Annual Review of Applied Linguistics (1999) **19**, 273–299. Printed in the USA.
Copyright © 1999 Cambridge University Press 0267-1905/99 $9.50

COMPUTER ADAPTIVE TESTING IN SECOND LANGUAGE CONTEXTS

Micheline Chalhoub-Deville and Craig Deville

INTRODUCTION

The widespread accessibility to large, networked computer labs at educational sites and commercial testing centers, coupled with fast-paced advances in both computer technology and measurement theory, along with the availability of off-the-shelf software for test delivery, all help to make the computerized assessment of individuals more efficient and accurate than assessment using traditional paper-and-pencil (P&P) tests. Computer adaptive testing (CAT)[1] is a form of computerized assessment that has achieved a strong foothold in licensure and certification testing and is finding greater application in many other areas as well, including education. A CAT differs from a straightforward, linear test in that an item(s) is selected for each test taker based on his/her performance on previous items. As such, assessment is tailored online to accommodate the test taker's estimated ability and confront the examinee with items that best measure that ability.

The measurement profession has been dealing with CAT issues since the early 1970s. The first CAT conference was held in 1975 and was co-sponsored by the office of Naval Research and the US Civil Service Commission (see Weiss 1978). Since then, the field has accumulated a range of research-based knowledge that addresses various psychometric and technological issues regarding the development of CAT instruments as well as the effect of various computer and CAT-specific features on test takers' performance. The second language (L2) field has only recently begun to deal with the practical aspects of CAT development and validation research. Perhaps the main reason why L2 testers are only now looking at CAT is that the L2 field has long promoted performance-based testing, whereas the general measurement researchers, especially those who have focused on CAT, have concerned themselves more with selected-response item types.

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

The purpose of the present paper is twofold: 1) It provides a broad overview of computerized testing issues with an emphasis on CAT, and 2) it furnishes sufficient breadth of coverage to enable L2 testers considering CATs to become familiar with the issues and literature in this area. The paper begins with a survey of potential CAT benefits and drawbacks; it then describes the process of CAT development; finally, the paper summarizes some of the L2 CAT instruments developed to assess various languages and skills. This last section explains approaches and decisions made by L2 researchers when developing CAT instruments, given their respective purposes for assessment and available resources. Much of the research reviewed in this paper comes from the general measurement field, as would be expected given the knowledge base accumulated in that area. The present paper, therefore, makes reference to that body of research and points out the issues that L2 CAT developers and researchers need to consider when exploring or implementing L2 CAT projects.

WHY COMPUTER ADAPTIVE TESTS?

Computer-based testing (CBT) and CAT have significantly altered the field of testing, especially for large-scale assessment, because of their notable advantages over conventional paper and pencil (P&P) tests. These advantages are due to computer capabilities as well as to the adaptive approach. The following section lists many of the potential benefits of computerized and CAT instruments, but finishes by noting several of the potential drawbacks as well. It is important to remember that any assessment approach or test method has its advantages and limitations. Moreover, depending on resources and needs, the potential advantages to some may be drawbacks to others.

1. Potential benefits of computer-based testing (CBT)

Below, we outline eight possible benefits for using Computer-Based Testing. There may be other benefits, though we believe that the following points provide a strong set of arguments.

1. Computer technologies permit individual administration of tests as opposed to requiring mass testing (Henning 1984). Individual administration reduces pressures relating to scheduling and supervising tests, and it enables more frequent and convenient testing.
2. CBT leads to greater standardization of test administration conditions.
3. CBT allows test takers to receive immediate feedback on their performance. Test takers can be provided with scores, pass/fail evaluations, placement decisions, etc., upon finishing the test.
4. The computer allows the collection and storage of various types of information about test takers' responses, for example, response time, item review strategies, items omitted or not reached, number of times examinee uses 'Help,' etc.

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

5. Computers can enhance test security. There is no need to worry about tests getting lost in shipment or test takers walking away with their test booklet. Stealing test materials would also be more difficult for test proctors. More sophisticated procedures are available to verify test taker identity. For example, a digitized picture of test takers can be captured and kept on file. Finally, computers make it relatively easy to control for item exposure.
6. CBT allows for the use of more innovative types of items and performance tasks such as dragging and dropping graphics, launching into other applications, incorporating multimedia, etc.
7. The computer permits the tracking of students' language development by storing information on students' performances over time. Students who take the test periodically can be more accurately observed and developmental profiles of students' language proficiencies can be charted.
8. CBT technologies enable the provision of special accommodations for test takers with disabilities. For example, large print or audio versions of tests can be provided to test takers who have vision impairment.

2. Potential benefits of computer adaptive testing (CAT)

In addition to the general benefits provided by CBT, tests that use computer adaptive testing approaches in particular offer further benefits, including at least the following four:

1. A CAT focuses immediately on a test taker's ability level. Whereas conventional P&P tests include a fixed number of items ranging across a broad spectrum of abilities, CAT selects a subset of items from a large item bank. Ideally, the subset of items corresponds to each test taker's ability level. Consequently, CAT requires fewer items in order to estimate test takers' abilities with the same degree of precision as conventional, linear tests, even when test takers vary widely in their ability levels (de Jong 1986, Tung 1986, Weiss and Kingsbury 1984). CATs can also lead to more accurate and reliable pass/fail decisions.
2. A CAT offers test takers a consistent realistic challenge as test takers are not forced to answer items that are too easy or too difficult for them (Henning 1991, Tung 1986).
3. The CAT algorithm enhances test security. Because each test taker is administered a different set of test items, depending on his/her language ability level, test takers sitting next to each other have minimal chance of copying from one other.
4. CAT instruments have been found to improve test-taking motivation on the part of minority and majority groups of test takers alike and to reduce the average test score differences between majority and minority groups that are frequently found in conventional P&P tests test taker (Pine, Church, Gialluca and Weiss 1979). As such, CAT instruments can be fairer or more equitable for diverse populations.

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

3. Potential drawbacks of CBT and CAT

Much like any other approach to testing, both CBT and CAT approaches have their drawbacks and limitations. There are obvious resource demands and technical expertise requirements, but other limitations also deserve consideration. Below, we outline six potential drawbacks of CBT and CAT:

1. CAT test developers must create a large number of items for the item pool and, therefore, need a large number of test takers for item piloting and calibration as well.
2. Converting a P&P exam to the computer requires conducting comparability studies to assess any potential test-delivery-medium effect. Also related to this issue is the need to develop tutorials to familiarize test takers with various aspects of taking the test on the computer.
3. Employing CAT as opposed to a linear approach means that test takers and users need to be educated about the adaptive process of the test.
4. Current CATs are unable to include extended response types of items, for example, essays, interviews, etc., that can be scored on-line. While test takers' performances on these types of items can be collected, human judgment is still required to score such performances. In a sense, CAT is often limited to the assessment of examinees' knowledge and skills, and not their performance, something that has important implications for construct validation.
5. CAT development is quite involved and costly. A high level of expertise and sophistication in computer technology and psychometrics related to CAT is required. As Dunkel (1997) points out: "It takes expertise, time, money, and persistence to launch and sustain a CAT development project. Above all it takes a lot of team work" (p. 3–4).
6. The logistics of administering a CAT are also more involved. Whereas with P&P tests a big room is needed to administer the exam to a large group of students, with CBTs and CATs a computer lab with appropriate hardware/software configurations is required. Additionally, because CBTs and CATs are touted as enabling individual test administration, the computer lab should have flexible hours to accommodate test takers' diverse schedules.

