level of ability.

3 Studying the relationship between the construct-irrelevant variable of computer familiarity and the construct-relevant variable of performance level on a set of test items in the computer-based Test of English as a Foreign Language (TOEFL), Taylor, Jamieson, Eignor, and Kirsch (1998) found "no practical differences" (p. 26) between computer-unfamiliar and computer-familiar examinees on the computer-based tests of listening, structure, reading, or total scores under the following conditions: (a) when examinees had been administered a TOEFL computer-based testing tutorial and (b) when the

language ability of the examinees had been taken into account. Brown (1997) raises a number of concerns about computer equipment, such as limited screen capacity and poor graphical capabilities, that could be viewed as potential disadvantages of using computers in language testing.

4 The Ohio State University is designing a multimedia CAT in French, German, and Spanish. CATs of listening comprehension proficiency in Hausa, ESL, and Russian have also been designed by Dunkel (1996) and a research team at Georgia State University and The Pennsylvania State University. Developers at Brigham Young University have for many years been actively engaged in developing CATs of second/foreign language proficiency (Madsen, 1991). More recently, the Defense Language Institute's English Language Center has been investigating and implementing computer-automated assembly of test forms, according to specific content and psychometric specifications, for use in its large-scale testing program. The automatic assembly of pre-equated language tests has cost-saving implications for large-scale testing programs and systematic test content variation (Henning, Johnson, Boutin, & Rice, 1994).

5 In a computer-based test (CBT), items are presented to the test taker in a fixed and linear fashion. They are not selected according to the examinee's previous right-wrong response patterns as are items in a CAT.

6 As Brown (1997) notes, in CAT "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" (p. 46).

7 The well-known unidimensional mathematical models in IRT (e.g., the one-, two-, and three-parameter models) handle dichotomously scored (right/wrong) data. Tung (1985) presents a rather accessible discussion of these commonly used IRT models for language teachers and testers. Briefly, one-, two-, and three-parameter models indicate the difficulty, discrimination, and guessing values, respectively, of test items in the item pool. A number of multidimensional IRT models (e.g., models for items with response formats other than right/wrong, or models that allow for multiple attempts or examinee responses for a single item) have been designed to handle open-ended response formats. For an indepth, albeit technical, discussion of unidimensional IRT models, see Hambleton and Swaminathan (1985); for an explication of multidimensional IRT models, see van der Linden and Hambleton (1997).

8 The Ohio State University is developing Web-based multimedia computer-adaptive language tests in French, German, and Spanish for placement purposes.

9 Listening comprehension CATs in Hausa, ESL, and Russian have been created with funding by the U.S. Department of Education and the National Endowment for the Humanities (Dunkel, 1996). The Hausa listening CAT is presently being used as a placement exam for Americans studying this language at the University of Kansas, and the ESL CAT is being trialed at Georgia State University.

10 Alderson, Clapham, and Wall (1995) suggest that developers of placement, progress, achievement, proficiency, and diagnostic tests offer extensive and comprehensive specifications for their assessment instruments so that users can determine exactly what abilities are being measured, what test methods are being employed, and what scoring and evaluation criteria are being used (p. 19). For a fuller discussion of how test specifications can be drawn up, see Bachman (1990) and Alderson, Clapham, and Wall (1995).

11 According to Bachman and Palmer (1996), reliability can be defined as consistency of measurement whereas construct validity pertains to the meaningfulness and appropriateness of the interpretations that can be made from the test scores. Authenticity is "the degree of correspondence of the characteristics of a given language test task to the characteristics of a TLU [Target Language Use] task" (p. 39).

12 Issues not taken up in this article, though highly deserving of careful consideration and further examination, include the trialing of items to obtain the CAT item calibrations (IRT statistics). See Wainer

(1990) and Alderson, Clapham, and Wall (1995) for an illuminating discussion of how to trial and determine item calibrations for items in the CAT item bank.

13 According to Wainer (1990, p. 114), an adaptive test terminates when one or more of the following stopping rules is met: (a) when a target measurement precision level has been attained, (b) when a pre-selected number of items has been given, or (c) when a predetermined amount of time has elapsed. Any of these stopping rules, or a mixture thereof, can be used to halt a CAT.

14 A cut-score is usually a predetermined criterion which divides scores into groups based on the examinees, level of performance. According to Green et al. (1995), a cut-score is often used to separate scores into pass/fail or mastery/non-mastery groups.