Overall, CBT and CAT offer exciting benefits that are worth pursuing in L2 testing. Some of the concerns about CBT and CAT, for example, the need to educate test takers and users about the adaptive testing process and the ensuing scores, or the limitations of item types with CAT, are issues that are likely to be of less concern with increased use and continued research.

Constructing CAT instruments requires making decisions about various issues along the way. CAT developers need to identify their technology resources, determine the appropriate content and makeup of the item bank, conduct a comparability study in certain cases, and decide which item selection algorithm to use. These issues are quite involved and require various types of expertise. The

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

following sections address these issues and examine the major considerations involved.

TECHNOLOGY

The 1996 volume of the *Annual Review of Applied Linguistics* includes a section that deals with technology in language instruction. The articles in that section address various issues from distance learning, to web-based instruction, to computer-assisted language instruction. The volume also includes an article by Burstein, Frase, Ginther and Grant that reviews hardware and software technologies for language assessment and provides an overview of technology issues when developing CATs. Additionally, a report published as part of the TOEFL Monograph Series, by Frase, *et al.* (1998), presents a comprehensive review of diverse technology topics, including CAT. These topics focus, among other things, on user-centered technology and on operating, authoring, and delivery systems for test development and test use. Given the extent of coverage of technology in the publications cited above and in other publications (e.g., Brown 1997, Mancall, Bashook and Dockery 1996, Sands, Waters and McBride 1997), the fast-paced changes in the industry, and the fact that many L2 CAT developers will be restricted to work within the constraints of onsite computer labs, the present paper will not delve into the technology issues in any depth. Readers are encouraged to refer to the referenced sources for more detailed information about CBT and CAT technology.

That being said, several commercial software and test delivery companies have been employed successfully in the L2 field, including Administrator (Computer Assessment Technologies) and MicroCAT (Assessment Systems Corporation). (See Appendix A.) Although both of these products can be purchased 'as is,' the distributors are also willing to modify their test engines to some degree. L2 testers who wish to construct tests in less commonly taught languages (e.g., Japanese or Arabic) need to make sure that the test engine can handle double-byte characters. Otherwise, the text must be stored and displayed as picture files, something that can substantially slow down transmission and online display of items. As for test delivery, especially for large-scale testing, companies such as Sylvan Prometric have been involved in delivering CBT and CAT instruments worldwide. ETS and CITO are two testing organizations that contract for selected services with Sylvan Prometric, although other delivery companies, such as National Computer Systems (NCS) and Assessment Systems Incorporated (ASI), also offer excellent services. (See Appendix A.)

Language CBTs and CATs have largely focused on receptive skills, mainly reading, and on discretely measured components such as grammar and vocabulary. The assessment of speaking via the computer has been largely ignored, mainly because of technology constraints. Recent advancements in speech recognition technology, however, have enabled the automated assessment of various aspects of the speaking skill. For example, PhonePass (Ordinate

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

Corporation 1998) is an assessment instrument that capitalizes on the technology to examine skills such as listening, fluency, pronunciation, syntax, and vocabulary—all discrete elements that sustain oral conversation (in this case, in American English). As the name indicates, PhonePass is administered over the phone via a computer system. Five types of items are included in PhonePass: reading aloud, repeating sentence, naming opposite words, providing short answers, and giving open responses. The first four response types are digitized and evaluated by an automated speech recognition system. The system includes "...an HMM-based speech recognizer that uses acoustic models, pronunciation dictionaries, and expected-response networks that were custom-built from data collected during administrations of the PhonePass test to over 400 native speakers and 2700 non-native speakers of English" (Ordinate Corporation 1998:4–5). The open response is stored and made available to interested test-users. Obviously, as *the test developers themselves point out, this instrument does not assess advanced* speaking skills, but measures the relatively more mechanical aspects of conversations. Nonetheless, progress in the technology of assessment, as exemplified by this instrument, is exciting and signifies the kind of CBT and CAT capabilities that we can expect before long.

Finally, computer technology can enhance test security in numerous ways. Item and score result files can be encrypted; separate text, picture, and multimedia files can be maintained and combined as needed; transmission and real time delivery of items, item panels, or item banks, can be accomplished in secure ways; 'enemy' item combinations (that is, where the information contained in one item will give away the answer to another item) can be avoided; records can be kept of which items examinees have seen; examinee registration databases can exercise control over repeated testing; etc. Again, although readers and test developers are encouraged to examine these issues with ready-made software, these security technologies may have to be set up within the constraints of local technologies.

CAT ITEM BANK

In addition to considering technology resources and needs, L2 CAT developers need to design and develop the CAT item bank. An item bank is a pool of items with established content specifications and item parameters intended to measure examinee abilities at various levels. The issues to consider in creating an item bank include, among other things, a planning stage, a pilot study and item calibration, and if needed, a comparability study.

A basic, first step in any test development, including CAT, is identifying and describing the L2 aspects being measured, that is, the L2 content domain. Next, similar to P&P tests, test specifications should be developed that include specific information about content and test methods. A difference between linear tests and CAT, however, arises from the fact that a somewhat unique set of items is administered to each test taker at various ability levels during a CAT. A linear test covers the specified content without much concern for item difficulty as all test

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

takers see the same items. Because each CAT will be unique, the content must be covered at all ability levels to ensure that examinees, regardless of ability, are exposed to items that adequately cover the content. For this reason (among others), it should be clear that relatively more items are required to construct a good item bank.

In addition to content coverage, test developers must consider other variables that impact the number of items needed in a CAT item pool. Stocking (1994) states that such factors include the measurement model, item content constraints, item exposure (the number of times an item is administered), and entry and exit rules (see below). Weiss (1985) suggests that a "...CAT operates most effectively from an item pool with a large number of items that are highly discriminating and are equally represented across the difficulty-trait level continuum" (p. 786). In general, the more attributes and properties added to the CAT item bank design, the larger the item pool that is required. Stocking (1994) provides tables that help predict the item pool size necessary for simpler item-bank designs. In any case, readers should realize that developing large item pools is quite costly and may prove to be prohibitive for some. In addition, items that are accompanied by graphics, sound, or video are even more costly to develop, and they put increased demands on the technology required to store and deliver such tests.

Yet another issue involves whether to conceive of an item as one question or as an item bundle, or 'testlet.' Wainer and Kiely (1987) describe a testlet as "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow" (p. 190). Testlets are found in many kinds of tests, but are especially prevalent in language tests where multiple items evaluate a reader's or listener's comprehension of a single passage. Because such a group of items is linked to one stimulus, the items share a common context and are likely to be dependent, requiring that the test taker's score on the testlet be considered, and not the score on the individual items separately.

After developing test specifications, items and testlets are created and pilot tested. Classical and item response theory (IRT) analyses need to be performed to identify good items, to revise items, and to discard bad items. A variety of dichotomous IRT models are available. The most popular models are the one-, two-, and three-parameter logistic models (1PLM, 2PLM, and 3PLM *respectively). (For more information on IRT, see Hambleton and Swaminathan* 1985, Hambleton, Swaminathan and Rogers 1991.)

Items retained are administered to test takers for calibration. Test takers' performance on these items is used to estimate item properties, such as difficulty level, discrimination power, and guessing index. These properties are subsequently utilized as item parameters to help determine item selection in CAT. As compared to P&P tests, considerably more items need to be calibrated through

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

piloting, which mandates larger numbers of test takers and complex field-test designs. Sample size depends on the number of items, the measurement model chosen, and the quality of the test taker sample. Data and model fit also need to be established. Additionally, dimensionality analyses such as factor analysis and *DIMTEST (Nandakumar and Stout 1993, Stout 1987) need to be performed to* ascertain the unidimensionality of the different test components, a critical *assumption for CAT measurement models* (Chalhoub-Deville, Alcaya and Lozier 1997). For a comprehensive review of the psychometric issues involved in developing a CAT item pool, readers are referred to Wainer (1990).

Although evident, it is worth repeating that when piloting items, test-developers should make sure they have a representative sample of their intended future test takers. Brown and Iwashita (1996; 1998) document problems that can arise in piloting when CAT developers do not have a representative sample of test takers. In their investigation of a Japanese CAT used for placement into a Japanese language program at an Australian university, the researchers document how test takers are found to misfit when item difficulty is computed based on the performance of test takers of a different L1 background than those used in the pilot test.

Finally, if test developers are converting existing P&P tests into CAT, they need to conduct research to obtain evidence of performance and score comparability (see below). Otherwise, test developers need to conduct their piloting on the computer to avoid factors that might mitigate test taker performance.

TEST SCORE COMPARABILITY

The introduction of CBT has been accompanied by concerns for the comparability of scores obtained with CBTs/CATs and their P&P counterparts. Bunderson, Inouye and Olson (1989), in a review of the literature investigating this issue, indicated that P&P test scores were often higher than scores from the CBTs. Nevertheless, this difference in scores was "generally quite small and of little practical significance" (1989:378). Mead and Drasgow (1993), in a meta-analysis of 29 equivalence studies, concluded the following: "[The results] provide strong support for the conclusion that there is no medium effect for carefully constructed power tests. Moreover, no effect was found for adaptivity. On the other hand, a substantial medium effect was found for speeded tests" (1993:457). The authors conclude, nevertheless, by cautioning against taking the equivalency of computer and P&P scores for granted. Comparability of scores needs to be documented in local settings.

It is worth noting here that studies looking into score comparability issues have focused on assessments that typically use selected response item types (e.g., multiple-choice). Investigations with more open-ended types of items, however,

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

are not as well documented. In conclusion, test developers need to be cautious in generalizing score equivalency research findings to constructed response items.

When developing a CAT, language testers are advised to gather evidence of the comparability of item parameters across mediums of delivery and not ignore this issue. One example of the type of research that can be carried out to investigate score comparability is that of Fulcher (in press). He has undertaken a study that examines score comparability when converting a P&P ESL placement test at the University of Surrey to CBT. He points out that, while score comparability, which has been the focus of test conversion studies, is critical, it is not the only variable to be considered. He maintains that test takers' previous experiences with and attitudes towards computers, as well as their backgrounds, also need to be considered. These variables may confound the measure of the L2 proficiency construct when using the computer as the medium of delivery.

In the L2 field, the most expansive research investigating test takers' experiences with computers and the subsequent effect on L2 test performance has been carried out by the TOEFL Program. As part of its effort to launch CBT TOEFL and to prepare for TOEFL 2000, ETS has recently undertaken a large-scale research agenda to document TOEFL test takers' familiarity with computers and examine the relationship between computer familiarity and CBT TOEFL performance. Based on an extensive survey of the literature, researchers have developed a questionnaire that probes test takers' access to, attitude toward, and experience in using computers (Eignor, *et al.* 1998, Kirsch, *et al.* 1998). The questionnaire was administered to a representative sample of 90,000 TOEFL test takers. Survey results show that approximately 16 percent of test takers in the sample can be classified as having low computer familiarity, 34 percent had moderate familiarity, and 50 percent had high familiarity. Several background variables were considered in the computer familiarity research. Findings show that:

> computer familiarity was unrelated to age, but was related to gender, native language, region of the world where the examinee was born, and test-center region. Computer familiarity was also shown to be related to individuals' TOEFL [P& P] test scores and their reason for taking the test [graduate versus undergraduate] but unrelated to whether or not they had taken the test previously (Kirsch, *et al.*, p. i).

Considering that very large numbers of persons take the TOEFL each year, 16 percent of test takers reporting low computer familiarity represents a substantial group, and these results have prompted the researchers to find a way to help address the issue. A tutorial has been developed that test takers see before starting the test (Jamieson, *et al.* 1998). A representative sample of 1100 TOEFL test takers, grouped according to high and low computer familiarity, were administered the tutorial and a 60-question CBT TOEFL. Subsequently, the relationship between level of computer familiarity and TOEFL CBT was examined,

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

controlling for L2 ability. In short, results show no practical differences between computer-familiarity and computer-unfamiliar test takers on TOEFL and its subparts (Taylor, *et al.* 1998). Nevertheless, as the researchers themselves write, more research is needed to examine the relationship between various background variables and CBT performance.

In order to further enhance test takers' familiarity with computers and help reduce the computer medium effect on test performance, a TOEFL Sampler, which is an instructional CD-ROM that includes seven tutorials, is being disseminated free of charge. Three of these tutorials familiarize potential test takers with how to scroll, use a mouse, and use the various testing tools such as 'Help.' The other four tutorials provide information about the exam and practice questions that focus on the four sections of the TOEFL: listening, reading, structure, and essay. It is likely that test preparation companies will provide materials and strategies to familiarize TOEFL test takers with the new technologies. Currently, Kaplan Testing Centers disseminate information about GRE CAT test-taking strategies on its web page: www.kaplan.com/gre/grecat/catwords.html and will probably develop a similar web page for the TOEFL.

In conclusion, research in L2 is still scarce regarding the comparability of P&P and computer scores. As a result, it would be unwise to generalize findings described above to local settings without first examining the test taker population and other variables.

ITEM SELECTION ALGORITHM

Another major issue to consider in CAT construction is the choice of an adaptive or item selection algorithm. The adaptive algorithm is a procedure that selects from the CAT item pool the most appropriate item for each test taker during the test depending on the questions seen and answers given. Items are selected (and sometimes sequenced) based on content and item parameters. An algorithm must specify starting and stopping rules for the CAT, and will likely account for content balancing and item exposure.