15 Alderson, Clapham, and Wall (1995) contend that the one-parameter (Rasch) model requires a minimum of 100 students for pretesting of the item bank. The Rasch model is concerned with two aspects of a test: person ability and item difficulty. The two-parameter model requires a sample of at least 200 students for trialing of the test, adding item discrimination, as well as person ability and item difficulty, to the analysis of trialed test items. In addition to everything the one- and two-parameter models do, the three-parameter model takes examinee guessing into account to determine item calibrations (IRT statistics). The three-parameter model requires a data set of at least 1,000 students.

16 For discussion of a heuristic for selecting CAT items from the item bank, based on both content and statistical properties, see Stocking, Swanson, and Pearlman (1993).

17 Green et al. (1995) note that if the purpose of the CAT is to estimate the examinees, proficiency levels, the stopping rule is usually a function of the conditional standard error of measurement for the proficiency estimate, whereas with a stopping rule used to make pass/fail decisions, the "stopping rule may instead focus on whether the [examinee's] score falls outside a pre-specified confidence band" (p. 10).

18 Wainer (1990) notes that in the Educational Testing Service's Computerized Mastery Test, examinees are initially administered two of the 10-item testlets (20 items); if they receive very low (or very high) scores, they pass (or fail). Examinees who receive less definitive scores on the two initial testlets are administered additional randomly chosen testlets until either a pass/fail decision can be reached or the pool is exhausted. "There is no adaptation of difficulty in the CMT model; its only adaptive features involve the stopping rule. Nevertheless, the computerized testlet version shortens the test for many examinees without reducing the precision of a pass-fall decision" (p. 129).

19 It should be noted that some examinees may present unusual item-response patterns as a result of unstable performance or random guessing at answers. When this occurs, calculation of adequate estimates of the examinees' proficiency levels can be difficult. CAT systems can anticipate such problems by tracking and recording the number of items attempted, the response choices and examinee patterns of response, and the final score achieved. Administrators can then examine individual test results for unusual patterns and make provision for them, for instance, by allowing examinees to retake the test (Green et al., 1995).

ABOUT THE AUTHOR

Patricia A. Dunkel is Professor of Applied Linguistics and English as a Second Language at Georgia State University. Her major areas of research and publication include L2 listening comprehension and computer-adaptive testing. She has designed models and prototype CATs in Hausa, ESL, and Russian.

E-mail: eslpad@panther.gsu.edu

REFERENCES

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

Alessi, S., & Trollip, S. (1991). *Computer-based instruction* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

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

Bergstrom, B., & Gershon, R. (Winter, 1994). Computerized adaptive testing for licensure and certification. *CEAR Exam Review*, 25-27.

Bergstrom, B., & Stahl, J. (1992). *Assessing existing item bank depth for computer adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Bernhardt, E. (1996, March). If reading is reader-based, can there be a computer adaptive reading test? In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (p. 18). Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 44-59. Retrieved November 1, 1998 from the World Wide Web: http://polyglot.cal.msu.edu/llt/vol1num1/brown/default.html.

Carslon, R. (1994). Computer-adaptive testing: A shift in the evaluation paradigm. *Journal of Educational Technology Systems, 22*, 213-224.

Chaloub-Deville, M. (1996, March). Constructing an operational framework for a CAT of L2 reading proficiency. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 1-3). Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

Chapelle, C. (1997). CALL in the year 2000: Still in search of research paradigms? *Language Learning and Technology, 1*(1), 19-43. Retrieved October 15, 1998 from the World Wide Web: http://polyglot.cal.msu.edu/llt/vol1num1/chapelle/default.html.

Cohen, A. (1994). *Assessing language ability in the classroom* (2nd ed.). Boston: Heinle & Heinle.

de Jong, J. (1996, March). Towards integrated learning and testing using structured item banks for CAT. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 38-41). Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

Dunkel, P. (1991). Research on the effectiveness of computer-assisted instruction and computer-assisted language learning. In P. Dunkel (Ed.), *Computer-assisted language learning and testing* (pp. 5-36). New York: Newbury House.

Dunkel, P. (1996). Checking the utility and appropriacy of the content and measurement models used to develop L2 listening comprehension CATS: Implications for further development of comprehension CATs. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 27-37). Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

Dunkel, P. (1997). Computer-adaptive testing of listening comprehension: A blueprint for CAT development. *The Language Teacher, 21*, 7-13, 49.