Test developers can either custom design the adaptive algorithm or purchase a software package that includes an adaptive procedure. (The reader is referred to the vendors listed in Appendix A who distribute such software.) Whether the software is custom designed or purchased off-the-shelf, CAT developers still have the responsibility of making informed decisions regarding the adaptive algorithm. As stated in the American Psychological Association's (APA) computer testing guidelines, "...none of these applications of computer technology is any better than the decision rules or algorithm upon which they are based. The judgment required to make appropriate decisions based on information provided by a computer is the responsibility of the test user" (APA 1986:8).

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

1. CAT entry point

Although an off-the-shelf software has adaptive capabilities, the specific entry point, (i.e., the first item to be administered) needs to be determined and incorporated into the algorithm. With an appropriate entry point close to the test taker's ability, the examinee will more quickly face challenging items that provide useful information in order to estimate his/her final ability level more accurately.

How can we obtain an individual's initial L2 ability estimate for selecting the appropriate first item(s) in CAT? Typically, items of average difficulty or, in criterion-referenced contexts, items near the cut point are administered first (Stevenson and Gross 1991). Another possibility is to have the test takers do a self-assessment of their abilities and use their estimates as a starting point. Yet another possibility is first to present each test taker with the same set of items of varying difficulty and, based on their performance on these items, choose the initial 'real' item for the test taker. The CAT developer might also consider using demographic information (e.g., number of years the test taker has studied the language) or previous test scores to gauge an appropriate entry point for the test taker.

2. Exit point and test length

The exit level, the point at which the computer algorithm terminates the test, also needs to be set. The exit point is critical because it impacts test takers' scores. Very often a CAT is terminated when a prespecified accuracy level of an ability estimate is reached (Henning 1987). Depending on the response pattern of individual test takers, however, this means that the test length, or number of items delivered, differs from one examinee to the next. In setting the exit point, CAT developers need to decide whether to have the test be variable length, as just mentioned, or fixed length. Fixed length CATs can be comforting to test takers who know they all took the same number of items. But their drawback is that not all examinees are measured with the same degree of precision. For longer CATs, however, where error is quite low, this drawback can be trivial.

When a test taker is confronted with an item at his/her ability level, s/he theoretically has a 50 percent probability of answering the item correctly. The algorithm can be set so that examinees have a higher probability of getting items right (e.g., 70 percent). Such an approach can compromise the efficiency and measurement precision somewhat, the effect of which, nonetheless can be calculated. This will also lead to longer CATs. The advantage is that test takers are less likely to guess at items (Linacre 1999) and will experience less frustration.

Last, test length can be determined by fixing the allowable time an examinee can have. Most CATs, however, are not designed to be speeded tests, but are power tests, whereby most examinees have sufficient time to finish.

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

3. Content balancing and item exposure

Testing efficiency, that is, reducing testing time, was of primary importance to many who first worked on CATs (see Weiss 1982). Researchers soon realized, however, that other very important considerations impinged on efficiency to some degree. An item selection algorithm that chooses items in order to efficiently maximize the precision of estimating an examinee's ability will likely sacrifice content coverage and representativeness. Content balancing assures that the content domain operationalized within each administered CAT is covered adequately and represented appropriately (Lunz and Deville 1996). While balancing content compromises testing efficiency to some extent, it provides critical content validity evidence and maintains the primacy of content over all other considerations.

Controlling for item exposure can also lead to slightly longer CATs but helps ensure that items are not seen by too many candidates and thus become compromised (Sympson and Hetter 1985). Item over-exposure is especially grievous when item pools are not sufficiently large, when content balancing is built in without content coverage at various ability levels, when new items are not rotated in regularly, when tests are administered on an ongoing basis, and when large numbers of examinees are rather homogeneous in ability. An item selection algorithm that neglects to control for item exposure will likely deliver some items over and over again, while other items are hardly ever seen by test takers.

Stocking and her colleagues at ETS (Stocking 1992, Stocking and Lewis 1995, Stocking and Swanson 1993, Swanson and Stocking 1993) have devoted considerable attention to these issues and have developed very sophisticated item selection algorithms. Luecht (1998) and his colleagues at the National Board of Medical Examiners (Luecht, Nungester and Hadadi 1996), in their work developing a CAT system for the high-stakes medical licensure field, have provided a most innovative solution to these issues. Luecht devised CAST (computer adaptive sequential testing), a CAT algorithm that adapts at the *subtest* rather than the item level. Some of the many advantages of CAST include the following:

> First, it is adaptive and therefore, efficient... Second, it allows explicit control over many different features of content-balance, including the possibility to conduct quality reviews at the level of subtests or panels. Third, statistical test characteristics are determined via the target test information functions in order to control rather precisely where and how much score precision is allocated to various regions of the score scale. Fourth, it makes concrete use of existing automated test assembly procedures.... Fifth, it can exploit the same and even additional exposure and item randomization controls used in CAT to minimize risks of unfair benefit to examinees having prior access to memorized items. Sixth, only item data...for active panels are at risk on any particular day of testing at

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

> test delivery sites. Finally,..examinees can actually review and change answers within a particular subtest (Luecht, Nungester and Hadadi 1996:18).

Although Stocking's algorithms and Luecht's CAST model may not be practical solutions to the issues of content balancing and item exposure for many language testers today, their research is finding implementation now and will likely be more accessible in the future.

4. Innovations

While the first generation of computerized tests and CATs engendered legitimate excitement with regard to simplified administration procedures and testing efficiency, there was little innovative thinking about item types beyond the selected response format (for an exception, see Weiss 1978). Linear, multiple-choice, P&P tests were simply delivered via the computer, and those testing agencies with the resources to develop CATs often did so primarily to reduce testing time. More recently, however, research is being devoted to the development, delivery, and scoring of complex performance tasks on the computer. In addition, testing efficiency is now viewed as a positive side benefit of CAT, and rarely as the primary benefit.

A limitation of computerized tests and CATs until recently has been the difficulty of administering and scoring constructed response items and performance tasks. Even now, such test development requires a very high degree of computer and psychometric expertise (not to mention other resources such as money and time) that it is prohibitive for many. Nevertheless, interesting and worthwhile test development and research in this arena (e.g., Sands, Waters and McBride 1997) will likely lead to more practical and affordable solutions in the near future.

Davey, Godwin and Mittelholz (1997) report on the development, administration, and scoring of an innovative writing ability test, COMPASS, used for placement purposes. Examinees are presented with a writing passage and asked to edit any or all segments of the passage for grammar, organization, or style. The examinee has full freedom to choose what to edit by clicking on a section and choosing from a set of alternatives. The examinee can thus edit a correct segment or insert an incorrect alternative to an already incorrect segment. When the examinee chooses an alternative, the computer inserts that into the passage. In this fashion, the test taker can essentially 'rewrite' the entire passage.

Not only the item type but also the psychometric model is somewhat innovative. Because COMPASS is a placement test, the assignment of test takers to an appropriate ability group, and the differentiation among groups, is more important than accurately differentiating one examinee from another. The authors use a classification-based measurement model that utilizes the sequential probability of ratio test (SPRT) (Wald 1947). SPRT estimates the probability

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

whether the test taker has exceeded a performance threshold or not and continues to administer items until a specified criterion level of confidence has been reached, and the examinee can then be classified.

Work has also been undertaken examining the use of CAT with open-ended responses in mathematical reasoning (Bennet, *et al.* 1997). Examinees had to produce mathematical expressions for which there was one correct response but which could take innumerable forms. The authors developed a special computer interface and automatic scoring algorithm to obtain and evaluate examinee responses. This study demonstrates how certain kinds of examinee-constructed tests can be delivered via the computer and scored accurately.

5. Miscellaneous CAT issues

Most CATs do not allow omissions. Examinees must respond to an item before being allowed to go on to the next item. Obviously, if a test taker skips items, the CAT algorithm cannot estimate ability, so the presentation of items can become somewhat random. In addition, students can skip items until they find questions they can answer, resulting in overestimated scores (Lunz and Bergstrom 1994).

With paper-and pencil tests, test takers have the opportunity to review items. Because CAT item sequence depends on student performance on every item, item review might jeopardize the adaptive test process. Lunz, Bergstrom and Wright (1992), however, indicate that item review is something examinees prefer and has very little influence on test takers' scores. One way around this potential dilemma is with the CAST model described above.

L2 CAT PROJECTS

Several interesting CAT instruments have been developed in the L2 field. The purpose of this section is to present some of these CATs and describe their basic features. Briefly, the projects portray the variety of decisions made as well as the approaches to creating CATs to accommodate diverse purposes and available resources.

1. The TOEFL

In July 1998, ETS launched CBT TOEFL in the US and countries around the world, except in a few Asian regions. The CBT is scheduled to be offered in the remaining regions in the year 2000. The CBT TOEFL enables year-round testing at over 300 centers worldwide. The test begins with a mandatory but untimed tutorial that familiarizes test takers with crucial computer functions, such as how to scroll, use the mouse, click on answers, etc. At the beginning of each section of the TOEFL, a tutorial is presented to familiarize test takers with the

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

directions, format, and question types found in that section. The average time to finish all the tutorial components is 40 minutes.

The listening and structure parts of the CBT TOEFL are adaptive. The IRT model adopted for the adaptive algorithm is the three parameter IRT model. Test takers are presented with one item at a time and no omissions are allowed. Exposure control parameters are employed to help ensure that very 'popular' items (in terms of information and test design) are not overexposed. The CAT algorithm in the listening section samples various content types of listening material, including dialogues, conversations, academic discussions, and minilectures. The questions examine comprehension of main ideas, supporting ideas, key details, and text-based inferences. Questions include four types: multiple-choice, selection of a visual/part of a visual, selection of two choices (out of four), and matching or ordering objects or text. Two kinds of visuals, content- and context-based, accompany the passages. The content-based visuals often complement the topics in the minilectures. The context-based visuals accompany all types of listening passages and help establish the setting and the roles of the speakers.

The structure CAT section includes two types of multiple-choice questions, similar to the P&P TOEFL: selecting the option that completes a sentence and identifying the incorrect option. The algorithm is designed to sample these two types of questions randomly.

With regard to the TOEFL reading test, TOEFL researchers have decided against adopting an adaptive algorithm because of the relatively large number of items associated with any given reading text and the interrelatedness of these items. The argument is that such interrelatedness violates the assumption of local independence required for the IRT model underlying the adaptive algorithm. Furthermore, if an adaptive testlet model were adopted, little if anything would be gained in terms of efficiency or accuracy. As a result, the reading section is CBT in format. Specifically, test takers are administered linear sections of reading passages that have been "constructed 'on the fly'" to meet test design requirements, thus ensuring that tests are parallel in both content and delivery (Eignor 1999:173). As such, each test taker receives an individualized combination of reading passages and items. Exposure control parameters are also employed in this reading section to help ensure that items are not overexposed.

The CBT TOEFL includes a writing essay and the writing score is added to that from the structure section. Test takers have the option of either handwriting or typing their essays. The handwritten essays are then scanned and scored by two independent readers. Research is being conducted on the feasibility and validity of using automated scoring of the essays.

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

2. ESL listening comprehension

The lead developer of this listening comprehension CAT is Patricia Dunkel, Georgia State University. The purpose of the instrument is to examine ESL students' listening comprehension ability for placement into or exit from adult ESL programs. The CAT includes topics and authentic listening excerpts that vary in their extensiveness and cultural references to accommodate the various language-ability and cultural-awareness levels of the test takers. The CAT item bank includes items ranging from comprehension of discrete words/phrases to variable length monologues and dialogues, authentic radio segments, and scripted texts. Four listener functions, as identified by Lund's (1990) taxonomy of listening skills, form the framework of test tasks for the listening items: recognition/identification, orientation, comprehension of main ideas, and understanding and recognition of details in the texts heard. Item formats include multiple-choice, matching a graphic to what was heard in a text, and identifying appropriate elements in a graphic. Students' results are reported using a nine-level scale representing the ACTFL scale continuum (Novice-Superior).

The hardware used to deliver the test is a Macintosh IIci with at least 8 megabytes of RAM and speech output capabilities. The software was created using the C++ language and a CAT testing shell created by programmers and instructional designers at Pennsylvania State University. The CAT shell was custom designed for the project and is presently being updated to be cross-platform and capable of displaying full-motion video as well as graphics and sound. The CAT algorithm used is based on that developed by Henning (1987). The item selection algorithm is based on Rasch estimation and is structured to provide the test taker with an initial item of median difficulty followed by an item one logit of difficulty above or below the first item, depending on the performance of the test taker. The algorithm estimates ability and provides the associated error of estimate after four items are attempted. The CAT terminates once the error of estimate falls below 0.5. For more information on this CAT, see Dunkel (1991) and (1997).

3. Hausa listening comprehension

Patricia Dunkel is also the lead developer of a Hausa listening CAT. The purpose of the instrument is to evaluate the listening comprehension of American University students studying Hausa. The Hausa CAT follows to a large extent the content and item specifications, algorithm design, and scoring described above for the ESL listening CAT. Instructions and the items, however, are presented in English. The Hausa CAT is presently being used for placement and exit purposes at the University of Kansas Hausa Program. (For more information on this CAT see Dunkel 1999.)

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

4. The CAPT

Michel Laurier, University of Montreal, is the lead developer of the French Computer Adaptive Proficiency Test (CAPT) used to place English speakers enrolled in French language courses at the post-secondary level in Canada. The shell and algorithm were developed locally using an IBM platform. The CAPT includes multiple-choice items that assess the following abilities: 1) reading comprehension of short-paragraphs typically encountered in daily life, 2) sociolinguistic knowledge, 3) lexical and grammatical knowledge, 4) listening comprehension of two-minute "semi-authentic" passages, and 5) self-assessment of oral skills. Items pertaining to each of these five skills are kept in separate item banks, and five distinct subtests are administered to each test taker.