Dunkel, P. (in press). Research and development of a computer-adaptive test of listening comprehension in the less-commonly taught language Hausa. In M. Chaloub-Deville (Ed.), *Trends in computer adaptive testing.* New York: Cambridge University Press.

Eignor, D. (1996, March). Adaptive assessment of reading comprehension for TOEFL. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 46-52). Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

Eignor, D., Way, W., Stocking, M., & Steffen, M. (1997). *Case studies in computer adaptive test design through simulation* (TOEFL Research Report #93-56). Princeton, NJ: Educational Testing Service.

Green, B., Bock, R., Humphreys, L., Linn, R., & Reckase, M. (1984).Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Green, B., Kingsbury, G., Lloyd, B., Mills, C., Plake, B., Skaggs, G., Stevenson, J., Zara, T., & Schwartz, J. (1995). *Guidelines for computerized-adaptive test development and use in education.* Washington, DC: American Council on Education Credit by Examination Program.

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

Henning, G., Johnson, P., Boutin, A., & Rice, H. R. (1994). Automated assembly of pre-equated language proficiency tests. *Language Testing, 15*, 15-28.

Johnston, C. (1996, March). Computerized testing on a large network: Issues for today and tomorrow. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 7-10). Symposium conducted at the Center for Advanced Research on Language Acquisition, University of Minnesota, Bloomington, MN.

Kaya-Carton, E., Carton, A., & Dandonoli, P. (1991). Developing a computer-adaptive test of French reading proficiency. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 259-284). New York: Newbury House.

Kingsbury, G., & Houser, R. (1993). Assessing the utility of item response models: Computer adaptive testing. *Educational Measurement: Issues and Practice. 12*, 21-27.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees* (TOEFL Research Report 98-6). Princeton, NJ: Educational Testing Service.

Larson, J. (1996, March). Content considerations for testing reading proficiency via computerized-adaptive tests. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 24-26). Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

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

Linacre, J. M. (1996, March). Constructing a reading strength profile with computer adaptive testing. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 42-45). Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139-183). New York: Harper & Row.

Lord, F. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Luecht, R. (1996, March). The practical utility of Rasch measurement models. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 53-60). Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

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

McNamara, T. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing, 8*, 139-159.

McNamara, T. (1996, March). Computer adaptive testing: An outsider,s view. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 19-23). Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Bruans (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nagata, N. (1998). Input vs. output practice in educational software for second language acquisition. *Language Learning & Technology, 1*(2), 23-40. Retrieved October 15, 1998 from the World Wide Web: http://polyglot.cal.msu.edu/llt/vol1num2/article1/default.html.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice. 8*, 11-15.

Stenson, H., Graves, P., Gardiner, J., & Dally, L. (1991). *Collected works on the legal aspects of computerized adaptive testing.* Chicago: National Council of State Boards of Nursing.

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* (pp. 223-235). New York: Newbury House.

Stocking, M., Swanson, L., & Pearlman, M. (1993). Application of an automated item selection method to real data. *Applied Psychological Measurement, 17*, 167-176.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Taylor, J., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer-familiarity and performance on computer-based TOEFL test tasks* (TOEFL Research Report #61). Princeton, NJ: Educational Testing Service.

Tung, P. (1985). Computerized adaptive testing: Implications for language test developers. In C. W. Stansfield (Ed.), *Technology and language testing* (pp. 11-28). Washington, DC: Teachers of English to Speakers of Other Languages.

van der Linden, W., & Hambleton, R. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Wainer, H. (1990). *Computer adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D. (1996, March). A perspective on computerized second language testing. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency*. Symposium conducted at the Center for Advanced Research on Language Acquisition of the University of Minnesota, Bloomington, MN.

West, R. (1995). Computer-inspired testing: Multiple-matching. *Language Testing Update. 17*, 44-46.

Yao, T.-C. (1995). A computer-adaptive test for reading Chinese (CATRC): A preliminary report. *Journal of the Chinese Language Teachers Association, 30*, 75-85.

Yoes, M. (1996, March). Exploring new item-types for computerized testing: New possibilities and challenges. In M. Chaloub-Deville (Chair), *Issues in computer-adaptive testing of second language reading proficiency*. Symposium conducted at the Center for Advanced Research on Language Acquisition, University of Minnesota, Bloomington, MN.