The test begins by asking the learner questions about his/her language background, (e.g., number of years of French studied, time spent in a French environment, and self-assessment of French ability). This information is pooled to determine the entry point for the test taker. Ability and error are first estimated after the student has answered five items. Additional items are presented until the error of estimate falls below .25 logits or until the student answers the allowable maximum number of items. The score from each subtest is then used to provide the entry point into subsequent subtests.

The algorithm is somewhat different, however, for the listening and self-assessment subtests. With regard to the listening comprehension subtest, three questions are presented on screen before the student hears the passage once. Passages differ in their difficulty level and altogether a test taker is presented with three to five passages, depending on his/her ability level. The oral self-assessment component is also adaptive, where students rate their ability on "can do" items using a six-step scale.

An overall ability level is estimated by simply obtaining an average of the five scores. Laurier (1999) points out, however, that "...should a given institution have specific needs, the weight of the subtests could be changed in the program." The IRT model selected for the first three subtests is a three-parameter model using BILOG. MULTILOG, designed to handle graded-response items, is used for the other subtests. For more information on this CAT see Laurier (1999).

5. Dutch reading proficiency CAT

The Language Training Division of the CIA collaborated with Brigham Young University (BYU) to produce a reading proficiency CAT in Dutch. The test includes an orientation component that checks test takers' familiarity with computers and introduces them to the computer layout of the keyboard and the key strokes necessary during the test. Practice items are also presented. The test simulates three phases of the oral proficiency interview: level check, probe, and wind-down. The CAT begins by administering nine selected response items that

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

span the entire ILR reading proficiency scale, levels 1 to 5, so that all examinees are confronted with items from all levels. (The ILR levels were further subdivided into levels 0 to 41 based on Rasch calibrations.) The CAT then embarks on the level check by starting at a low level item and "...advanc[ing] six levels for each subsequent item if the previous item was answered correctly, or go[ing] back five levels if the previous item was answered incorrectly. This branching continues for seven iterations" (Larson 1999:85). Test takers are then presented with items to help determine their reading proficiency "ceiling" level. The ceiling level is defined as the level at which the test taker misses four items. At that point the reading ability estimate of the test taker is computed along with the standard error estimate. The algorithm, nevertheless, continues to provide students with questions. In the wind-down section, higher difficulty items are first presented in order to "...satisfy the concern that might be expressed by some examinees that they had not been given a sufficient number of items to determine accurately their true performance" (Larson 1999:86). Then items below the estimated ability are provided, allowing test takers to leave with positive feelings about their performance.

The CAT algorithm is also designed to ensure content representativeness by balancing content, context, abstract/concrete passages, and cultural understanding. Items can be flagged as "enemies"; that is, if content in one item might give away the answer to another item, the two will not be presented to a test taker. Five inventive item types are included in the test: best meaning, best misfit, best restatement, best summary, and best logical completion. Each CAT typically includes 15 items that are being field-tested. These items do not contribute to the test takers' ability estimation, but provide useful information for further test development. For more information about this CAT see Larson (1999).

6. French, German, Spanish, Russian, and ESL placement CATs

These placement tests have been developed by a research team at Brigham Young University to evaluate test takers' ability levels in grammar, reading, and vocabulary. These tests were among the first L2 CATs developed and are typically used to place incoming students in language curricula at universities in the US. The ESL CAT is a relatively new addition and contains a listening component. The item selection algorithm for the CATs is based on Rasch estimation. Content sampling is random within each of three identified skills listed above. Additionally, the stopping rule adopted for these CATs is a standard error of estimate below .4 logits. These CATs are available to run on either PCs or Macintosh computers. Demo disks are available by contacting BYU.

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

7. ESL reading comprehension (Young, Shermis, Brutten and Perkins 1996)

The purpose of this CAT is to assess test takers' reading comprehension as they move from one level to the next in a four-course ESL program at Southern Illinois University. The CAT was converted from a battery of four P&P tests and includes variable-length reading passages on diverse topics. Items are calibrated using RASCAL (i.e., the Rasch model). While the multiple-choice items, presented one at a time, are provided on the computer, the reading text itself is not. The text is included in a printed booklet and the computer screen refers the test taker to the designated text in that booklet. This approach is done to "...minimize the test method differences between the pencil-and-paper and computer adaptive tests...[and because it is] considerably easier for the reader to scan the whole of long passages on paper than to scroll through it a few lines at a time on a computer monitor" (Young, *et al.* 1996:29).

Macintosh HyperCAT, developed by Shermis, is the development system used for this CAT. The starting point for the CAT is based on the course level just completed by the test taker. The CAT terminates when the test information function is equal to or less than a prespecified value, or when the number of items administered reaches a prespecified limit.

Item difficulty is the only parameter considered in the adaptive algorithm. No constraints are placed on the repeated presentation of the same passage. As such, the test taker is likely to encounter the same passage repeatedly during the test, each time with a different item of appropriate difficulty at that point. The authors point out that they are considering bundling items together to allow for a testlet-based approach and thereby avoid this repetition.

8. Other L2 CAT instruments

A number of other CAT instruments are under development by various institutions around the world. For example, Ohio State University has been creating multi-media CAT placement instruments for various languages, including French, German, and Spanish. These CATs assess reading, listening, and grammar skills. Likewise, The University of Minnesota has been involved in the development of CATs to assess students' reading proficiencies in French, German, and Spanish, mainly for entrance into and exit from these language programs at the post-secondary level. Michigan State University is also developing placement CATs for French, German, and Spanish to assess university students' reading, vocabulary, and listening skills. Language testers at UCLA have also expressed interest in developing a placement CAT there. Finally, the Defense Language Institute has converted its P&P English Language Proficiency test, focusing on reading and listening, to CAT. This CAT is restricted to U.S. government use.

Similar interest in developing CATs is growing in Europe. The University of Cambridge Local Examinations Syndicate has been involved in developing CAT

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

instruments for various languages and purposes. For example, CommuniCAT, intended primarily for language schools and university language centers, and BULATS, more appropriate for the corporate sector, have been produced in several languages: English, French, German, Spanish, Italian, and Dutch. These CATs have been piloted with an international group of test takers. They include audio and graphics, and provide on-screen help. They also offer the flexibility of providing test instructions in a different language than that of the test (e.g., an English test with Spanish instructions and on-screen help). Another European project is DIALANG, a CBT/CAT (that eventually will be delivered on the Internet), coordinated by the Centre for Applied Language Studies at the University of Jyväskylä in Finland in cooperation with various European universities and research institutes. DIALANG will include 14 languages (the official EU languages plus Irish, Icelandic, and Norwegian). The instruments will assess all four language skills, vocabulary, and structure. Finally, CITO has been delivering CATs in listening for some time. It should be clear that CBTs and CATs have already made inroads in the area of language testing, and they will likely find even more widespread implementation in the near future.

CONCLUSION

Computer technology provides expanded possibilities for test development, administration, scoring, and thus decision-making regarding examinee abilities. In this paper, we have presented many of the issues related to the decisions that accompany the development, administration, and scoring of L2 CBTs and CATs, all of which influence how scores will be interpreted and used (i.e., the validation process). With regard to validation, developers are reminded that CAT is but one form of assessment and must be subjected to analyses that buttress their validity argument. In language testing, we know all too well that test methods can and do influence scores and thus alter our subsequent use and interpretation of the scores. At the risk of sounding overly trite—no method is a panacea; each comes with its own set of advantages and disadvantages. Our job is to discern wisely when and how to make use of the various methods in order to obtain accurate and fair measures of our test takers' abilities.

NOTES

1. In the language testing field, the acronym 'CAT' sometimes refers to the field, 'Computer Adaptive Testing,' and sometimes to the test instrument itself, 'Computer Adaptive Test.'

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

ANNOTATED BIBLIOGRAPHY

Brown, J. D. 1997. Computers in language testing: Present research and some future directions. *Language Learning & Technology*. 1.44–59. [Retrieved August 15, 1998 from the World Wide Web: http://polyglot.cal.msu.edu/llt/vol1num1/brown/default.html]

In this article, Brown provides an overview of various developments related to the use of computers in language testing. He addresses item banking uses, technology and computer-based language testing, and the effectiveness of computers in language testing. The article also reviews some of the major issues discussed in the general measurement field regarding CAT. Finally, Brown highlights research issues that need to be undertaken by L2 test developers and researchers in order to further CAT research in the L2 field.

Chalhoub-Deville, M. (ed.) 1999. *Issues in computer adaptive testing of reading proficiency*. New York: Cambridge University Press.

The book addresses the fundamental issues regarding the development of, and research on, L2 CAT for assessing the receptive skills, mainly reading. The chapters by the various authors in this edited volume are grouped into three major sections: the L2 reading construct, L2 CAT applications and considerations, and item response theory (IRT) measurement issues. Discussion chapters are included in each of the three sections. These chapters highlight and discuss the issues raised by the authors in their respective sections as well as those of immediate relevance in the other sections. The book also provides a critical discussion of CAT practices from the point of view of performance assessment.

Dunkel, P. (ed.) 1991. *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House.

This edited volume includes two major sections. The first section presents several chapters on computer-assisted language instruction and learning research and applications. The second section focuses for the most part on CAT and includes various studies that explore the design and effectiveness of CAT for assessing L2 proficiency. The chapters address a range of technical and logistical considerations for the development, maintenance, and use of CATs; they describe different L2 CAT instruments for assessing students' L2 proficiency in schools as well as at universities; and they discuss a range of issues that impact CAT research validation agendas.

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

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

This publication introduces the basic concepts of IRT and describes how IRT approaches can be utilized for various purposes, including test development, test bias identification, and CAT. Additionally, the volume provides thorough discussions of various procedures for IRT parameter estimation (e.g., maximum likelihood estimation and Bayesian estimation). The book provides many examples that illustrate the topics discussed. The authors have succeeded in presenting complex measurement concepts and procedures that are accessible to those with limited mathematical backgrounds. Finally, the book explores new directions in IRT development and research.

Wainer, H. (ed.) 1990. *Computerized adaptive testing: A primer*. Hillsdale, NJ: L. Erlbaum.

This volume is a classic publication on CAT. It summarizes and discusses over two decades of work on CAT research and development, and charts the future of the CAT industry. The book includes a collection of articles on various CAT-related topics, including the history of CAT, fundamentals of IRT, system design and operations, item pools, testing algorithms, test scaling and equating, reliability, validity, and future directions in this area. The book also presents a discussion of testlets.

UNANNOTATED BIBLIOGRAPHY

American Psychological Association. 1986. *Guidelines for computer-based tests and interpretations*. Washington, DC: American Psychological Association.

Bennet, R. E., M. Steffen, M. E. Singley, M. Morley and D. Jacquemin. 1997. Evaluating an automatically scoreable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement*. 4.162–176.

Bernstein, J. 1997. Speech recognition in language testing. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.) *Current developments and alternatives in language assessment*. Jyväskylä, Finland: University of Jyväskylä. 534–537.

Brown, A. and N. Iwashita. 1996. Language background and item difficulty: The development of a computer-adaptive test of Japanese. *System*. 24. 199–206.

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

Brown, A. and N. Iwashita. 1998. The role of language background in the validation of a computer-adaptive test. In A. Kunnan (ed.) *Validation in language assessment.* Mahwah, NJ: L. Erlbaum. 195–207.

Bunderson, C. V., D. K. Inouye and J. B. Olson. 1989. The four generations of computerized educational measurement. In R. L. Linn (ed.) *Educational measurement*. Washington, DC: American Council on Education. 367–407.

Burstein, J. 1997. Scoring rubrics: Using linguistic description to automatically score free-responses. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.) *Current developments and alternatives in language assessment*. Jyväskylä, Finland: University of Jyväskylä. 529–532.

_____, L. Frase, A. Ginther and L. Grant. 1997. Technologies for language assessment. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics, 16. Technology and Language*. New York: Cambridge University Press. 240–260.

Chalhoub-Deville, M., C. Alcaya and V. M. Lozier. 1997. Language and measurement issues in developing computer-adaptive tests of reading ability: The University of Minnesota model. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.) *Current developments and alternatives in language assessment*. Jyväskylä, Finland: University of Jyväskylä. 546–585.

Davey, T., J. Godwin and D. Mittelholz. 1997. Developing and scoring an innovative computerized writing assessment. *Journal of Educational Measurement*. 34.21–41.

de Jong, J. 1986. Item selection from pretests in mixed ability groups. In C. Stansfield (ed.) *Technology and language testing*. Washington, DC: TESOL. 91–107.

Dunkel, P. 1991. Computerized testing of nonparticipatory L2 listening comprehension proficiency: An ESL prototype development effort. *Modern Language Journal*. 75.64–73.

_____ 1997. Computer-adaptive testing of listening comprehension: A blueprint for CAT development. *The Language Teacher Online*. 21.1–8. [Retrieved August 15, 1998 from the World Wide Web: http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/97/oct/dunkel.html.]

_____ 1999. Research and development of computer-adaptive test of listening comprehension in the less-commonly taught language Hausa. In M. Chalhoub-Deville (ed.) *Issues in computer adaptive testing of reading proficiency*. New York: Cambridge University Press. 91–118.

Eignor, D. 1999. Selected technical issues in the creation of computer adaptive tests of second language reading proficiency. In M. Chalhoub-Deville (ed.) *Issues in computer adaptive testing of reading proficiency*. New York: Cambridge University Press. 162–175.

_____, C. Taylor, I. Kirsch and J. Jamieson. 1998. *Development of a scale for assessing the level of computer familiarity of TOEFL examinees*. Princeton, NJ: Educational Testing Service. [TOEFL Research Report No. 60.]

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

Frase, L., B. Gong, E. Hansen, R. Kaplan, R. Katz and K. Singley. 1998. *Technologies for language testing*. Princeton, NJ: Educational Testing Service. [TOEFL Monograph Series No. 11.]

Fulcher, G. In press. Computerizing an English language placement test. *English Language Teaching Journal*.

Hambleton, R. K. and H. Swaminathan. 1985. *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Henning, G. 1984. Advantages of latent trait measurement in language testing. *Language Testing*. 1.123–133.

_____ 1987. *A guide to language testing: Development, evaluation, research*. Cambridge, MA: Newbury House.

_____ 1989. Does the Rasch model really work for multiple-choice items? Take another look: A response to Divgi. *Journal of Educational Measurement*. 26.91–97.

_____ 1991. Validating an item bank in a computer-assisted or computer-adaptive test: Using item response theory for the process of validating CATs. In P. Dunkel (ed.) *Computer-assisted language learning and testing*. New York: Newbury House. 209–222.

Jamieson, J., C. Taylor, I. Kirsch and D. Eignor. 1998. *Design and evaluation of a computer-based TOEFL tutorial*. Princeton, NJ: Educational Testing Service. [TOEFL Research Report No. 62.]

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

Larson, G. 1999. Considerations for testing reading proficiency via computer adaptive testing. In M. Chalhoub-Deville (ed.) *Issues in computer adaptive testing of reading proficiency*. New York: Cambridge University Press. 71–90.

Laurier, M. 1999. The development of an adaptive test for placement in French. In M. Chalhoub-Deville (ed.) *Issues in computer adaptive testing of reading proficiency*. New York: Cambridge University Press. 119–132.

Linacre, J. M. 1999. A measurement approach to computer adaptive testing of reading comprehension. In M. Chalhoub-Deville (ed.) *Issues in computer adaptive testing of reading proficiency*. New York: Cambridge University Press. 176–187.

Luecht, R. M. 1996. Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*. 20.389–404.

_____ 1998. Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*. 22.222–236.

_____, R. J. Nungester and A. Hadadi. 1996. Heuristic-based CAT: Balancing item information, content and exposure. Paper presented at the Annual Meeting of the National Council of Measurement in Education. New York, 1996.

Lund, R. 1990. A Taxonomy for teaching second language listening. *Foreign Language Annals*. 23.105–115.

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

Lunz, M. E. and B. A. Bergstrom. 1994. An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement*. 31.251–263.

_____ and B. D. Wright. 1992. The effect of review on student ability and test efficiency for computer adaptive tests. *Applied Psychological Measurement*. 16.33–40.

_____ and C. Deville. 1996. Validity of item selection: A comparison of automated computerized adaptive and manual paper and pencil examinations. *Teaching and Learning in Medicine*. 8.152–157.

Mancall, E. L., P. G. Bashook and J. L. Dockery (eds.) 1996. *Computer-based examinations for board certification*. Evanston, IL: American Board of Medical Specialties.

Mead, A. D. and F. Drasgow. 1993. Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*. 114.449–458.

Nandakumar, R. and W. Stout. 1993. Refinement of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*. 18.41–68.

Ordinate Corporation. 1998. *PhonePass test validation report*. Menlo Park, CA: Ordinate.

Pine, S. M., A. T. Church, K. A. Gialluca and D. J. Weiss. 1979. *Effects of computerized adaptive testing on black and white students*. Minneapolis, MN: University of Minnesota. [Research Rep. No. 79–2.]

Sands, W. A., B. K. Waters and J. R. McBride (eds.) 1997. *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Stevenson, J. and S. Gross. 1991. Use of a computerized adaptive testing model for ESOL/bilingual entry/exit decision making. In P. Dunkel (ed.) *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House. 223–236.

Stocking, M. L. 1992. *Controlling item exposure rates in a realistic adaptive testing paradigm*. Princeton, NJ: Educational Testing Service. [Research Report No. 93–2.]

_____ 1994. *Three practical issues for modern adaptive testing item pools*. Princeton, NJ: Educational Testing Service. [Research Report No. 94–5.]

_____ and C. Lewis. 1995. *Controlling item exposure conditional on ability in computerized adaptive testing*. Princeton, NJ: Educational Testing Service. [Research Rep. No. 95–24.]

_____ and L. Swanson. 1993. A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*. 17.277–292.

Stout, W. 1987. A nonparametric approach for testing latent trait unidimensionality. *Psychometrika*. 52.589–617.

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

Swanson, L. and M. L. Stocking. 1993. A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*. 17.151–166.

Sympson, J. B. and R. D. Hetter. 1985. Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association*. San Diego, CA: Navy Personnel Research and Development Center. 973–977.

Taylor C., J. Jamieson, D. Eignor and I. Kirsch. 1998. *The relationship between computer familiarity and performance on computer-based TOEFL test tasks*. Princeton, NJ: Educational Testing Service. [TOEFL Research Report No. 61.]

Tung, P. 1986. New developments in measurement theory: Computerized adaptive testing and the application of latent trait models to test and item analysis. In C. Stansfield (ed.) *Technology and language testing*. Washington, DC: TESOL. 11–27.

Wainer, H. and G. L. Kiely. 1987. Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*. 24.185–201.

Wald, A. 1947. *Sequential analysis*. New York: Wiley.

Weiss, D. J. (ed.) 1978. *Proceedings of the 1977 computerized adaptive testing conference*. Minneapolis, MN: University of Minnesota.

_____ 1982. Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*. 6.473–492.

_____ 1985. Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*. 53.744–789.

_____ and G. Kingsbury. 1984. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*. 22.361–375.

Young, Y., M. D. Shermis, S. R. Brutten and K. Perkins. 1996. From conventional to computer-adaptive testing of ESL reading comprehension. *System*. 24.23–40.

Appendix A

CAT Software Vendors:

Assessment Systems Corporation — MicroCAT
2233 University Ave., Suite 200
St. Paul, MN 55114–1629
USA
Phone: (612)–647–9220
Fax: (612)–647–0412
E-mail: info@assess.com

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press

Computer Adaptive Technologies, Inc. CAT Administrator
2609 W. Lunt Avenue
Chicago, Illinois 60645-9804
USA
Phone: (773)-274-3286
Fax: (773)-274-3287
E-mail: chorwit@catinc.com

Computer Adaptive Technologies, Inc. is also involved in computerized test delivery.

Calico Assessment Technologies, Inc.
Phone: (602)-267-9354
Website: http:\\www.calicocat.com

Computerized Test Delivery Companies

National Computer Systems (NCS)
2510 N. Dodge Street
Iowa City, IA 52245
USA
Phone: (319)-354-9200; (800)-627-0365
E-mail: info@ncs.com

Assessment Systems, Inc. (ASI)
3 Bala Plaza
Bala Cynwyd, PA 19004
USA
Phone: (800)-274-3444
E-Mail: webmaster@harcourtbrace.com

Sylvan Prometric
1600 Lancaster St.
Baltimore, MD 21202
USA
Phone: (410)-843-8000; (800)-627-4276
E-Mail: webmaster@prometric.com

https://doi.org/10.1017/S0267190599190147 Published online by Cambridge University Press